Title
Restricted sample variance reduces generalizability.

Journal
Psychological Assessment, 25(2)

ISSN
1939-134X 1040-3590

Author
Lakes, Kimberley D

Publication Date
2013

DOI
10.1037/a0030912

Published in final edited form as:
Psychol Assess. 2013 June ; 25(2): 643–650. doi:10.1037/a0030912.

Restricted Sample Variance Reduces Generalizability

Kimberley D. Lakes
University of California, Irvine

Abstract

One factor that affects the reliability of observed scores is restriction of range on the construct measured for a particular group of study participants. This study illustrates how researchers can use generalizability theory to evaluate the impact of restriction of range in particular sample characteristics on the generalizability of test scores and to estimate how changes in measurement design could improve the generalizability of the test scores. An observer-rated measure of child self-regulation (Response to Challenge Scale; Lakes, 2011) is used to examine scores for 198 children (Grades K through 5) within the generalizability theory (GT) framework. The generalizability of ratings within relatively developmentally homogeneous samples is examined and illustrates the effect of reduced variance among ratees on generalizability. Forecasts for *g* coefficients of various *D* study designs demonstrate how higher generalizability could be achieved by increasing the number of raters or items. In summary, the research presented illustrates the importance of and procedures for evaluating the generalizability of a set of scores in a particular research context.

Keywords

generalizability theory; reliability; restriction of range; sample characteristics; restricted variance

It is common for researchers to select a psychological assessment measure based on prior evidence of reliability and validity of scores and, seemingly, to assume that the measure will, therefore, produce reliable and valid scores for their study (Vacha-Haase, Kogan, & Thompson, 2000; Yin & Fan, 2000). Although prior evidence is important and should be considered in the selection of research measures, it should not be interpreted as a guarantee that a measure will produce reliable and valid scores for participants in a particular study. As much as is possible given research design constraints, researchers should examine and report on the reliability and validity of scores obtained in their research; however, this is not yet common practice (most methods sections refer to prior estimates of reliability and validity rather than to estimates obtained in their research sample; for a review and meta-analysis in which this trend is empirically documented, see Yin & Fan, 2000). Yin and Fan stated:

> It is important to emphasize that reliability is a characteristic related to "a set of scores," not an inherent characteristic of a test or an instrument. Despite the popular and yet misleading use of "test reliability" both in the measurement literature and in some less formal settings, as if reliability were an inherent characteristic of an instrument itself, many authors, and researchers (e.g., Crocker & Algina, 1986, p. 144; Linn & Gronlund, 1995, p. 82; Pedhazur & Schmelkin, 1991, p. 82; Traub,

Correspondence concerning this article should be addressed to Kimberley D. Lakes, Division of Development and Behavior, Department of Pediatrics, 101 Academy Way, Suite 150, University of California, Irvine, Irvine, CA 92617. klakes@uci.edu.

Supplemental materials: http://dx.doi.org/10.1037/a0030912.supp

> 1994) have emphatically pointed out that it is the scores from a test administration, not the test itself, that we are concerned about when we discuss measurement reliability. (p. 203)

Many factors, including sample characteristics, impact the reliability of scores (Cohen & Cohen, 1983; Henson, Kogan, & Vacha-Haase, 2001). When a measure of a psychological construct is administered within different samples, the reliability of scores will be affected by sample characteristics, such as restriction of range on the construct measured (Thompson, 1994). For example, if a researcher administers a measure of self-regulation in a sample of children ages 5 through 13, scores may yield different reliability estimates than if the same instrument were administered only with 12- and 13-year-olds. Restriction of range (in this case, most likely due to developmental similarities on the particular items in the 12- and 13-year-old sample) and corresponding problems related to reliability or generalizability of test scores are common and understudied problems in psychological testing (Thorndike, 1949; Wiberg & Sundstrom, 2009). Restricted range in homogeneous samples is likely to lower estimates of reliability (Henson et al., 2001), and such reductions in reliability can have an adverse impact on research findings. Observed relationships between variables may be smaller and may not be significant (contributing to Type II errors) as a result of this artifact of measurement (see Lakes & Hoyt, 2009, for examples and a discussion of this issue). Therefore, it is important for researchers to consider the dependability of test scores within the sample selected for their research and to understand that reliability estimates will be affected by characteristics of their sample. As Yin and Fan (2000) noted, “The dependence of reliability on the observed score variance and sample characteristics may cause the reliability of scores in an application to deviate considerably from those reported in the original validation studies (e.g., those reported in the manual)” (p. 209).

