Published in final edited form as:

J Clin Child Adolesc Psychol. 2009 January ; 38(1): 144–165. doi:10.1080/15374410802575461.

Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research

Kimberley D. Lakes and

Department of Pediatrics, University of California, Irvine

William T. Hoyt

Department of Counseling Psychology, University of Wisconsin, Madison

Abstract

Using generalizability theory to evaluate the reliability of child and adolescent measures enables researchers to enhance precision of measurement and consequently increase confidence in research findings. With an observer-rated measure of child self-regulation, we illustrate how multiple sources of error variance (e.g., raters, items) affect the dependability (replicability) of scores and demonstrate methods for enhancing dependability of observer ratings. Using ratings of 181 children, we illustrate the use of two-facet (i.e., raters and items as sources of error) and three-facet (i.e., raters, items and occasions) analyses to optimize design features of future studies using this measure. In addition, we show how generalizability theory provides a useful conceptual framework for thinking about determinants of scores on acquaintance (e.g., teacher or parent) ratings, as well as observer ratings, and sheds light on the strengths and limitations of both types of data for child and adolescent clinical research.

A researcher evaluating a school-based clinical intervention for behavioral problems compares a small group of students receiving the intervention ($n = 25$) to a matched group of students who receive no intervention ($n = 25$) on a teacher-rated behavioral checklist. Although the treatment group scores lower than the control group (by an amount indicating moderately fewer behavior problems; Cohen's $d = 0.45$), this difference between groups does not attain statistical significance, $t(49) = 1.59, p = .12$.

What went wrong here? Low statistical power is one well-known explanation for failure to obtain significant findings in small sample studies. Checking her power tables, the researcher finds that, with 25 students per group, her power to detect a medium effect (i.e., $d = 0.5$) is only .41 (Cohen, 1988). Clearly, a larger sample size would be desirable to test a medium-strength intervention. However, based on clinical testimonies to the effectiveness of this intervention, and knowledge of effect sizes for other similar treatments (e.g., Lipsey & Wilson, 1993), she had expected a strong treatment effect (i.e., $d = 0.8$).

The researcher is also aware that measurement error can contribute to attenuation of measures of association between variables (e.g., Cohen, Cohen, West, & Aiken, 2003, pp.

Copyright © Taylor & Francis Group, LLC

Correspondence should be addressed to Kimberley Lakes, Child Development Center, Department of Pediatrics, University of California, Irvine, 19722 MacArthur Boulevard, Irvine, CA 92612. klakes@uci.edu.

⁷Conger (1981) showed that when both two-facet and three-facet generalizability analyses are conducted on the same sample (which is not the case here), each variance component in the two-facet analysis (e.g., σ_{PR}^2) is partitioned in the three facet analysis into two components, one reflecting stable variance and one reflecting unstable variance (e.g., σ_{PR}^2 and σ_{PRO}^2 respectively).

This article is based on a doctoral dissertation by Kimberley D. Lakes (William T. Hoyt, Chair).

55–57). In a group comparison design, the independent variable (treatment vs. no-treatment) is generally “measured” with near-perfect reliability, but the dependent variable contains some error variance. This results in some attenuation of treatment effects. To estimate the magnitude of this attenuation, it is important to use a reliability coefficient that takes all relevant sources of error into account (Schmidt & Hunter, 1996). The researcher consults previous studies of her dependent measure and finds that this teacher-rated checklist has strong internal consistency reliability (coefficient $\alpha = .85$) and moderate test–retest reliability (3-month test–retest $r = .72$). Which coefficient provides the best indication of measurement error in her data? And what about the fact that students in her study are rated by different teachers, who may vary in their severity or leniency? Is rater error a source of error that should be considered, both in interpreting her past findings and in planning future studies using this measure?

We have four main goals for this article. First, we hope to help readers consider the effects of measurement error on findings in clinical child and adolescent research. Second, we want to help readers understand the limitations of common methods of measuring reliability. We note that multiple sources of error are relevant to most types of measures used in child and adolescent research, such that traditional reliability coefficients (which usually omit one or more relevant error sources from consideration) do not provide a good indication of the effects of measurement error on study findings. Third, we provide a primer on generalizability theory (GT), which simultaneously examines the effects of multiple sources of error, and show how this approach constitutes both a useful computational framework (for quantifying the effects of error on study findings) and a useful conceptual framework (for understanding the strengths and weaknesses of different measurement approaches) for researchers studying children and adolescents. Finally, we discuss specific steps that researchers can take to begin utilizing GT to improve measurement in their fields of research.

SCORE RELIABILITY AND ERROR OF MEASUREMENT

Measurement error is a fact of life in psychological research and is usually defined by reference to the classical theory of reliability. As Stanley (1971) put it,

Two sets of measurements of the same features of the same individuals will never exactly duplicate each other. ...[This] is what is meant by *unreliability*. At the same time, however, repeated measurements of a series of objects or individuals will ordinarily show *some* consistency. The block of wood that was the heaviest the first time the set of blocks was weighed will tend to be among the heaviest blocks the second time, and consistency will be the rule among all the blocks of the set. The same, to a degree, will be the case for the weights of the boys in a classroom or for their performance on a test of reading comprehension. This tendency toward consistency from one set of measurements to another is called *reliability*. (p. 356)

The presence of measurement error in predictor or criterion variables adversely affects our confidence in test scores and distorts research findings (e.g., *attenuates* effect sizes such that observed correlations between scores on error-prone measures will systematically underestimate the true correlation between the constructs of interest in the population; Schmidt & Hunter, 1996). Almost always, measurement error reduces the statistical power of a research design, increasing the probability of Type II errors (Hoyt, 2000). For these reasons, researchers and practitioners employing psychological measures are advised (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Wilkinson and APA Task Force on Statistical Inference, 1999) to report on the reliability of scores based on these measures and

to take unreliability of measurement into account in interpreting test results and research findings.

Typically, users of psychological assessments have employed *reliability coefficients* to quantify the proportion of systematic or replicable (i.e., nonerror) variance in a set of scores. In this article, we briefly describe traditional reliability coefficients and show how these often overestimate the contribution of replicable (“true score”) variance and underestimate the contribution of error variance in a set of scores. This is a problem even for ability tests and self-report questionnaires (which have been the subject of the preponderance of the psychometric literature) and tends to be more of a problem for measures in which raters (e.g., teachers, parents, or trained observers) provide scores for participants based on observation of behavior. Because other sources such as parent, teacher, observer, and peer reports are frequently used in the assessment of child and adolescent clinical problems, the inadequacy of traditional reliability coefficients is particularly acutely felt in this research area. We introduce GT (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) as a more flexible and hence more accurate means of quantifying the extent to which observed scores reflect error of measurement rather than the characteristics of the persons under study.

Limitations of Traditional Methods of Assessing Reliability

By the 1940s, numerous coefficients had been proposed to quantify the replicability or dependability of scores on psychological tests (Cronbach, 1947). In classical test theory, the reliability coefficient is interpreted as the proportion of the observed score variance that is attributable to *true scores* (i.e., true differences between persons on the measured construct):

$$r_{XX} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}, \quad (1)$$

where r_{XX} is the reliability coefficient, σ_T^2 is the variance attributable to true scores, and σ_E^2 is error variance. The denominator represents the variance of the observed scores (true score variance plus error variance). Thus, for a measure with an internal consistency reliability (e.g., coefficient alpha) $r_{XX} = .80$, it is understood that 80% of the variance reflects true differences among persons on the measured construct and the remaining 20% reflects measurement error.

Although all represent estimates of r_{XX} in Equation 1, different classes of reliability coefficients are not equivalent to one another, because each coefficient embodies a different definition of measurement error (Schmidt, Le, & Ilies, 2003; see also Brennan, 2001, pp. 127–129). For example, *coefficients of stability* (i.e., test–retest reliability) define as true score variance all variance that replicates over testing occasions—that is, all variance that cannot be attributed to either *transient error* or *random response error*. Transient errors arise because of ephemeral conditions within the person (e.g., mood) or in the environment (e.g., idiosyncratic daily events) that affect scores on a measure. Such ephemeral factors cause inconsistencies in persons’ scores on the measure between occasions (i.e., test–retest $r_{XX} < 1.0$), which remind users that scores based on a single measurement occasion are a somewhat unreliable measure of each person’s true standing on the construct of interest.

Coefficients of equivalence (i.e., internal consistency reliability) define as true score variance all variance that replicates over items—that is, all variance that cannot be attributed to *specific factor error* or *random response error*. Specific factor error arises because the selection of items on a measure does not perfectly reflect the underlying construct. Thus, two parallel forms of the same assessment procedure will not correlate perfectly even if administered at the same time (to minimize inconsistencies due to transient error), because

each measures a somewhat different range of item content (as reflected in the differential sampling of items).

Because all three sources of error (i.e., transient error, specific factor error, and random response error) are problematic for virtually all psychological measures, both coefficients of stability and coefficients of equivalence underestimate the proportion of error in observed score variance, and therefore overestimate reliability of measurement (Cronbach, 1947). For example, when a given measure has been shown to have internal consistency $r_{XX} = .85$ and test-retest $r_{XX} = .72$, there is ambiguity about the proportion of variance (85% vs. 72%) in observed scores that reflects true differences among persons.

In fact, when the investigation focuses on relations among trait constructs, as is often the case, both of these percentages are too high. As Schmidt et al. (2003) pointed out, the former coefficient is inflated in that transient error variance (which replicates over items) is erroneously included in the estimate of true score variance; the latter coefficient fails to differentiate specific factor variance (which replicates over occasions) from variance in true scores. Thus, the actual reliability for this measure (taking all three sources of error into account), is somewhat less than the lower of these two coefficients (i.e., $r_{XX} < .72$; Schmidt et al., 2003).

The inadequacy of traditional reliability coefficients becomes even more acute when observer ratings (e.g., parent and teacher ratings) rather than self-ratings are the source of scores to be analyzed. Observer ratings introduce yet a fourth¹ source of error, *rater bias*, along with an additional design complication. Specifically, it is often the case that ratings for different participants are provided by different observers (i.e., a nested, rather than crossed rating design). Rater bias can be a substantial source of error in scores, especially when observers receive little training and are asked to rate attributes requiring inference (rather than readily observable behaviors; Hoyt & Kerns, 1999). Rater biases reduce reliability most acutely when raters are assigned to participants in a nested rating design (as in studies where each child or adolescent is rated by his or her own parent, or where different teachers provide ratings of different children), and when each participant is rated by only one or two raters (Hoyt, 2000). Although coefficients of interrater reliability are available (Shrout & Fleiss, 1979), these estimate error attributable to rater bias and random response errors but ignore transient error and specific factor error. More precisely, they treat transient and specific factor errors, which replicate over raters, as parts of true score variance. Thus, users of observer ratings are well advised to conduct generalizability (G) studies that address the combined impact of these multiple sources of error.

A PRIMER ON GT

GT was developed by Cronbach et al. (1972; for earlier formulations, see Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965) as an expansion of classical approaches to estimating reliability of measurement. Recognizing that error of measurement is not unitary or undifferentiated, they sought to develop a psychometric theory that would encompass multiple sources (or *facets*) of error. In any instance of psychological measurement,

the [observed] score is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be

¹Actually, rater bias encompasses two distinct sources of error—rater main effects, which contribute to error variance in some rating designs, and Person $\times$ Rater interactions, which always contribute to error variance in ratings (Hoyt, 2000; Longford, 1994).

altered without making the score any less acceptable to the decision maker. (Cronbach et al., 1972, p. 15)

The items on a questionnaire are not sacrosanct. They are interchangeable with a *universe* of alternative items that embody the same construct. Similarly, the observer who provides a behavioral rating is in principle interchangeable with a universe of other observers with equivalent skill and training. And when we measure characteristics that are theorized to be enduring, we presume that measurements taken at several not-too-distant points in time would yield equally valid scores.