Although it is increasingly common practice for researchers to conduct power analyses to determine appropriate sample sizes for their research, less attention has been paid to the importance of planning measurement designs to optimize the reliability of the set of scores to be obtained in the research. Attending to the reliability of scores is arguably as important as the size of the sample and should be addressed in study proposals and in reports of findings. The purpose of this report is to describe how restricted range affects the reliability of scores and to illustrate how a generalizability theory decision (*D*) study can be used to plan for measurement designs in which sample characteristics such as restricted variance may become an issue. Using an observer-rated measure of children’s self-regulation (Response to Challenge Scale, or RCS; Lakes, 2011), I report the results from generalizability analyses with homogeneous groups (in terms of grade level). Next, forecasts for various *D* study designs are presented, providing estimates of dependability in situations where it may be feasible to use different numbers of raters or items to increase estimates of dependability and offset the reduction associated with restricted variance in the sample.

Method

Participants for the data set used in this study have been described previously (Lakes & Hoyt, 2004) and will be only briefly described here. Participants were 198 students (Grades K through 5) at a private lower school in the midwestern United States. All 208 children in the school were invited to participate in a research study that was approved by the Institutional Review Board of the University of Wisconsin, Madison. Parental written consent was obtained for 207 children. Eighty-three percent were White, 73% were from upper middle class and affluent families, and 51.5% were girls.

Eight raters (seven psychology upper level undergraduate and graduate students and one postgraduate professional) were trained by the author to rate children in a physical context. The 1-hr training session included a review of subscale items and their definitions, examples

of performances that would fit different points on the scale, and the opportunity to practice and compare ratings. Raters received clipboards and sat in the school gymnasium, separated so that they could not compare their ratings. They were told to refrain from comment and interaction with children. A physical education instructor, who was not a rater, created a physical challenge course that was the same for all of the children in a given grade level; difficulty was increased incrementally for higher grades. While the challenge course followed the same steps and used the same targets, targets were adjusted (i.e., raised higher) to increase difficulty. Raters were instructed to anchor their ratings on the first child of each grade level, so that they would compare students to their grade-level peers. Children were rated during the first week of the academic year.

The Response to Challenge Scale (RCS; Lakes, 2011; Lakes & Hoyt, 2004, 2009) is an observer-rated measure of children's self-regulation in three domains (affective, cognitive, motor) and includes 16 items, each consisting of two bipolar adjectives rated on 7-point scales (e.g., 1 = *inattentive*, 7 = *attentive*). Some items were reversed (i.e., the response intended to measure low regulation was a 7 rather than a 1) to discourage raters from globally high or low responding, but these scores were reversed prior to analyses so that in the database higher scores indicate greater self-regulation. As part of a multimethod, multidimensional assessment, students completed a challenging physical obstacle course and were rated on the RCS by research-trained observers, whose scores were aggregated to yield single scores for each item. Subscale scores were obtained for each domain by calculating the average score across the items for that subscale (six Cognitive, seven Affective, and three Motor items).

The RCS has been discussed in three prior published studies: an intervention study where it proved to be a useful tool for measuring self-regulation outcomes (Lakes & Hoyt, 2004); a methodological primer on generalizability theory (GT), where the authors analyzed RCS data to illustrate the utility of generalizability theory (Lakes & Hoyt, 2009); and an instrument development study (Lakes, 2011). Lakes (2011) described the rationale for and content of the RCS and its three subscales (Cognitive, Affective, and Motor Regulation) and reported results that indicated that in an elementary school sample, ratings of self-regulation in different domains (cognitive, affective, motor) were strongly related but distinguishable. Lakes and Hoyt (2009) reported results from a two-facet generalizability analysis in which participants (*persons*) were crossed with *raters* and *items* (*PRI*), treating raters and items as the primary sources of error. With a fully crossed measurement design and five raters, *PRI* generalizability coefficients (a g coefficient is interpreted like a reliability coefficient, but takes multiple sources of error into account) were $g = .89$ (Cognitive Regulation), $g = .91$ (Affective Regulation), and $g = .89$ (Motor Regulation). Lakes and Hoyt (2009) also examined generalizability over two occasions 4 months apart (i.e., which is similar to test–retest reliability but different in that g coefficients consider multiple sources of error simultaneously and are always lower than estimates based on a single source of error) as well as items and raters in a three-facet (*PRIO*) analysis using study control group participants only. Test–retest reliabilities using conventional reliability analyses were .64, .84, and .80 for scores obtained from the Cognitive, Affective, and Motor scales, respectively; the g coefficients were .37, .71, and .58 for scores obtained from the Cognitive, Affective, and Motor subscales, respectively. Although the test–retest reliability coefficients, measured using standard methods and compared with general standards for adequate test–retest reliability, were sufficient (.84 and .80) for Affective and Physical scale scores, the coefficient for Cognitive scale scores (.64) was lower, which could suggest that this domain may be more difficult to rate in the context used or may just be the result of differential maturation on this factor over the course of the school year (pretesting was conducted the first week of the academic year, and posttesting was conducted 4 months later). Thus, the evidence to date indicates that that the RCS is a measure of self-regulation that in an

elementary school sample, has produced scores that are consistent with self-regulation theory (Lakes, 2011), responsive to intervention change (Lakes & Hoyt, 2004), and adequately reliable (Lakes & Hoyt, 2009).

The analyses conducted for this research were dependability or generalizability analyses (see Brennan, 2001; Shavelson & Webb, 1991). Generalizability analyses examine sources of error in ratings and provide estimates of generalizability of scores for future studies (*decision* or *D studies*) using varying numbers of raters and items. This study reports results from multiple two-facet analyses (*PRI* for homogeneous groups) of subgroups of participants who were the targets of measurement (grouped by grade level), treating raters and items as the primary sources of error (see Table 5 in the online supplemental material for a description of the variance components in *PRI* analyses). Analyses for this report were conducted on a fully crossed subsample, where all raters rated all participants (i.e., if one participant was absent during an assessment or was not rated by all raters for some reason, this participant's data were excluded). The fully crossed subsample consisted of 181 participants with five raters. Generalizability coefficients[1] were used as the RCS scores are to be interpreted relative to group performance.

These analyses provide the same type of information yielded by the RCS *G* analyses in Lakes and Hoyt (2009), but rather than providing dependability estimates for an elementary school sample, they provide researchers with generalizability estimates for more developmentally homogeneous samples. For example, a researcher working with a sample of one age or grade level (e.g., first grade) would be likely to see restricted *P* (person) variance, which should affect the dependability of the scores obtained. The results of these analyses provide estimates of generalizability in three subgroups (e.g., kindergarteners and first graders, second and third graders, and fourth and fifth graders); analyses were conducted on three subgroups rather than individual grade levels to maintain an adequate and approximately equivalent number of participants per subgroup. All analyses were conducted using the program GENOVA (generalized analysis of variance; Crick & Brennan, 1982).

Results

G Studies: Generalizability Analyses (PRI)

Variance estimates for persons (*P*) represent variance in ratings that is attributed to individual differences on the construct measured. The restricted variance in homogeneous groups had a greater effect on the *P* variance components for the Cognitive subscale scores (Table 1) than on variance components for the Affective and Motor subscale scores (Table 2 and Table 3, respectively). *P* variances for the Cognitive subscale scores were 24%–26% with the homogeneous samples and 32% with the full sample. *P* variances for the Affective subscale scores were 37%–41% with the homogeneous samples, compared with 46% with the full sample, and *P* variances for the Motor subscale scores were 39%–51% compared with 47% in the full sample. As the proportion of variance that is attributed to persons decreases, the *g* coefficient decreases as well.

[1]In generalizability theory, two types of coefficients are common. Brennan (2001) described a *g* coefficient (generalizability coefficient) as "the ratio of universe score variance to itself plus relative error variance" and a phi coefficient (dependability coefficient) as "the ratio of universe score variance to itself plus absolute error variance." Phi coefficients are generally slightly lower than *g* coefficients. Brennan recommended the use of phi coefficients when scores have "absolute interpretations, as in domain-referenced or criterion-referenced situations" (p. 13). Because the RCS was not designed to be domain- or criterion-referenced (the person's score is not interpreted in isolation, but in reference to group performance) this report focuses on *g* coefficients, rather than phi coefficients.

Variance estimates for raters (*R*) represent between-rater variance (e.g., systematic differences in rater severity or leniency) for all raters over persons and items. *R* variance was small (ranging from 3% to 8%) for all three subscales and was similar for heterogeneous and homogeneous samples. *R* variance was lowest for the Affective subscale, which indicates that raters interpreted items on this subscale more consistently.

Variance estimates for items (*I*) represent between-item variance, averaged over persons and raters. For the Cognitive and Motor subscales, *I* variance was negligible in all analyses (ranging from 0% to 1%). *I* variance was small for the Affective scale (5%–6%) and was similar across samples. Higher *I* variance for the Affective scale suggests that there are greater mean differences among items on this scale than on the other two scales.