That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. *The question of "reliability" thus resolves into a question of accuracy of generalization, or generalizability.* (Cronbach et al., 1972, p. 15)

Thus, in GT, as in classical test theory, *error* is equated with variance in observed scores that is attributable to facets (e.g., particular items, raters, or measurement occasions) that are irrelevant to the construct of interest, as well as to unexplained variations in responding (i.e., random error). The classical concept of a "true" score is replaced by the notion of the *universe score*—the hypothetical mean of all acceptable (i.e., interchangeable) observations. The undifferentiated error term in classical test theory is partitioned in GT into components attributable to the main effect of each facet, as well as interactions between facets, and between facets and the object of measurement (here referred to as "persons").

Finally, similar to classical reliability theory, GT quantifies dependability of measurement by a coefficient of generalizability (or *g* coefficient) that represents the ratio of variance attributable to universe scores to the total observed score variance (i.e., universe variance plus variance attributable to all sources of error that contribute to variance in observed scores). Intermediate to the computation of this coefficient is a variance partitioning procedure (analogous to factorial analysis of variance [ANOVA]), the results of which are highly informative as they enhance understanding of the nature of measurement error and suggest ways of optimizing dependability of measurement by a judicious choice of measurement procedures in future studies.

HOW IS VARIANCE PARTITIONED IN THE G STUDY?

Example Data Set

To illustrate the variance partitioning procedure, we consider the simple data set shown in Table 1, in which each of five targets is rated by each of three observers or "raters." The grand mean rating for this data set (average of all 15 ratings in the table) is 18.24. We use the term *targets* rather than persons, to remind readers that the object of measurement in a G study is not necessarily a person. For example, in meta-analysis, raters frequently code study characteristics; the target of ratings is a study, rather than a person in this context. As this G study includes a single source of error (i.e., raters), it is designated as a *one-facet* G study.

Examining the row marginals in Table 1 (found in the column labeled Target Mean), we can see that there is variability in how raters, on average, perceive targets. These mean ratings range from 9.67 (Target 4) to 27.67 (Target 1). This variability is the basis for the estimate of target (*T*) variance in the G study. The column means also vary, indicating that Rater 2 is somewhat more lenient (i.e., gives higher average ratings) than Raters 1 and 3. This variability is the basis for the estimate of rater (*R*) variance in the G study.

Finally, if we were to predict the values in each cell from the rater mean and the target mean, our predictions would not be perfectly accurate. For example, consider the top left cell in Table 1, which represents the rating of Target 1 by Rater 1. Target 1's mean score was 27.67, which is 9.43 points higher than the grand mean of 18.24. The mean rating given by Rater 1, averaged across all targets, was 16.40, which is 1.84 points below this grand mean. Using only this information, our best prediction of Rater 1's rating of Target 1 is the grand mean plus these two deviation scores, or $18.24 + 9.43 - 1.84 = 25.83$. This predicted rating is not exactly equal to the observed rating in the top left cell, which was 27. The difference between the predicted and observed ratings is called the *residual*, and the variance of these residuals (if computed for each of the 15 cells in Table 1) is the basis for the estimate of the variance attributable to the target $\times$ rater interaction. To remind us that this variance estimate is confounded with error, it can be designated as $\text{var}(TR,e)$.

Computation of Variance Estimates

A two-way ANOVA provides the basis for estimating the variance due to each of the sources in this one-facet G study design. Table 2 shows the sum of squares (SS), degrees of freedom (df), and mean square (MS) for Targets, Raters, and residuals (TR,e). G studies typically use a random effects model, in which levels of each facet (i.e., the three raters, in our example data set) are treated as representative of a broader universe of admissible observations to which one seeks to generalize. In random effects ANOVA, the mean squares for Targets and Raters also contain variance attributable to the TR,e interaction (cf. Kirk, 1982, p. 247). Thus, the variance estimate for T (also called the *variance component* for T , or $\text{var}(T)$) is computed as a linear combinations of $MS(T)$ and $MS(TR,e)$, and similarly for $\text{var}(R)$. The variance estimate for TR,e is identical to $MS(TR,e)$, so does not require additional computation.

Standard errors for variance estimates are not always reported in G studies. However, this is good practice, as it gives an indication of the precision of each variance estimate. As in any statistical analysis, when a G study is performed on a small sample, effect sizes are estimated with limited precision, so caution should be exercised in generalizing these findings to future studies that will use this same measurement procedure. Variance percentages are also not always reported, but these can be an aid to interpretation of findings, especially in more elaborate G studies involving multiple facets. This column represents each variance estimate as a percentage of the total variance (i.e., of the sum of all variance estimates in the study). It should be noted that this total variance (52.29 in Table 2) is often larger than the observed variance in scores (45.69 for the 15 scores in Table 1). The reason for this is that not all variance components in the G study contribute to observed variance in scores. This fact is important in choosing a measurement design (e.g., raters crossed with targets versus raters nested within targets) and in computing appropriate g coefficients for this design and is discussed in detail with reference to the real data set considered in the following sections.

Estimating Variance Components: Practical Considerations

Variance estimates for simple G study designs can be computed as linear combinations of ANOVA mean squares. Formulas for deriving variance estimates from mean squares in basic designs, along with a discussion of the basis for this approach, are available in Shavelson and Webb (1991). For more extensive technical discussions, and an approach to deriving corresponding formulas for any G study design, see Brennan (2001) or Cronbach et al. (1972).

For routine work involving GT, it is almost always easier to make use of a statistical software package to compute estimated variance components. GENOVA (Crick & Brennan,

1982) is the best known software package specifically designed to conduct generalizability analyses. It is available for free download (http://www.education.uiowa.edu/casma/computer_programs.htm#genova) for either Macintosh or PC platforms and includes a manual describing basic and advanced applications. GENOVA outputs both variance estimates and their estimated standard errors but requires complete data. An alternative is available in the SAS software package, in the VARCOMP procedure. This procedure uses the standard SAS interface, which is an advantage for researchers already familiar with SAS. It also is tolerant of missing data, which eliminates the need to impute missing values. However, PROC VARCOMP provides variance estimates but not their standard errors, so information about the precision of these estimates is lacking. In the appendix, we provide syntax for analyzing the example data set in both GENOVA and SAS.

Interpretation of Variance Components

In a subsequent section, we show how variance components from two-facet and three-facet G studies can be used to estimate g coefficients for a variety of measurement designs. In this section, we confine ourselves to observing that the one-facet G study yields findings equivalent to a conventional reliability study. Hoyt and Melby (1999, pp. 335–339) provide an extended discussion of a one-facet G study with raters as the facet of interest. They show how two g coefficients computed from such a study are equivalent to Shrout and Fleiss's (1979) ICC(2, k) and ICC(3, k) and clarify how to choose the coefficient that is appropriate for particular measurement designs. Hoyt and Melby also noted (see their footnote 2) that if the G study involves items rather than raters, the usual g coefficient is equivalent to Cronbach's (1951) coefficient alpha for this collection of items.

Although the g coefficients derived from these one-facet G study designs are equivalent to conventional reliability coefficients, the G study may provide additional information that will help users to optimize measurement design in future studies. However, the unique benefits of GT are most evident when (as is so often the case in psychological research) more than one source of error contributes to variance in observed scores. We turn now to illustrative G studies, using actual data, examining two and three sources of error.

TWO FACET G STUDY (*PRI*)

Variance Partitioning

We illustrate this variance partitioning procedure in a G study in which a group of trained observers (or raters; $n_r = 5$) rated elementary school children (or persons; $n_p = 181$) on three multi-item scales designed to measure different domains of self-regulation ($n_i = 6, 7$, and 3, respectively, for cognitive, affective, and physical regulation scales).² After discussing the variance partitioning from this G study, we illustrate how these data would be used to estimate g coefficients for future studies (usually termed decision studies or *D studies*) using this rating scale in a number of different possible rating designs.

Table 3 lists the components whose contributions to observed scores can be estimated in the *PRI* design. As in a factorial ANOVA, the score assigned to person p by rater r on item i is conceptualized as a deviation from the grand mean (over all persons, raters, and items), with the degree of deviation determined by p 's person effect (i.e., universe score), r 's rater effect, and i 's item effect. In addition to these main effects, each facet (R or I) interacts with the object of measurement (P) in a two-way interaction (PR and PI , respectively), as well as

²For illustrative purposes, we provide separate G analyses of each of the three scales. To understand the relationships among these sub-scales (which have interscale correlations ranging from $r = .75$ to $r = .90$), one would ideally use a *multivariate* G analysis, which estimates component covariances as well as variances. A detailed treatment of this multivariate generalization of GT is available in Brennan (2001, chap. 9–10; see also Cronbach et al., 1972, chap. 9).

with the other facet (RI). Finally, there is also a three-way interaction, which is confounded with error (PRI_e). The right column in Table 3 describes the interpretation of each effect as it contributes to a given rating. Generalizability analysis uses random effects ANOVA, with the raw scores (ratings) as input, to estimate the variance of each effect in the model—a process called *variance partitioning*. We briefly describe the procedures from our example study, then discuss the interpretation of variance components from the G study in some detail, to illustrate how the interpretations in Table 3 inform our understanding of G study results.

Example Study Procedures

Students enrolled in K–5 classrooms in a private lower school completed a challenge course of increasing difficulty and were rated one at a time using the Response to Challenge Scale (RCS), a theory-derived, observer-rated measure of children's self-regulation in response to a physically challenging situation (obstacle course; Lakes & Hoyt, 2004). The RCS asks raters to make inferences, based on the target person's verbal and especially nonverbal behavior, about his or her self-regulatory abilities in three domains: physical, cognitive, and affective/motivational. Five raters evaluated each child's responding to a challenge course using 16 bipolar adjectives (e.g., *Distractible*—*Focused*) rated on 7-point scales. Negatively scored items were reversed prior to conducting generalizability analyses. Raters had no prior acquaintance with participants and received 30 min of training, which included examples of strong and weak performances and corresponding ratings. Raters were told to rate children on a developmentally appropriate level by anchoring ratings on the first child of a given grade level to rate all children of that grade category in comparison to their age-level peers. The challenge course was adapted for each grade level to increase the level of difficulty for the older children. Adaptations included increasing the number and difficulty of tasks.

Variance Estimates

In this G study, every person (child) was rated by all five raters, on all 16 RCS items. This makes it a fully crossed design denoted as Persons $\times$ Raters $\times$ Items or PRI . We used GENOVA (Crick & Brennan, 1982) to compute variance estimates. Results from PRI analyses for each of the three RCS subscales (based on preintervention ratings) are shown in Table 4.

What does P represent?—The variance estimate for Persons (P) represents variance in ratings attributable to differences in children's actual standings (universe scores) on the characteristic of interest. Psychometrically sound measures are designed to maximize $\text{var}(P)$. Other components, if they contribute to variance in observed scores, usually count as error. The percentage of variance attributable to Persons was largest for the Physical subscale (47%) and smallest for the Cognitive subscale (32%), suggesting that either (a) the former construct is easier to rate in the current context or (b) individual differences were somewhat larger for physical regulation than for cognitive regulation in this population of children.

What does R represent?—The R variance component estimates between-rater variance for all raters, averaged over persons and items. Because all items are scaled so that higher scores reflect superior self-regulation ability, substantial R variance can be interpreted as systematic differences in rater leniency. Raters with a positive R effect tended to evaluate participants favorably; those with a negative R effect have average ratings below (i.e., less favorable than) the mean. The percentage of variance attributable to R provides information regarding the extent to which rating variance is attributable to mean differences among raters. R variance for the RCS accounted for between 3% (Affective) and 6% (Cognitive) of ratings variance.

None of the R variance estimates was statistically significant, which is not surprising, given the relatively small n_r . Variance estimates for facet main effects (such as R or I in the $PR I$ design) tend to be the least stable and can fail to achieve statistical significance even when these account for an appreciable proportion of total score variance. Shavelson and Webb (1981; see also Smith, 1978) recommended G study designs with at least eight levels per facet (i.e., $n_r = 8$) to achieve reliable estimates of facet main effects. In generalizability analysis, the focus is on obtaining the best estimates possible of the variance of each component, not on testing whether this variance differs significantly from zero. Hence, our focus is on the numerical variance estimates, with attention to their standard errors as an indication of their precision.