Variance estimates for Person × Rater interactions (*PR*) represent differences between raters in their rank ordering of persons. For the Cognitive subscale, in the full heterogeneous group, *PR* was 14%; in homogenous groups, *PR* ranged from 1% (for ratings of fourth and fifth graders) to 18% (for ratings of kindergarteners and first-grade students). This indicates that there was greater agreement between raters on the rank order of older students. For the Affective subscale, in the heterogeneous groups, *PR* was 14%; in homogeneous groups, *PR* ranged from 12% to 18%. On the Motor subscale, *PR* was 19% in the heterogeneous sample and ranged from 12% (fourth- and fifth-grade sample) to 27% (kindergarten and first-grade sample). Substantial *PR* variance suggests that raters differ in their criteria for assigning high and low scores; in this case, raters were somewhat more idiosyncratic in evaluating self-regulation in younger children.

Inconsistencies in ordering of persons from one item to another, averaged over raters, is represented by variance estimates for Person × Item (*PI*) interactions. As Lakes and Hoyt (2009) noted, “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)” (p. 151). *PI* variance was minimal, ranging from 1% to 7% across samples and subscales. *PI* was highest for the Cognitive subscale, indicating that interitem consistency is somewhat lower on this subscale, which contains items designed to measure attentiveness, engagement in tasks, and ability to ignore distractions.

Substantial *RI* variance suggests that rater leniency varies across items. *RI* variance was minimal (1%–8%) and was fairly similar in the heterogeneous and homogeneous samples. *RI* variance was slightly larger for homogeneous samples (second- and third-grade sample as well as the fourth- and fifth-grade sample) for the Cognitive subscale, suggesting that among these students, rater leniency varied more across Cognitive control items.

Lakes and Hoyt (2009) described the final variance component (*PRI,e*) as “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)” (p. 151). *PRI,e* variance was highest for the Cognitive subscale and ranged from 39% to 49%, with higher estimates for analyses conducted with homogeneous samples. For the Affective subscale, *PRI,e* was 26% in the heterogeneous sample and ranged from 28% (second- and third-grade sample) to 41% (fourth- and fifth-grade sample). *PRI,e* variances were similar for the heterogeneous and homogeneous samples for the Motor scale, with a slightly higher estimate for the fourth- and fifth-grade sample (29% vs. 25% for the heterogeneous sample).

Predictions of G Coefficients for Various D Study Designs

Results from the *PRI* analyses indicate that researchers using the scale with populations that might have restricted variance may need to consider adjusting the numbers of raters or items

to increase the expected *g* coefficients for a particular study. To provide an example of how these adaptations may be made, I computed *g* coefficients for a variety of possible *D* study designs using the variance estimates from the homogenous groups. Table 4 provides estimates of expected *g* coefficients given a certain number of raters and items for use with homogenous populations, and Table 6 in the online supplement shows the estimated variance components for a subset of the predictions in Table 4. Figure 1 illustrates the trends for expected generalizability coefficients (for Cognitive Regulation scores, Grades K–1) as numbers of items and raters are increased or decreased, using results in Table 4. As the figure illustrates, gains level off at a certain point, indicating that further modifications are unlikely to increase generalizability of scores.

Discussion

The RCS was designed to measure cognitive, affective, and motor regulation as a child engages in a series of challenging tasks. This article reported analyses of scores using the generalizabiltiy theory (GT) framework and provided estimates of generalizability over raters and items in developmentally similar groups of children. Based on their *G* analyses of data from the full sample, Lakes and Hoyt (2009) recommended that with a large relatively heterogeneous sample (e.g., an entire elementary school), researchers should use the RCS in a fully crossed measurement design (where rater variance does not contribute to relative error variance because all raters rate all participants) with five raters and the scale's current format (16 items). Expected generalizability coefficients for this design are .89 for Cognitive, .91 for Affective, and .89 for Motor. In the present study, the generalizability of the RCS with developmentally homogeneous populations was analyzed to illustrate the effect of reduced variance among ratees on the generalizability of the subscale scores. In the same rating design, expected range for *g* coefficients for homogeneous populations would be between .82 and .87 for Cognitive, between .86 and .90 for Affective, and between .85 and .90 for Motor, with the slightly lower estimates expected for younger children. These *g* coefficients reach the acceptable level; however, in situations where a different measurement design is needed, forecasts for different *D* study designs demonstrated how increases in items or raters could lead to an increase in expected *g* coefficients; for example, increasing the number of raters from five to 10 would result in an increase in expected *g* coefficients of between .88 and .91 for Cognitive, between .92 and .94 (Affective), and between .91 and .93 (Physical) for the younger and older groups, respectively.