What does I represent?—The variance component for items (I) estimates between-item variance in ratings, averaged over persons and raters. When I variance is large, this means that Persons are rated consistently higher on some items than others. On ability tests, this represents variance in item difficulties. I variance in this study accounted for between 0% (Cognitive, Physical) and 5% (Affective) of rating variance. Thus, there was little evidence of mean differences among items on these scales.

Four interactions among P , R , and I are included in the two-facet analysis. Most important among these are the three interactions involving Persons: PR , PI , and $PR I$. When substantial variance is attributable to one or more of these components, the ordering of Persons will differ depending on which levels of a particular facet (R or I) or combination of facets (RI) are used to establish scores in the D study.

What does PR represent?—The PR variance component estimates the variance due to inconsistencies in different raters' evaluations of the same person, averaging over items. PR variance accounted for between 14% (Cognitive, Affective) and 19% (Physical) of total variance. Thus, raters differed appreciably in their rank ordering of persons. The PR component indexes a lack of consensus among raters on the standing of the target persons on the dimension of interest.

In a review of research on consensus in interpersonal perceptions, Kenny (1994) showed that the two most important obstacles to consensus are lack of overlap (i.e., different raters observe different target person behaviors) and dissimilar meaning systems (i.e., different raters interpret the same behavior in different ways). In ratings using the RCS, all raters had an opportunity to observe the same set of target person behaviors (high overlap). So it is most likely that the lack of consensus is attributable to raters' dissimilar meaning systems. It may be possible to reduce this source of error with further rater training. Otherwise, it will be important to aggregate ratings of multiple raters, as described later.

What does PI represent?— PI variance estimates the variance due to inconsistencies from one item to another for a person, averaging over raters. PI variance accounted for between 3% (Affective, Physical) and 5% (Cognitive) of total variance. This indicates that ordering of persons differs somewhat on different items. Findings indicate that although all items in a given subscale may share variance attributable to a common underlying factor (e.g., cognitive self-regulation), items also embody some specific factor variance (e.g., attentive, involved in task).

As with PR variance, PI variance can be reduced as a source of error by aggregation (i.e., using multiple items to evaluate each person and aggregating scores across items). When items are examined as the only source of error in a reliability study, the well-known Spearman-Brown prophecy formula can be used to forecast the improvement in internal consistency reliability from increasing the scale length by some desired number of items.

Next we describe procedures for simultaneously optimizing the number of levels of two or more facets (sources of error: here, the facets considered are R and I), given the raw variance estimates from a G study.

What does RI represent?—The RI component examines evidence that ordering of raters differs for different items, averaging over persons. When RI variance is substantial, rater leniency varies from item to item. That is, the rater with the highest average on one item may have only a midrange average on another item in the same subscale. RI variance accounted for between 2% (Physical) and 4% (Cognitive, Affective) of variance in ratings.

What does PRI,e Represent?—The final variance component (PRI,e) encompasses the three-way interaction between P , R , and I and the residual variance due to random error and other factors (i.e., any unanalyzed facets of measurement that varied among persons). The three-way interaction (PRI) means that PR interactions are inconsistent across items (I) within the subscale. Because each item is rated only once for each P - R pair, it is impossible to determine what part of the PRI variance is stable over replications and what part is variable (i.e., residual or random error variance). This confounding is usually present for the highest order interaction in G analyses, because (typically) there is only a single observation per cell of this interaction. (Later we consider a three-facet G analysis of a reduced data set that suggests that these PRI interactions are not stable over rating occasions.) Because of this confounding, the highest order interaction in a G study is also referred to as the *residual* component. PRI,e variance accounted for between 25% (Physical) and 39% (Cognitive) of the variance in ratings.

GENERALIZABILITY COEFFICIENTS

In a classical reliability study, the investigator computes a reliability coefficient, which estimates the proportion of variance in a set of scores that is attributable to actual differences in true scores (see Equation 1). Similarly, an important result of a G study is a generalizability (g) coefficient, which estimates the proportion of variance in a set of scores that is attributable to universe score variance. However, each G study yields many different possible g coefficients, corresponding to the different possible measurement design of future studies (called D studies) that may use this measurement procedure. Because the distinction between G and D studies is likely to be unfamiliar to readers new to GT, we clarify the definition of these terms here.

G Studies and D Studies

Among researchers working within the framework of classical reliability theory, it is not uncommon to introduce new measures using what might be called a *measurement development* study. Such a study may or may not involve an investigation of a substantive research question, but an important focus of the measurement development study is to provide reliability and validity data on the measure, for a particular population and research intent. A measurement development study may include a factor analysis to assess the structure of a scale, and reports of internal consistency and test–retest reliability, as well as correlations with existing measures of similar and dissimilar constructs. Future researchers wishing to use this measure will refer back to the measurement development study for evidence on the reliability and interpretation of scale scores, which will be used in designing and interpreting findings of their own studies.

In GT, the G study plays a similar role to the investigation of reliability in a traditional measurement development study. It is a study of primarily psychometric import, designed to address questions about the generalizability of scores across one or more relevant facets of measurement. Investigators may choose to address substantive questions as well as

psychometric questions in this initial investigation, but a primary purpose of the G study is to provide information about the dependability of scores on the measure—information that will guide future users in the design and interpretation of their research. In GT, these studies are referred to as D studies, to indicate that their focus is substantive (i.e., on a theoretical or practical decision) rather than primarily psychometric. Cronbach et al. (1972) recommended that investigators developing a new measurement procedure first conduct a G study, which would then guide the design and interpretation of future D studies.

In Table 4, the percentage of variance for $P(\%var)$ is interpretable as a g coefficient, albeit for a D study design that would likely never be used in practice. In this hypothetical design, each person is evaluated by a single rater on a single item, and a unique rater-item pair is assigned to each person. As just noted, the appropriate g coefficient for a given D study will depend on the design of that D study—a topic to which we turn in the next section.

Importance of D Study Design for Computing the G Coefficient

In GT, it is recognized that investigators conducting a D study may not use the same measurement design (e.g., number of raters, number of items) that was used in the G study. Indeed, one appeal of GT is that it can help to inform choices about D study design. Consider this illustration from Brennan (2001):

Although all measurements are fallible to some extent, scientists seek ways to increase the precision of measurement. To do so, they frequently average measurements over some subset of predefined conditions of measurement A substantive question becomes, “How many instances of which conditions of measurement are needed for an acceptably precise measurement?” For example, if prior research has demonstrated that the choice of ruler has little influence on measurements of the length of certain objects, but considerable variability is associated with the persons who record measurements, then it is sensible to average measurements over many persons but few rulers. (p. 1)

Thus, the expected g coefficient for a given set of scores will be a function of the measurement design to be used in the D study. There are two main design decisions made in planning the D study: (a) how many levels of each facet to include (e.g., n'_r refers to the number of raters who will provide data on each Person in the D study; n'_i refers to the number of items on which each Person will be rated), and (b) whether the measurement design is crossed or nested for each facet. Figures 1 and 2 illustrate the distinction between crossed and nested facets in a one-facet study involving Persons and Raters. When Persons are *crossed* with Raters (a PR design, shown in Figure 1), every person in the study is rated by the same set of raters. This design is often possible when trained observers provide ratings of live or recorded behavior. When Persons are *nested within* raters (a $P:R$ design, shown in Figure 2), each rater evaluates multiple persons, but each person is evaluated by only one rater (or one *set* of raters, for nested designs with $n'_r > 1$). This design is used when teachers rate multiple students in their classes, but each student is rated by only a single teacher. Other possible designs include an $R:P$ design (e.g., each person is rated by two parents) and hybrid or *incomplete* rating designs (e.g., each person is rated by 2 randomly selected raters from a pool of 10 trained raters).

Thus, a D study using the RCS could potentially use more or fewer items per rating dimension than our illustrative G study, could use more or fewer raters than our G study, and could nest rather than cross either facet with persons in the measurement design. In practice, it is almost always the case that Items are crossed with Persons. In procedures employing trained observers, raters are usually either completely crossed or incompletely crossed (i.e., a hybrid design). Because a variety of hybrid designs are possible, and no

analytical procedures are available for estimating a g coefficient for these designs, we consider two possible designs for future D studies using the RCS: the fully crossed PRI design, which optimizes generalizability, and also the inferior (from the standpoint of generalizability) $(P:R)I$ design in which persons are nested within raters. As we describe later, g coefficients for hybrid designs are expected to fall between these optimal and inferior estimates, for given values of n'_r and n'_i .

Computation of G Coefficients

As mentioned earlier, the raw proportion of variance for Persons in Table 4 is interpretable as a g coefficient, although this coefficient is applicable to a minimal D study design that would not likely be used in practice ($n'_r=1, n'_i=1$; and a different rater-item pair is used for each person). When scores for each person are derived as an aggregate over multiple items, and often over multiple raters as well, there is a reduction in (error) variance expected because of items and raters, whereas the Person variance remains the same. This results in a reduction in the total variance, and a higher coefficient of generalizability, compared with the minimal design. In this section, we show how to use the raw variance estimates obtained in a G study to estimate g coefficients for D studies in which $n'_r>1, n'_i>1$, or both.

The g coefficients for D studies are conventionally notated as $E\hat{\rho}^2$ (Brennan, 2001), to signify that the g coefficient estimates the squared correlation between observed scores and universe scores. As in classical reliability theory, this coefficient also corresponds to the expected correlation between two sets of scores, both collected using the specified D study design, but with different levels of the facets of interest. In our example, $E\hat{\rho}^2$ for a fully crossed design with $n'_r=2, n'_i=5$ estimates the correlation between one set of scores based on ratings by two raters using a five-item scale, and a second set of scores based on ratings by two new raters using a different (but parallel) five-item scale.

G coefficients and D study design (nested vs. crossed)—To simplify our notation, instead of $E\hat{\rho}^2$, we refer to the g coefficient³ from the nested $[(P:R)I]$ rating design as $G1$ and the g coefficient from the fully crossed (PRI) rating design as $G2$.

The formulas⁴ for $G1$ and $G2$ are as follows:

$$G1 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_R^2 + \sigma_{PR}^2 + \sigma_{PI}^2 + \sigma_{RI}^2 + \sigma_{PRI,e}^2} \quad (2)$$

³ $G1$ reflects the proportion of observed score variance that is attributable to persons (universe scores) when persons are nested within raters, as is the case when several teachers at each grade level provide ratings on each student within that grade level. This same coefficient applies when raters are nested within persons $[(R:P)I]$, as when data are based on parent ratings. Both of these rating designs are unlikely to be used with the RCS. $G1$ is computed here for illustrative purposes, to illustrate the cost (in terms of added measurement error) when rating designs include nesting involving persons (object of measurement).

In computing these $G1$ coefficients, we presume that the variance partitioning in Table 4 (derived from lightly trained observers) would also apply to teacher (or parent) ratings. Although one would prefer to use variance components derived from ratings of actual teachers (or parents), such estimates are often impossible to derive from D study data for rating designs in which raters are selected on the basis of their acquaintance with target persons (Hoyt, 2000). For practical reasons, researchers often recruit only a single rater for each target person in such D studies. In the extreme case, a different rater provides scores on each target person, as when each child in a study is rated by his or her parent. In a D study with this design (i.e., $R:P$ with $n'_r=1$), σ_p^2 is confounded with both σ_R^2 and σ_{PR}^2 , so that no pure estimate of variance attributable to persons is available from the D study data.

$$G2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{PR}^2 + \sigma_{PI}^2 + \sigma_{PRI,e}^2} \quad (3)$$

Like the reliability coefficient in classical test theory (see Equation 1), each g coefficient represents a ratio of person-based variance to total variance. However, in contrast to Equation 1, Equations 2 and 3 partition error variance into variance from multiple sources that contribute to variance in observed scores. In addition to estimated universe score variance (σ_p^2), the denominator for $G1$ (see Equation 2) includes estimates of (error) variance attributable to R , PR , PI , RI , and PRI,e components. I variance creates a constant effect for all persons (and hence does not contribute to error variance) as long as all persons are rated on the same set of items.⁵ The denominator for $G2$ (see Equation 3) includes estimated variance attributable to universe scores (P) and three sources of error (PR , PI , and PRI,e). Because both raters and items are crossed with persons in this D study design, neither R nor I nor their interaction (RI) contributes to error variance. Each of these components contributes an effect that is constant over persons, so they do not contribute to variance in observed scores, and can therefore be omitted from the denominator of Equation 3. Comparing Equations 2 and 3, one can see why the PRI design is preferable to the $(P:R)I$ design from the standpoint of generalizability. Variances must be positive (or zero), so Equation 2 will almost always yield a larger denominator (and hence a smaller g coefficient) than Equation 3. Nonetheless, for some D studies a fully crossed design is not an option, and Persons must be nested within Raters (or vice versa). The decrease in generalizability of ratings will depend on the size of the components present in Equation 2 and absent in Equation 3 (i.e., σ_R^2 and σ_{RI}^2). This decrease may be minimal when a rating scale requires little inference (cf. Hoyt & Kerns, 1999), so that these components are relatively small.

⁴Textbooks on GT (i.e., Brennan, 2001; Cronbach et al., 1972) use a more elaborate notational system, in which lowercase subscripts (p , r , and i) denote facets for which only a single level contributes to the designated effects (or variances), and uppercase subscripts (R , and I) denote facets that are aggregated across multiple levels (as when scores are averaged over several raters and items for each person). The equivalent formulas in the conventional notation are:

$$G1 = \frac{\sigma_p^2}{\sigma_{p:R}^2 + \sigma_R^2 + \sigma_{RI}^2 + \sigma_{pI:R,e}^2} \quad (3)$$

$$G2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pR}^2 + \sigma_{pI}^2 + \sigma_{pRI,e}^2}$$

In these formulas, R and I are uppercase subscripts, to indicate that the variance component reflects aggregation over several levels of each of these facets. Next (see Table 5), we explain in detail the effect of aggregation on each variance component in the PRI design. For simplicity, we use a more basic notation here, with uppercase subscripts for both G study variance components (which are computed without aggregating across levels within facets) and D study variance components (which are computed with aggregation). Because our variance estimates are derived from a G study using a design in which raters are crossed with persons, we do not notate

“nested” (i.e., confounded) variance components in the denominator of $G1$. $\sigma_{p:R}^2$ in the conventional notation includes both σ_p^2 and σ_{pR}^2 (which has been aggregated over raters) in our Equation 2; $\sigma_{pI:R,e}^2$ in the conventional notation includes both σ_{pI}^2 and $\sigma_{pRI,e}^2$ (which have been aggregated over both raters and items) in our Equation 2.

⁵This discussion (and Equations 2 and 3) assume that the D study investigator will use scores to make what Cronbach et al. (1972) called a *relative* decision. Relative decisions involve comparisons of the relative standing of persons within the D study. Examples of relative decisions include correlation or regression analyses involving aggregate scores, or ANOVA analyses using scores as a dependent variable. Virtually all research applications of observer ratings will fall into this category. Cronbach et al. also described a second class of decisions, termed *absolute*, in which scores are compared to some external criterion (e.g., a cutoff score to determine a standard level of competence). When scores are to be used to make an absolute decision, main effects of crossed facets contribute to error variance, so that (e.g.) σ_I^2 would be included in the denominator in Equation 2 (and σ_I^2 , σ_R^2 , and σ_{RI}^2 would be included in the denominator of Equation 3). Thus, dependability coefficients for absolute decisions are generally somewhat lower than the comparable g coefficients for relative decisions. To reiterate, it is our belief that virtually all uses of observer ratings in clinical child research involve relative decisions, so our presentation of GT focuses on this case.

G coefficients with multiple levels of each facet in the D study—When multiple raters and items are used (as is usually the case in a D study), variance components involving R and I are reduced proportionally. To estimate its contribution to the total variance in this design, each component is assigned a multiplier (i.e., a weight). Person (universe) variance does not change with changes in n'_r and n'_i , and so always receives a weight of 1. Components that do not contribute to the total variance, because some facets are crossed with Persons in the rating design, receive a weight of 0. Variance attributable to components containing a single facet (i.e., R , I , PR , and PI here) are reduced in proportion to the number of levels of that facet to be included in the D study, and so receive weights of $1/n'_r$ (for included facets involving R) or $1/n'_i$ (for included facets involving I). Variance from components involving both R and I (i.e., PR , PI , and also RI if it contributes to the total variance) are reduced by factors of both n'_r and n'_i and so receive a weight of $1/(n'_r n'_i)$.

Table 5 illustrates these procedures for the RCS Affective scale in a D study design with five raters ($n'_r=5$) rating all target persons on the same set of six items ($n'_i=6$), to arrive at a $G1$ of .91 and a $G2$ of .92. This is interpretable as a highly dependable measure of affective self-regulation in this D study, where rater variance is reduced by using multiple raters and where persons are fully crossed with raters and items.

Tabulating Predicted G Coefficients for Various D Study Designs

Impact of design decisions on predicted generalizability of scores—Table 6 provides estimated g coefficients ($G1$ and $G2$) for each subscale of the RCS based on mean variance estimates from Table 4, for various numbers of raters (n'_r) and items (n'_i). Using a table of estimated g coefficients such as Table 6 allows researchers to better allocate their resources without necessarily sacrificing generalizability. For example, a researcher who would like to have a scale reliability of $G2 = .85$ but has limited economic resources (i.e., does not have a large budget to hire multiple raters) may use the table to determine how to best apply resources while maintaining reliability. Using 5 raters and 6 items per sub-scale, the researcher could expect g coefficients between $G2 = .86$ and $G2 = .92$ across the three sub-scales. However, using only 3 raters and increasing items to 12 per subscale would still provide the researcher with g coefficients in the desired range ($G2 = .84$ to $G2 = .89$). It is important to note that the possibility of increasing items to 12 per scale is based on the assumption that suitable additional items could be generated. Generating additional items would require resources for item creation and pilot testing that could equal or be greater than those required to hire additional raters. Table 6 also assists the researcher in identifying the point of diminishing returns with respect to gains in generalizability. For example, on the Cognitive subscale, with 10 raters and 12 items ($G2 = .94$), increasing the number of items from 12 to 20 provides a gain of only .01 and is likely not worth the additional time required for the rating and data entry.

Finally, because both $G1$ and $G2$ are included in this table, researchers can examine the differences in g coefficients yielded by using a fully crossed research design. Because both $\text{var}(R)$ and $\text{var}(RI)$ are relatively modest for these three scales (see Table 4), the difference between $G1$ and $G2$ is relatively small, except for studies with a single rater per target ($n'_r=1$). As Table 6 indicates, crossing raters with targets provides a substantial gain in generalizability for these single-rater studies but may be less important for studies using two or more observers per research participant. It is interesting to note that if the RCS were used as a basis for teacher ratings or parent ratings (where $n'_r=1$ and persons are nested within raters), the expected g coefficient for scores would be quite modest ($G1 = .53$, .68, and .59 for Cognitive, Affective, and Physical scores, respectively). The combination of the nested

design with the availability of only a single rater for each participant can sharply curtail generalizability of rating data.⁶

Implications for choosing a research design for the RCS—Based on these findings, if using the current version of the RCS (6 Cognitive, 7 Affective, and 3 Physical items) a fully-crossed design with five raters ($n_r'=5$) yields expected g coefficients of $G2 = .86$ (Cognitive), $.92$ (Affective), and $.88$ (Physical). In an ideal world (i.e., the researcher has considerable control over both raters and ratee schedules), a research design involving five trained raters is more than adequate, relative to conventional standards of generalizability. From a pragmatic viewpoint, however, it may be desirable to overdesign rating studies, so that generalizability of scores is protected in the event of expectable or unexpected deviations from the research plan.

It is likely that in many research studies, even when trained observers serve as raters, a perfect, fully crossed rating design may not be possible for a number of reasons (e.g., some raters may not be able to be present to evaluate all ratees). This situation results in a hybrid design in which every person is rated by a subset of the available observers. It is not possible to estimate g coefficients for hybrid designs such as these directly. However, the expected generalizability of ratings for this type of hybrid design is likely to be greater than or equal to that for the fully-nested design and less than or equal to that of the fully crossed design. Thus, $G1$ and $G2$ represent lower and upper bounds on generalizability for a hybrid design with a given average number of raters per ratee (i.e., a given value of n_r'). Because such deviations from the planned rating design may result in some or even many ratees receiving scores from fewer than the full number of available raters, it may be wise to plan to train one or two more raters than are necessary to achieve the desired dependability of scores. Thus, researchers using the RCS may wish to plan for $n_r'=6$ or even 7, to safeguard dependability against scheduling mishaps.

A note on the magnitude of G coefficients—In the behavioral sciences it is generally considered sound psychometric practice to use measures with reliability estimates of $r_{XX} = .8$ or higher. Because investigators typically examine classical reliability coefficients, which usually overestimate the proportion of person (true score) variance by failing to consider multiple sources of error, the actual dependability of scores (when all relevant sources of error are taken into account) used in psychological research is undoubtedly somewhat lower than this ideal (Schmidt et al., 2003). Researchers using GT are advised to consider all relevant sources of error in the G study, so that g coefficients presumably do not suffer from this type of inflation. Thus, $G2$ derived from a two-facet G study is more conservative as an assessment of score dependability than the traditional (one-facet) reliability coefficient for the same measure.

THREE FACET G STUDY (PRIO)

Variance Partitioning

Researchers are often interested in variance in a child's behavior across contexts or occasions. For instance, to warrant a diagnosis of attention deficit hyperactivity disorder,

⁶The limitations of this rating design often go unnoticed by researchers using acquaintances (e.g., teachers, parents, or spouses) as sources of observer ratings. This lack of awareness is understandable. Although the generalizability of ratings in such a D study can be estimated from a fully crossed G study from a table of estimates such as Table 6, assuming that the variance partitioning in Table 4 is relevant to acquaintance raters, no such estimate can be made from the data of the D study itself, because P variance, R variance, and PR variance are completely confounded when $R:P$ and $n_r'=1$ (Hoyt, 2000). Because acquaintance ratings are quite popular in clinical research with children and adolescents, we return to this issue in the Discussion section.

researchers and clinicians have emphasized that the symptoms must be pervasive (i.e., occurring similarly over various contexts such as school and home). Stability over occasions is important for measures of constructs that are predicted to be stable over time, but for which scores may vary over testing occasions as a function of transient factors such as mood or recent life events. Using the GT framework, researchers can examine the variance that context or occasion contributes to behavior ratings, while still considering variance due to the person, rater, and item.

Table 7 illustrates the variance decomposition for the *PRIO* design. In this analysis, observed score variance is partitioned into variance attributable to main effects (i.e., *P*, *R*, *I*, and *O*), two-way interactions (i.e., *PR*, *PI*, *PO*, *RI*, *RO*, and *IO*), three-way interactions (i.e., *PRI*, *PRO*, *PIO*, and *RIO*), and the four-way interaction, which is confounded with error (*PRIO,e*).

Variance Estimates

We assessed the stability of RCS scores during a 5-month period among participants ($n_p = 36$) in the control or comparison group in our sample study (i.e., those who received no intervention). The three-facet (*PRIO*) analysis, as shown in Table 8, yields variance estimates for 15 components, making interpretation somewhat overwhelming at first glance. To facilitate interpretation of these results, we first consider the 7 components that are already familiar to readers from the *PRI* analysis. Comparisons of parallel components between Table 4 (*PRI*) and Table 8 (*PRIO*) clarify that components that have the same name and symbol in the two analyses actually have a different meaning in these two contexts. For example, *PR* variance in Table 8 refers to differential perceptions among raters of the same target person (i.e., person–rater interactions) that are stable over both items and occasions. This same component in Table 4 estimates person–rater interaction variance that is stable over items, but not necessarily over occasions. Thus, some of the variance in each component in Table 4 is stable over occasions (i.e., likely to replicate on a second rating occasion), but some is not, so that it is expected that variance estimates in Table 4, which include both temporally stable and temporally unstable variance, will be larger than the comparable variance estimates in Table 8 (which include only temporally stable variance).⁸

Comparison of Variance Estimates Included in Both *PRI* and *PRIO* Analyses

Accordingly, we find that *P* variance in Table 8 is smaller in magnitude, and accounts for a smaller proportion of total variance, than in Table 4. This is the case especially for Cognitive ratings (9% variance accounted for vs. 22% or 24% for Affective and Physical ratings). Although one should be cautious about overinterpreting findings from relatively small samples, this suggests that cognitive self-regulation may be relatively unstable over time (i.e., a person's levels of attention and concentration may vary more over occasions than his or her emotional and physical self-regulation). This finding bears further scrutiny in future investigations. As expected, for all three subscales, variance attributable to universe scores

⁸The formulas for *G1* and *G2* in the three-facet (*PRIO*) design are as follows:

$$G1 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_R^2 + \sigma_I^2 + \sigma_O^2 + \sigma_{PR}^2 + \sigma_{PI}^2 + \sigma_{PO}^2 + \sigma_{RI}^2 + \sigma_{RO}^2 + \sigma_{PRIO,e}^2} \quad (4)$$

$$G2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{PR}^2 + \sigma_{PI}^2 + \sigma_{PO}^2 + \sigma_{PRIO,e}^2} \quad (5)$$

in Table 8 (i.e., performance that is stable over *R*, *I*, and *O*) is smaller than the comparable variance estimates in Table 4 (where universe scores represent variance stable over *R* and *I* facets only).

R variance in Table 8 (as in Table 4) accounts for a substantial proportion of variance (12–17%). This hints at the possibility of systematic differences in rater leniency that are stable over *P*, *I*, and *O*. As noted earlier, these are not too serious when raters will be fully crossed with persons (i.e., a *PR* design in the D study). *R* variance becomes a concern when persons are nested within raters (*P:R*) or vice versa (*R:P*), as when using teacher or parent ratings. *I* variance in Table 8 is relatively small. There was little evidence of systematic mean differences among RCS items on a single occasion (Table 4), so the amount of *I* variance that is stable across occasions (Table 8) must also be small.

The proportion of *PR* variance in Table 8 is small relative to what was found in Table 4 (1–7% as compared with 12–17%). When variance estimates are sharply reduced in *PRIO* analyses relative to *PRI* analyses, this indicates that the corresponding effects are relatively unstable across replications (i.e., occasions). That is, for *PR*, the rater who gave person A the highest rating at Time 1 did not necessarily give him or her the highest rating again at Time 2.

PI and *RI* make only minor contributions to score variance for all three subscales. Of interest, *PRI* variance in Table 8 is also trivial in magnitude for all three subscales. This suggests that *PRI* variance in Table 4 consisted primarily of residual (error) variance—there is little stable *PRI* interaction variance when we replicate over occasions.

Variance estimates unique to PRIO analysis—If *O* main effect is significant it suggests higher scores on some occasions than others, averaged over persons, raters, and items. *O* variance in Table 8 is trivial in magnitude. *PO* variance reflects differences in the ordering of persons over occasions and directly reflects temporal instability of scores (i.e., perfectly stable universe scores over the 5-month period would yield zero *PO* variance). *PO* in Table 8 accounts for 6% to 11% of total variance, suggesting appreciable variability in performance over the two occasions. In other words, *P* variance in Table 4 is somewhat limited in terms of generalizability over time—at least over the relatively long time interval examined here. Because some instability in self-regulation ability is expected in children because of developmental changes, it would be useful to conduct G studies with closer intervals to determine how much of the *PO* variance detected here is attributable to differences in maturation rates (valid variance for most research purposes), and how much is due to occasion-specific variation in performance (error variance for most research purposes).

RO variance is relatively small (2–4% of total variance) for all three scales. Rater leniency appears to be relatively stable over time (i.e., raters who were most lenient at Time 1 tended to be among the most lenient at Time 2). *IO* variance is near zero for all subscales. This indicates that mean differences among items (which were relatively small in Table 4) were quite stable over testing occasions.

PRO accounts for a substantial proportion of variance (12–15%) in all three subscales. Thus, unlike *R* effects (which appear to be relatively stable over occasions), *PR* effects were largely occasion specific, suggesting that they were probably based on differential rater interpretations of occasion-specific behaviors rather than raters' idiosyncratic reactions to more stable attributes of target persons.

PIO and *RIO* account for relatively small proportions (4% or less) of score variance. For the Cognitive scale, especially, persons' uniquely high (or low) scores on particular items (*PI* component accounted for 2% of score variance) were relatively unstable over occasions (*PIO* component accounted for 4% of score variance). Thus there was little evidence of idiosyncratic performance on particular scale items that was replicable over testing occasions.

Finally, *PRIO,e* variance is substantial. A comparison to *PRI* variance in Table 8 reinforces the notion that the *PRI,e* component in Table 4 is mostly “*e*”. That is, there is no evidence of *PRI* interactions that replicate over occasions. Likewise, it is reasonable to conclude that *PRIO,e* in Table 8 is mostly error or “noise” in the rating procedure.

Tabulating Predicted G Coefficients for Various D Study Designs

To better illustrate the impact of temporal instability on generalizability of RCS scores over the three facets of *R*, *I*, and *O*, we used procedures⁸ analogous to those described for the *PRI* design (Table 5) to compute *g* coefficients for D study designs in Table 9. Table 9 demonstrates that adding raters, items, and occasions will increase the generalizability of the subscales and displays the changes in *g* coefficients (*G1* and *G2*) as the numbers of raters, items, and occasions are adjusted. The importance of assessing cognitive self-regulation (even more so than affective and physical self-regulation) over occasions is clear as the increases in *g* coefficients from 1 to 4 occasions illustrate. For example, if using a design with 10 raters and the RCS in its current form (7, 6, and 3 items for the Cognitive, Affective, and Physical subscales, respectively), a researcher could expect *g* coefficients of *G2* = .47 (Cognitive), *G2* = .83 (Affective), and *G2* = .76 (Physical). Increasing the number of testing occasions to four increases predicted *g* coefficients to *G2* = .71, *G2* = .90, and *G2* = .85 for the Cognitive, Affective, and Physical subscales, respectively. In summary, researchers interested in measuring behavior traits, either to predict long-term outcomes or to estimate stable levels, should use *PRIO* analyses to ensure an adequate number of observations to yield reliable scores.

We note in conclusion that the *PRIO* analysis, because it focuses on stability of scores over testing occasions, is similar to the examination of test–retest reliability. However, *g* coefficients are always lower than the corresponding classical reliability coefficient because *G* analyses consider multiple sources of error. To illustrate, the *g* coefficients for the RCS in a D study using the same design as our example study are .37, .71, and .58 for the Cognitive, Affective, and Physical subscales, respectively ($n'_r=5$, $n'_i=6$, 7, and 3, respectively). The same data yield conventional test–retest reliability estimates of .64, .84, and .80 for the Cognitive, Affective, and Physical subscales, respectively. This illustrates the overestimation of score dependability that results when all sources of error are not considered and the importance of taking all relevant sources of error into account, if we wish to understand the consequences of measurement error for research findings.

DISCUSSION

To study dependability of ratings, it is almost always desirable to consider multiple sources of error simultaneously and hence to use *G* analyses in preference to traditional reliability coefficients, which inevitably fail to consider all of the relevant sources of error (i.e., raters, items, and occasions; Hoyt & Melby, 1999). GT (Cronbach, et al., 1972) also provides an important conceptual framework for understanding the potential liabilities of measurement methods and for recognizing that weak correlations between ratings from different sources can have artifactual as well as substantive explanations. In this section, we consider the implications of the GT framework and of the present findings as well as those of other

generalizability studies (cf. Hoyt & Kerns, 1999) for some common measurement methods used in child and adolescent research: parent and teacher ratings, performance measures, and observer ratings.

GT Perspective on Common Measurement Procedures in Child and Adolescent Research

Among the most important lessons of GT for researchers are the need to consider what sources of error are relevant to different measurement methods and to take measurement error into account in interpreting effect sizes within and between studies. Hoyt, Warbasse, and Chu (2006) noted that self-reports are often more reliable than other methods. However, they may be less valid, because their reliability is attributable to stable response styles, as well as construct relevant variance. Thus, correlations between self-report measures (mono-method correlations) can be expected to be substantially larger than correlations between one self-report scale and a second set of scores derived from observer ratings. This can lead to effect size comparisons both within and between studies that are difficult to interpret at best, and misleading at worst (cf. Hunter & Schmidt, 1990, 1994). Thus, to level the playing field in comparing mono-method effect sizes to multimethod effect sizes, it is critical to (a) optimize measurement designs (ideally using G study findings) to minimize error variance to the extent possible or practical, (b) compute dependability coefficients that take relevant sources of error into account, and (c) disattenuate observed effect sizes to correct for the biasing effects of measurement error.

In addition, researchers need to be aware when practical constraints may disguise psychometric problems with measurement methods. This is often the case for acquaintance ratings, where the number of raters per target is typically small (usually $n_r'=1$) and raters are nested within targets (or vice versa). In this research design, error variance attributable to raters (and Rater $\times$ Target interactions) cannot be estimated in the D study and is, therefore, often overlooked as a source of measurement error. Thus, researchers using these measurement methods may overestimate the dependability of scores and be surprised by the disappointingly small effect sizes when these acquaintance ratings are correlated with scores on other measures. In this section, we consider lessons from GT for investigators using parent or teacher ratings (which are common in child and adolescent research) and approaches to using performance measures and ratings by trained observers (which are somewhat common and, in our view, could profitably be used even more frequently in this research area).

Implications for Parent Ratings

The strengths and weaknesses of parent ratings have been considered at length in the literature on child and adolescent assessment. The limitations noted include differential rater tolerance of behavior, differential interpretations of the scale (Elliot, Busse, & Gresham, 1993), impact of parental symptoms on perceptions (Patterson & Forgatch, 1995; Phares, Compas, & Howell, 1989), and other rater characteristics and bias (Patterson & Forgatch, 1995). However, parent ratings arguably have an important place in child and adolescent research as parents often have the most experience with the child. Other benefits include brief time demands, relatively easy administration and scoring, and the need for parents to describe behavior in naturalistic settings and to evaluate generalization of treatment effects in the home (Elliot et al., 1993).

From the standpoint of GT, however, the psychometric weakness of parent ratings goes beyond the specific sources of bias, which are inherent to any rating task. Bias becomes especially problematic in the case of parent ratings because the relationship of raters to target persons constrains the rating design. When parents provide ratings of children, $n_r'=1$ (or at most 2) for the rating design, and raters are nested within target persons ($R:P$). As the

estimated g coefficients in Tables 4 and 7 make clear, such a rating design can be highly problematic in terms of the expected generalizability of ratings, because aggregation over multiple raters is the primary means for minimizing the contribution of measurement error to score variance.

When considering the generalizability of RCS ratings over raters and items (*PRI*; Table 6), we found g coefficients for nested designs with $n_r=1$ of $GI = .53$, $.68$, and $.59$ for Cognitive (6 items), Affective (7 items), and Physical (3 items) subscales, respectively. The impact of this relatively weak dependability of scores on observed effect sizes is usually quantified in terms of *attenuation* (Hoyt, 2000; Schmidt & Hunter, 1996):

$$r_{XY} = \rho_{XY} \sqrt{r_{XX} r_{YY}} \quad (6)$$

where r_{XY} is the observed correlation, ρ_{XY} is the (hypothetical) correlation between error-free measures of constructs X and Y , and r_{XX} and r_{YY} are reliability (dependability) coefficients. Because $r_{XX} < 1$ and $r_{YY} < 1$, r_{XY} is expected to underestimate the true (population) correlation. The square root of the product of the reliabilities of X and Y ($\sqrt{r_{XX} r_{YY}}$) on the right side of equation determines the magnitude of attenuation. That is, it shows the degree to which the observed effect size underestimates the population correlation. To get an idea of impact on effect size, imagine correlating parent ratings of a child's cognitive self-regulatory ability with scores on an outcome measure with reliability $r_{YY} = .80$. Because GI is a more accurate coefficient of dependability than any single-facet reliability coefficient, we can substitute GI for r_{XX} to estimate the magnitude of attenuation: $\sqrt{GI \cdot r_{YY}} = \sqrt{(.57)(.80)} = .68$. Thus, if the population correlation is moderate (i.e., $\rho_{XY} = .30$), the observed correlation is expected to be smaller by a factor of $.68$ (i.e., $r_{XY} = .20$). Attenuation has adverse consequences for statistical power (Hoyt, 2000), and for our understanding of the importance of predictor variables (because it reduces the observed magnitude of validity correlations).

Although we used Table 6 values for this illustration, parents would not likely be asked to serve as raters on a performance task like the RCS. Instead, parents usually are asked to provide global impressions of their children or adolescents that are presumed to generalize over time (occasions) and over performance domains. As Hoyt (2000) pointed out, the dependability of such ratings cannot be estimated directly, because G studies with an arbitrarily large number of parents for each child cannot be conducted. However, using a meta-analysis of bias in observer ratings Hoyt and Kerns (1999) estimated that when untrained raters are asked to rate global attributes requiring inference, about 34% of the variance in ratings is due to rater bias.

Implications for Teacher Ratings

Strengths and weaknesses of teacher ratings are similar to those of parent ratings. Most studies collect ratings from a single teacher, so the expected dependability is rather modest, probably similar to that for parent ratings. However, it is often the case that ratings could be collected from multiple teachers for each child to enhance dependability of ratings, by reducing teacher-related bias components. In our example study (Lakes & Hoyt, 2004), the interteacher correlation for the total score on a child behavior scale was $.41$ ($p < .01$), indicating that scores derived from a single teacher's ratings would have modest reliability. For our analyses of program effectiveness, we aggregated the scores of the two teachers to create a single score for outcome analyses. Nonetheless, when 59% (i.e., $1 - r_{XX}$) of score variance for one teacher is attributable to error, the reliability of an aggregate rating from two teachers is only $r_{XX} = .58$, still a major source of attenuation in predictor-criterion correlations.

Congruence of Parent and Teacher Ratings

It is well documented (e.g., Achenbach, McConaughy, & Howell, 1987; Phares et al., 1989) that the correlations between teacher and parent ratings are modest (.26–.35), if these two ratings are considered as separate indicators of the same underlying construct. Achenbach et al. (1987) attributed the small associations between the two sets of ratings primarily to inconsistencies of the children's behavior between home and school. Our psychometric argument suggests that these inconsistencies are not as great as is implied by the observed correlations, which are severely attenuated because both sets of ratings contain substantial bias variance. To illustrate the effect of attenuation, consider Equation 6. Given that scores based on parent and teacher reports are both based on a single rater ($n_r=1$) in a nested design, Hoyt and Kerns's (1999) findings suggest that about 34% of score variance will be attributable to rater biases (σ_R^2 and σ_{PR}^2 combined). As a conservative estimate, we might expect another 20% of variance to be attributable to other error components, including the residual component. This would lead us to expect a g coefficient for each set of scores of no more than $(1 - .34 - .20) = .46$. The magnitude of attenuation when these two sets of scores are correlated will be $\sqrt{G1_x \cdot G1_y} = \sqrt{(.46)(.46)} = .46$. Given this severe level of attenuation, Achenbach et al.'s (1987) observed r_{XY} of .27 implies a disattenuated correlation of at least $\hat{\rho}_{XY} = .27/.46 = .59$. This population correlation between home and school behaviors suggests quite a bit more consistency in behavior and illustrates the potential of this type of GT analysis for clarifying existing findings as well as improving the quality of future child and adolescent research.

Implications for Performance Measures

Performance measures, such as computerized or paper-and-pencil tasks, are often used in child and adolescent research. Estimates of test–retest reliability vary per measure, but most indicate that a child's performance varies considerably across occasions. GT provides a useful tool for examining the variability of a child's performance on a task across occasions, and using G studies researchers could estimate how many performance occasions are needed to ensure adequate generalizability, given preferred or pragmatic levels of other facets, such as raters or items (tasks). It is not always the case that an identical performance measure can be administered multiple times because of the potential impact of practice effects. In such cases, it may be useful to examine performance across similar types of tasks or to explore adjustments to the numbers of items or length of time to increase generalizability.

Implications for Ratings by Trained Observers

As noted earlier, when multiple raters are used and all raters are crossed with targets, rater (R) variance does not contribute to error because the constant rater biases affect all targets equally. The R main effect accounted for a substantial proportion of variance in ratings in both *PRI* and *PRIO* analyses, a finding that appears typical from meta-analysis of the GT literature on observer ratings (Hoyt & Kerns, 1999). Thus, observer ratings have a significant advantage compared to acquaintance ratings, because sets of trained observers can be crossed with target persons to enhance generalizability of scores over raters. This is one noteworthy weakness of rating designs involving acquaintance ratings, as just discussed.

Brooks and Kutcher (2001; see also Frick & Loney, 2000; Kolko, 1993) argued that a critical consideration in the validity of observational ratings is their generalizability. Because trained observer reports are generally based on a relatively brief behavioral sample, it is important to move beyond basic questions of rater agreement, to consider generalizability of scores over time and across settings. GT provides a useful framework for conceptualizing and studying such questions. As Kenny and Zautra (1995) pointed out, single measures of any attribute contain both trait (stable or consistent) and state (unstable or

inconsistent) variance (see also Steyer, Ferring, & Schmidt, 1992). The three-facet (*PRIO*) analysis illustrated here gives useful information about the temporal stability not only of person's universe scores but also of error components of interest (such as *R* and *PR*). This analysis yields findings of theoretical interest, in terms of the stability of different dimensions of self-regulation, and findings of immediate practical interest to psychologists interested in developing dependable measures of these constructs for particular research applications.

Researchers have also been aware of the value of aggregating behavioral observations across situations to obtain more reliable (and valid) indicators of behavioral tendencies (Epstein, 1980; Rushton, Brainerd, & Pressley, 1983). More recently, Borkenau, Riemann, Angleitner, and Spinath (2001) developed an "assessment center" in which 300 pairs of adult twins participated separately in 15 diverse videotaped tasks (e.g., tell a joke to a confederate; build a stable paper tower in a short time). Trained observer ratings for this study both confirmed generalizability of personality traits across a broad spectrum of behavioral contexts and yielded heritability findings that contrast with those derived from parent- and self-ratings, raising questions about dependability and bias in these more typical rating modalities.

Thus, attention to generalizability across tasks or contexts is a fruitful area for research. Such studies may yield substantively useful information about similarities or distinctions between types of situations (Kelley et al., 2003) and about behavioral dimensions that are more or less variable within and between situations (Mischel & Shoda, 1995), as well as psychometrically useful information about generalizability of ratings. Although we did not do so in this illustrative G study, this question can be addressed within the GT framework by varying Setting (*S*) as a facet of future G studies, to examine replicability of universe scores (as well as of other effects of interest) across diverse behavioral contexts (see Lakes & Hoyt, 2008).

General Implications of GT for Improving Research Practices

To clarify how GT can inform design and interpretation of research in child and adolescent clinical psychology, we present a brief summary of a sample research study. We list four recommended practices then illustrate how each is implemented in this study.

Sample study—Researchers examine the impact of three family interventions on a child's behavior by randomly assigning caregivers of children with behavioral problems (based on parent and teacher ratings) to one of three treatments or a control condition. Researchers collect data during preintervention, midintervention, and postintervention visits.

Recommendation 1: Consider the generalizability of measurement methods and design a measurement strategy accordingly—The primary outcome measures in this study include a child behavior questionnaire, an observer-rated measure of child behavior, an observer-rated parent-child relationship assessment, and a parent-report questionnaire intended to evaluate relationship quality. The researchers will collect a standardized child behavior questionnaire from both parents and teachers (two teachers will complete the measure, and their ratings will be aggregated) to complement observer-rated measures of child behavior. As noted earlier, observer ratings may provide a dependable assessment based on a limited behavior sample, but their generalizability across occasions and settings (i.e., the *representativeness* of this behavior sample) is unclear. Although teacher and parent ratings may have limited generalizability across raters (because variance due to the unique biases of these acquaintance raters is likely to be substantial), they draw on

a much broader range of experience with the target person, so may be presumed to have greater generalizability over occasions and behavioral contexts.

The parent–child relationship also will be measured using two methods. Researchers will use an observer-rated measure of parent–child relationship quality, which will allow for systematic observation and use of multiple observers and will yield a dependable index of relationship quality. In addition, a self-report measure of parent–child relationship quality will provide data from the parent’s perspective that one cannot obtain through observation alone.

As this example illustrates, using multimethod, multimodal measurement procedures strengthens research designs (Shadish, Cook, & Campbell, 2002). From the perspective of GT, any single measurement method is likely to be generalizable over some facets (e.g., items, raters) but not others (e.g., settings, occasions). By combining multiple measures, using different methods, researchers can avoid falling prey to the limitations of particular measurement approaches and enhance confidence in the generalizability of their conclusions.

Recommendation 2: Quantify dependability of measurement using G coefficients rather than classical reliability coefficients—As discussed, classical (single facet) reliability coefficients almost always overestimate dependability of measurement (because they underestimate the proportion of error variance in observed scores). Generalizability coefficients provide a clearer indication of the strengths and weaknesses of different measurement methods and clarify interpretation of research findings.

Recommendation 3: G studies involving relevant error facets should be conducted prior to the D study—The illustrative study in this section is a D study, because the researchers will use the results to make decisions about treatment efficacy. For novel or complex measures, it is important to conduct pilot G studies to establish the psychometric properties of these measurement methods and to assist with D study design. Thus, prior to the start of the D study, researchers should conduct G studies on the observer measures to gain information to help plan for the main randomized clinical trial (D study) previously described (e.g., How many raters should be used? On how many occasions should observations be collected?).

Recommendation 4: Dissattenuate observed effect sizes to correct for biasing effects of measurement error—Probably the most controversial implication of the GT perspective on psychological measurement is the realization that it is not enough simply to report effect sizes (Wilkinson and the APA Task Force on Statistical Inference, 1999) to clarify the strength of association between variables in research studies. If we do not correct these effect sizes for distortion (usually attenuation) due to measurement error, we run the risk of misinterpreting our findings and misrepresenting them to others. Hoyt et al. (2006) noted that the potential for misinterpretation is acute in studies adopting multimethod measurement strategies, in that attenuation can be particularly severe in these studies.

Although some researchers may balk at reporting “corrected” effect sizes, these are actually more accurate representations of strength of association than the uncorrected versions, assuming accurate estimates of dependability of measurement, because they forecast the association that would be observed, if error-free measures of the constructs were available. It is important to note that, although correction for attenuation changes the magnitude of the effect size, it also changes the standard error and confidence interval proportionally, so that dissattenuated correlation coefficients for example are no more likely to be statistically

different from zero than the corresponding uncorrected correlations (Hunter & Schmidt, 1990). A sensible strategy, and one adopted in many meta-analyses, is to report both uncorrected and corrected values. The former may be useful for comparison with previous studies and the latter are more accurate representations of the strength of the relation between constructs.

Schmidt and Hunter (1996) provided a primer on the correction for attenuation, using the well-known formula provided here as Equation 6 to obtain an estimate of the “true” (unbiased) correlation, given an observed correlation (r_{XY}) and two reliability coefficients (r_{XX} and r_{YY}). Schmidt and Hunter emphasized the need to choose a reliability coefficient that takes into account the relevant sources of error, often the coefficient of stability and equivalence (Cronbach, 1947), in the case of self-report measures. Our suggestion here is that, for measures of other types, it is usually best to substitute a g coefficient for the reliability coefficient in this formula. The correction for attenuation is straightforward for correlation coefficients; for more complex analyses (e.g., multiple regression; structural equation modeling), a good approach is to use the attenuation-corrected correlation matrix (rather than the raw data) as input for the analysis.

As noted earlier, g coefficients often cannot be computed directly for acquaintance (e.g., parent) ratings. In the section on congruence of parent and teacher ratings, we estimate that 54% of variance in these scores may be attributable to various sources of error in acquaintance ratings (a g coefficient of .46). This figure is based on 34% rater variance (Hoyt & Kerns, 1999) plus 20% error due to other sources (which is conservative because reliability coefficients for a single source of error are often around .80). Thus, although no exact g coefficient is available due to constraints on conducting a G study with parent raters, to disattenuate effect sizes obtained using the parent rating scale, the researcher plans to use an estimated g coefficient of $GI = .46$. Because the illustrative study includes two teacher ratings per child, it is possible to compute the interrater reliability coefficient for this measure, using the SPSS Reliability procedure for the aggregate of two or more raters. This provides an empirically derived estimate of dependability for these scores. The resulting intraclass correlation coefficient would be used as a g coefficient for disattenuating effect sizes obtained with teacher ratings.

CONCLUSION

Measurement of psychological characteristics in children and adolescents is challenging. Common approaches include performance measures, child and adolescent self-report ratings, parent and teacher ratings, and observer ratings. GT elucidates strengths and weaknesses of each approach. A common weakness when ratings are provided by someone other than the child or adolescent is confounding due to rater biases, often compounded (in nested designs) by the inability to specify the nature or magnitude of these biases. Attenuation of observed effect sizes will usually be severe using these methods, unless multimodal measurement procedures are employed to minimize bias variance (Hoyt & McCullough, 2005).

Observer ratings have much to recommend them as an alternative (or supplement) to more common parent and teacher ratings, in that they offer researchers a variety of options to reduce the impact of rater bias. GT illustrates how crossing raters with targets reduces the effects of bias and offers guidance for optimizing observer measures. Observers generally have rather limited exposure to target behaviors on which to base their ratings, however, and more research is needed to understand the reliability-validity trade-off. Do observer ratings enhance reliability by reducing rater-based variance at the expense of validity (i.e., generalizability across settings/context)? The GT framework can be used to address this

important question, and generally offers a highly flexible system for addressing both psychometric and substantive questions of interest to researchers studying children and adolescents.

References

- Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*. 1987; 101:213–232. [PubMed: 3562706]
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
- Borkenau P, Riemann R, Angleitner A, Spinath FM. Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology*. 2001; 80:655–668. [PubMed: 11316228]
- Brennan, RL. Generalizability theory. New York: Springer; 2001.
- Brooks SJ, Kutcher S. Diagnosis and measurement of adolescent depression: A review of commonly utilized instruments. *Journal of Child and Adolescent Psychopharmacology*. 2001; 11:341–376. [PubMed: 11838819]
- Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, NJ: Erlbaum; 1988.
- Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3. Mahwah, NJ: Erlbaum; 2003.
- Conger AJ. A comparison of multi-attribute generalizability strategies. *Educational and Psychological Measurement*. 1981; 41:121–130.
- Crick, JE.; Brennan, RL. Computer Facilities. Dorchester, MA: University of Massachusetts; Boston: 1982. GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual). Available from http://www.education.uiowa.edu/casma/computer_programs.htm#genova
- Cronbach LJ. Test reliability: Its meaning and determination. *Psychometrika*. 1947; 12:1–16. [PubMed: 20293842]
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951; 16:297–334.
- Cronbach, LJ.; Gleser, GC.; Nanda, AN.; Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley; 1972.
- Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology*. 1963; 16:137–163.
- Elliot SN, Busse RT, Gresham FM. Behavioral assessment of children: School psychology. *School Psychology Review*. 1993; 22:313–321.
- Epstein S. The stability of behavior: II. Implications for psychological research. *American Psychologist*. 1980; 35:790–806.
- Frick PJ, Loney BR. The use of laboratory and performance-based measures in the assessment of children and adolescents with conduct disorders. *Journal of Clinical Child Psychology*. 2000; 29:540–564. [PubMed: 11126632]
- Gleser GC, Cronbach LJ, Rajaratnam N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*. 1965; 30:395–418. [PubMed: 5217607]
- 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, Kerns MD. Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*. 1999; 4:403–424.
- Hoyt, WT.; McCullough, ME. Issues in the multimodal measurement of forgiveness. In: Worthington, EL., Jr, editor. *Handbook of forgiveness*. New York: Brunner-Routledge; p. 109-123.(in press)
- 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 validation in counseling psychology research. *The Counseling Psychologist*. 2006; 34:769–805.

- Hunter, JE.; Schmidt, FL. *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage; 1990.
- Hunter, JE.; Schmidt, FL. Correcting for sources of artificial variation across studies. In: Cooper, H.; Hedges, LV., editors. *The handbook of research synthesis*. New York: Russell Sage; 1994. p. 323-336.
- Kelley, HH.; Holmes, JG.; Kerr, NL.; Reis, HT.; Rusbult, CE.; Van Lange, PAM. *An atlas of interpersonal situations*. Cambridge, UK: Cambridge University Press; 2003.
- Kenny, DA. *Interpersonal perception: A social relations analysis*. New York: Guilford; 1994.
- Kenny DA, Zautra A. The state-trait-error model for multiwave data. *Journal of Consulting and Clinical Psychology*. 1995; 63:52–59. [PubMed: 7896990]
- Kirk, RE. *Experimental design: Procedures for the behavioral sciences*. 2. Belmont, CA: Brooks/Cole; 1982.
- Kolko DJ. Further evaluation of inpatient child behavior ratings: Consistency across settings, time, and sources. *Journal of Emotional and Behavioral Disorders*. 1993; 1:251–259.
- Lakes KD, Hoyt WT. Promoting self-regulation through school-based martial arts training. *Journal of Applied Developmental Psychology*. 2004; 25:283–302.
- Lakes KD, Hoyt WT. What sources contribute to variance in observer ratings? Using generalizability theory to assess construct validity of psychological measures. *Infant and Child Development*. 2008; 17:269–284.
- Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*. 1993; 48:1181–1209. [PubMed: 8297057]
- Longford NT. Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*. 1994; 19:171–200.
- Mischel W, Shoda Y. A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*. 1995; 102:246–268. [PubMed: 7740090]
- Patterson GR, Forgatch MS. Predicting future clinical adjustment from treatment outcome and process variables. *Psychological Assessment*. 1995; 7:275–285.
- Phares V, Compas BE, Howell DC. Perspectives on child behavior problems: Comparisons of children's self-reports with parent and teacher reports. *Psychological Assessment*. 1989; 1:68–71.
- Rushton JP, Brainerd CJ, Pressley M. Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*. 1983; 94:18–38.
- Schmidt FL, Hunter JE. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*. 1996; 1:199–223.
- 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 difference constructs. *Psychological Methods*. 2003; 8:206–224. [PubMed: 12924815]
- Shadish, WR.; Cook, TD.; Campbell, DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin; 2002.
- Shavelson, RJ.; Webb, NM. *Generalizability Theory: A primer*. Newbury Park, CA: Sage; 1991.
- Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*. 1979; 86:420–428. [PubMed: 18839484]
- Smith PL. Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics*. 1978; 4:319–346.
- Stanley, JC. Reliability. In: Thorndike, RL., editor. *Educational measurement*. 2. Washington, DC: American Council on Education; 1971. p. 359-442.
- Steyer R, Ferring D, Schmitt MJ. States and traits in psychological assessment. *European Journal of Psychological Assessment*. 1992; 8:79–98.
- Wilkinson L. APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*. 1999; 54:594–604.

APPENDIX

GENOVA syntax for one-facet G study (raw data from Table 1):

```
STUDY Sample 1-facet for JCCAP paper
COMMENT
COMMENT # RECORDS = 5 (5 targets (persons), 3 raters)
COMMENT # VALUES PER RECORD = 3 (one for each rater)
COMMENT P = PERSON; R = RATER
COMMENT
OPTIONS RECORDS ALL
EFFECT * P 5 0
EFFECT + R 3 0
FORMAT (3F3.0)
PROCESS
27 31 25
15 19 21
15 23 22
7 12 10
18 20 11
FINISH
```

SAS syntax for one-facet G study (raw data from Table 1):

```
* SAS syntax for JCCAP sample varcomp;
data a;
input p r y @@;
cards;
1 1 27 1 2 31 1 3 25
2 1 15 2 2 19 2 3 21
3 1 15 3 2 23 3 3 22
4 1 7 4 2 12 4 3 10
5 1 18 5 2 20 5 3 11
;
proc varcomp method = mivque0;
class p r;
model y = p|r;
run;
```

	P1	P2	P3	P4	P5	P6
R1	X	X	X	X	X	X
R2	X	X	X	X	X	X
R3	X	X	X	X	X	X

FIGURE 1.
Fully crossed research design.

	P1	P2	P3	P4	P5	P6
R1	X	X				
R2			X	X		
R3					X	X

FIGURE 2.
Nested research design.

TABLE 1

Sample Data Set for 5 Targets and 3 Raters

	Rater 1	Rater 2	Rater 3	Target M
Target 1	27	31	25	27.67
Target 2	15	19	21	18.33
Target 3	15	23	22	20.00
Target 4	7	12	10	9.67
Target 5	18	20	11	16.33
Rater <i>M</i>	16.40	21.00	17.80	

TABLE 2
Two-Way Analysis of Variance and Variance Partitioning for Sample Data Set

Source	SS	df	MS	Variance Estimate	SE	%
<i>T</i>	507.0	4	127.0	39.03	24.40	75
<i>R</i>	55.6	2	27.8	3.63	4.02	7
<i>TR_e</i>	77.1	8	9.63	9.63	4.31	18
Total	640.0	14		52.29		

TABLE 3**Component Definitions for Two-Facet Generalizability Analysis (PRI)**

Persons (P)	Universe score for person p (deviation from grand mean, averaged over raters and items)
Raters (R)	Rater effect for rater r (rater leniency, averaged over persons and items)
Items (I)	Item effect for item i (deviation from grand mean, averaged over persons and raters)
PR	Idiosyncratic perception of person p by rater r (averaged over items)
PI	Idiosyncratic perception of person p on item i (averaged over raters)
RI	Idiosyncratic leniency of rater r on item i (averaged over persons)
$PRLe$	Idiosyncratic perception of person p by rater r on item i , confounded with random error

Note. Components contributing to each score X_{pri} , representing the rating for person p by rater r on item i of the Response to Challenge Scale. In this table, components are defined as *effects*—deviations from the grand mean for all persons, raters, and items. In GT, we are interested in the *variance* of each component, which provides information about the importance of the corresponding effect in determining observed variance in ratings.

TABLE 4

Raw and Percentage Variance Estimates (PRI)

Source of Variance	df	Variance Estimate	SE	% of Total Variance	p
RCS Cognitive					
<i>P</i>	180	0.3332	0.0405	32 .00**	
<i>R</i>	4	0.0619	0.0406	6 .11	
<i>I</i>	5	0.0020	0.0066	0 .40	
<i>PR</i>	720	0.1445	0.0114	14 .00**	
<i>PI</i>	900	0.0564	0.0068	5 .00**	
<i>RI</i>	20	0.0429	0.0136	4 .00**	
<i>PRLe</i>	3600	0.4152	0.0098	39 .00**	
RCS Affective					
<i>P</i>	180	0.6357	0.0723	46 .00**	
<i>R</i>	4	0.0376	0.0272	3 .19	
<i>I</i>	6	0.0619	0.0370	5 .13	
<i>PR</i>	720	0.1861	0.0125	14 .00**	
<i>PI</i>	1080	0.0421	0.0051	3 .00**	
<i>RI</i>	24	0.0561	0.0161	4 .01**	
<i>PRLe</i>	4320	0.3518	0.0076	26 .00**	
RCS Physical					
<i>P</i>	180	0.6630	0.0790	47 .00**	
<i>R</i>	4	0.0662	0.0444	5 .18	
<i>I</i>	2	0.0000	0.0024	0 .50	
<i>PR</i>	720	0.2651	0.0206	19 .00**	
<i>PI</i>	360	0.0375	0.0085	3 .00**	
<i>RI</i>	8	0.0246	0.0119	2 .09	
<i>PRLe</i>	1440	0.3543	0.0132	25 .00**	

Note. RCS = Response to Challenge Scale; *P* = Person; *R* = Rater; *I* = Item.

* $p < .05$.

$p < .01$
**

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

TABLE 5
Computing G1 and G2 for RCS-Affective Self-Regulation in a D Study With $n'_r = 5$, $n'_i = 6$

Source	Raw Variance	Multiplier [(P:R)] ^a	Adjusted Variance [(P:R)] ^a	Multiplier [PRI]	Adjusted Variance [PRI]
<i>P</i>	.6357	1	.6357	1	.6357
<i>R</i>	.0376	1/5	.0075	0	0
<i>I</i>	.0617	0	0	0	0
<i>PR</i>	.1861	1/5	.0372	1/5	.0372
<i>PI</i>	.0421	1/6	.0070	1/6	.0070
<i>RI</i>	.0561	1/30	.0019	0	0
<i>PRLe</i>	.3518	1/30	.0117	1/30	.0117
var (Tot)	1.3711		.701		.6916
var (<i>P</i>)/var (Tot)			.6357/.701 = .91		.6357/.6916 = .92

Note. Raw variance estimates are derived from G study (see Table 4); var = variance. G1 = estimate of var(*P*)/var(Tot) when the D study will use a (*P:R*) design; G2 = estimate of var(*P*)/var(Tot) when the D study will use a *PRI* design; Tot = total.

^aThe same multipliers, adjusted variances, and G1 values would apply for a D study design in which raters will be nested within persons [(*R:P*)]_j.

TABLE 6

Predicted G Coefficients for Various Possible D Study Designs (PRI)

n'_i	RCS Cognitive		RCS Affective		RCS Physical	
	G1	G2	G1	G2	G1	G2
$n'_r=1$						
1	.32	.34	.49	.50	.47	.49
3	.47	.51	.63	.65	.59	.62
6	.53	.59	.68	.71	.62	.67
12	.57	.64	.71	.74	.65	.69
20	.59	.66	.72	.75	.65	.70
$n'_r=2$						
1	.46	.48	.64	.65	.63	.65
3	.63	.67	.77	.78	.73	.76
6	.69	.74	.81	.83	.77	.80
12	.72	.78	.83	.85	.78	.81
20	.74	.79	.84	.86	.80	.82
$n'_r=3$						
1	.55	.56	.72	.73	.71	.73
3	.71	.74	.83	.84	.80	.82
6	.76	.80	.86	.87	.83	.85
12	.79	.84	.88	.89	.84	.87
20	.81	.85	.88	.90	.85	.87
$n'_r=5$						
1	.64	.65	.79	.80	.79	.80
3	.79	.81	.88	.89	.87	.88
6	.84	.86	.91	.92	.89	.90
12	.86	.89	.92	.93	.90	.91
20	.87	.90	.93	.94	.90	.92
$n'_r=8$						
1	.71	.72	.84	.85	.84	.85
3	.84	.86	.91	.92	.91	.92
6	.88	.90	.94	.94	.92	.93

n_i'	RCS Cognitive		RCS Affective		RCS Physical	
	G1	G2	G1	G2	G1	G2
12	.90	.92	.95	.95	.93	.94
20	.91	.93	.95	.96	.94	.95
1	.73	.74	.86	.86	.86	.87
$n_r'=10$						
3	.86	.87	.86	.87	.92	.93
6	.90	.91	.90	.91	.94	.94
12	.92	.94	.92	.94	.94	.95
20	.93	.95	.93	.95	.95	.96

Note. RCS = Response to Challenge Scale; $G1$ = predicted generalizability coefficient for a D study with the specified number of items and raters, using $(P;R)$ design; $G2$ = predicted generalizability coefficient for a D study with the specified number of items and raters, using PR design; n_i' =number of items, n_r' =number of raters.

TABLE 7

Component Definitions for Three-Facet Generalizability Analysis (PRIO)

Persons (P)	Universe score for person p (deviation from grand mean, averaged over raters, items, and occasions)
Raters (R)	Rater effect for rater r (rater leniency, averaged over persons, items, and occasions)
Items (I)	Item effect for item i (deviation from grand mean, averaged over persons, raters, and occasions)
Occasions (O)	Occasion effect for occasion o (deviation from grand mean, averaged over persons, raters, and items)
PR	Idiosyncratic perception of person p by rater r (averaged over items and occasions)
PI	Idiosyncratic perception of person p on item i (averaged over raters and occasions)
PO	Idiosyncratic perception of person p on occasion o (averaged over raters and items)
RI	Idiosyncratic leniency of rater r on item i (averaged over persons and occasions)
RO	Idiosyncratic leniency of rater r on occasion o (averaged over persons and items)
IO	Idiosyncratic effect for item i on occasion o (averaged over persons and raters)
PRI	Idiosyncratic perception of person p by rater r on item i (averaged over occasions)
PRO	Idiosyncratic perception of person p by rater r on occasion o (averaged over items)
PIO	Idiosyncratic perception of person p on item i on occasion o (averaged over raters)
RIO	Idiosyncratic leniency of rater r on item i on occasion o (averaged over persons)
$PRIO,e$	Idiosyncratic perception of person p by rater r on item i on occasion o , confounded with random error

Note. Components contributing to each score X_{prj} representing the rating for person p by rater r on item i on occasion o . As in Table 3, components are defined as *effects*—deviations from the grand mean for all persons, raters, and items. GT analyses estimate the *variance* of each component, which provides information about the importance of the corresponding effect in determining observed variance in ratings.

TABLE 8

Raw and Percentage Variance Estimates (PRIO)

	Cognitive					Affective					Physical					
	Var Src	df	Var Est	SE	% Var	p	df	Var Est	SE	% Var	p	df	Var Est	SE	% Var	p
<i>P</i>		35	0.1024	0.0491	9	.04*	35	0.3310	0.0995	22	.01**	35	0.3631	0.1217	24	.03*
<i>R</i>		4	0.1683	0.1176	14	.15	4	0.1805	0.1280	12	.20	4	0.2671	0.1746	17	.18
<i>I</i>		5	0.0224	0.0184	2	.18	6	0.0686	0.0400	5	.15	2	0.0000	0.0036	0	.50
<i>O</i>		1	0.0000	0.0073	0	.50	1	0.0000	0.0094	0	.50	1	0.0057	0.0166	0	.50
<i>PR</i>		140	0.0166	0.0201	1	.24	140	0.0783	0.0305	5	.04*	140	0.1141	0.0383	7	.03*
<i>PI</i>		175	0.0204	0.0125	2	.09	210	0.0170	0.0085	1	.09	70	0.0027	0.0103	0	.43
<i>PO</i>		35	0.1155	0.0393	10	.01**	35	0.0834	0.0335	6	.05*	35	0.1732	0.0573	11	.03*
<i>RI</i>		20	0.0250	0.0133	2	.06	24	0.0445	0.0189	3	.06	8	0.0308	0.0152	2	.11
<i>RO</i>		4	0.0479	0.0331	4	.14	4	0.0489	0.0351	3	.20	4	0.0339	0.0244	2	.21
<i>IO</i>		5	0.0037	0.0064	0	.32	6	0.0000	0.0026	0	.50	2	0.0004	0.0017	0	.45
<i>PRI</i>		700	0.0000	0.0171	0	.51	840	0.0296	0.0118	3	.01*	280	0.0	0.0187	0	.50
<i>PRO</i>		140	0.1442	0.0264	12	.00**	140	0.1524	0.0319	15	.00**	140	0.2168	0.0395	14	.00**
<i>PIO</i>		175	0.0473	0.0154	4	.01**	210	0.0213	0.0093	2	.01**	70	0.0160	0.0144	1	.24
<i>RIO</i>		20	0.0197	0.0098	2	.05*	24	0.0308	0.0108	2	.03*	8	0.0	0.003	0	.50
<i>PRIO,e</i>		700	0.4535	0.0242	38	.00**	840	0.2971	0.0145	20	.00**	280	0.3211	0.027	21	.00**

Note. Var src = source of variance; Var est = variance estimate; % var = percentage of total variance; *P* = Person; *R* = Rater; *I* = Item; *O* = Occasion.

* $p < .05$.

** $p < .01$.

TABLE 9

Predicted G Coefficients for Various Possible D Study Designs (PRIO)

n_r	n_i	n_o		RCS Cognitive		RCS Affective		RCS Physical	
				G1	G2	G1	G2	G1	G2
1	3	1	.13	.19	.39	.56	.28	.37	
		2	.19	.30	.45	.66	.36	.50	
		3	.22	.38	.47	.70	.39	.56	
		4	.24	.44	.49	.72	.41	.60	
	6	1	.15	.22	.42	.60	.30	.39	
		2	.21	.35	.47	.68	.37	.52	
		3	.24	.43	.49	.72	.40	.58	
		4	.26	.49	.51	.74	.42	.62	
	12	1	.16	.24	.43	.61	.30	.41	
		2	.22	.38	.41	.70	.37	.53	
		3	.25	.46	.50	.73	.40	.59	
		4	.27	.52	.52	.75	.42	.62	
5	3	1	.29	.34	.61	.67	.52	.58	
		2	.41	.49	.70	.78	.64	.71	
		3	.48	.59	.74	.83	.69	.78	
		4	.52	.64	.77	.85	.72	.81	
	6	1	.32	.37	.64	.71	.54	.59	
		2	.44	.53	.73	.81	.65	.72	
		3	.51	.62	.77	.85	.70	.78	
		4	.55	.68	.79	.88	.73	.82	
	12	1	.33	.39	.67	.74	.54	.60	
		2	.46	.55	.75	.83	.65	.73	
		3	.52	.64	.78	.87	.71	.79	
		4	.56	.70	.80	.89	.73	.82	
10	3	1	.35	.38	.65	.69	.59	.62	
		2	.48	.54	.76	.80	.71	.76	
		3	.56	.62	.80	.85	.77	.82	
		4	.60	.68	.83	.87	.80	.85	

n_r	n_i	n_o	RCS Cognitive		RCS Affective		RCS Physical	
			G1	G2	G1	G2	G1	G2
6	1		.37	.41	.69	.73	.60	.63
	2		.51	.57	.79	.83	.72	.76
	3		.59	.66	.83	.87	.77	.82
	4		.63	.71	.85	.90	.80	.85
12	1		.39	.42	.72	.75	.60	.63
	2		.53	.59	.81	.85	.72	.77
	3		.60	.68	.84	.89	.77	.82
	4		.65	.73	.86	.91	.80	.86

Note. RCS = Response to Challenge Scale; G1 = predicted generalizability coefficient for a D study with the specified number of items and raters, using *PRIO* design; n_i =number of items, n_r =number of raters, n_o =number of occasions, coefficient for a D study with the specified number of items and raters, using *PRIO* design; n_i =number of items, n_r =number of raters, n_o =number of occasions.