In addition to demonstrating the decrease in *g* coefficients with restricted person variance, Table 4 demonstrates a trend across subscales and subgroups. Raters using the RCS showed greater consensus (i.e., higher generalizability) in their ratings of older children than younger children across all three subscales. This provides important information for researchers. Assessing self-regulatory skills in children ages 5–6 may be more difficult than assessing self-regulation in children older than age 8, at least in the context used for this study. For a homogeneous population of fourth and fifth graders, most coefficients were only minimally different or remained the same as those obtained with the full elementary school sample. However, for a sample of kindergarten and first graders, the coefficients were slightly lower, suggesting that ratings for developmentally homogeneous groups of younger children may have lower *g* coefficients in the recommended design, and researchers would be prudent to increase items or raters to increase the expected generalizability. Increasing the number of items or perhaps increasing the length of time in which the children are observed may also positively affect the *g* coefficients of RCS scores for homogeneous groups of younger children.

Limitations

There are other factors that likely restricted the variance within the sample, including similarities between participants in characteristics such as socioeconomic status, race, and geographic location. The impact of these restrictions on generalizability of scores was not possible to measure in the current study. The analyses conducted for this research suggest that restriction of range on participant characteristics can reduce the generalizability of scores obtained. However, in this study, restriction of range was studied by grouping participants into categories based on similarities in age, and self-regulation is known to vary over developmental stages, with children developing greater self-regulation as they mature. Because the variable studied (self-regulation) is expected to correlate with age, use of the scale in samples homogeneous in age will restrict the range of scores and attenuate reliability (and generalizability) coefficients.

Moreover, it is possible that subscales with highly correlated, narrowly defined items could have higher estimates of generalizability simply because of the relationship between items. This is a possibility with the Motor Regulation subscale, where items measuring motor coordination and athleticism are expected to highly correlate with one another. It is possible that an overly narrow range of items on a scale could increase generalizability, while decreasing validity. Therefore, in addition to studying the dependability of scores, researchers should simultaneously attend to the validity of scores for the purposes of their research. It is also important to note that the dependability of estimates forecasted for various *D* study designs is limited by the original measurement design. For example, when estimating dependability coefficients for item increases for a scale that in the example data set had three items, researchers should be aware that the *D* study produces only estimates and that the proximity of obtained generalizability coefficients to those that were predicted will depend on a number of factors, including the relationships between items.

Future research could examine the impact of variance restriction in *G* studies with additional facets, such as occasion. In a heterogeneous sample (Grades K–5), Lakes and Hoyt (2009) reported results from *PRIO* analyses, examining occasion variance in addition to variance due to persons, raters, and items. *PRIO* analyses could be repeated with homogeneous groups. Moreover, though it was beyond the scope of the current report, as the RCS has several subscales, a multivariate generalizability study could also be used in this research, allowing for the examination of variance as well as covariance.

The Impact of Restricted Variance in Test Scores: Implications for Researchers

In his discussion of the effect of group heterogeneity on the reliability of test scores, Gulliksen (1950) noted, “If a test administered to two groups is the same and the same standard conditions are observed, it is highly unlikely that error variance is affected” (p. 108). He concluded that the effect of group heterogeneity on score reliability (higher reliability coefficients) is attributed to differences in true variance rather than differences in error variance. Results of the GT analyses presented in this report confirmed this—in relatively developmentally homogenous samples, *P* variance was smaller, and, therefore, *g* coefficients were smaller as well. Thus, the reduced generalizability coefficients observed in this study do not suggest that there is greater error in scores when the measure is administered in groups that are homogeneous in some characteristic; rather, it suggests that there is less true variance among persons, as measured by the rating scale.

While observed restrictions in variance do not inherently change a scale, they do affect the generalizability of a particular set of scores and impact the success of the scale in accomplishing the research aims in a particular sample. As demonstrated in the present study, restricted variance in a sample may produce smaller *g* coefficients. Weaker

generalizability of scores for a particular sample in turn will affect important statistics, such as coefficients of validity and estimates of intervention change. As noted previously, poor dependability of scores can contribute to Type II errors (Lakes & Hoyt, 2009). Therefore, when administering a measure within a group that is relatively homogeneous in regards to a particular characteristic, researchers would be well advised to consider the impact of restricted range on scores and to make adjustments to the measurement design prior to implementing the study in order to maximize generalizability. This report illustrated how GT can be applied to address this issue. One of the major advantages of GT is the potential to forecast *g* coefficients for alternative *D* study designs. This procedure demonstrates multiple ways to increase generalizability; for example, when using a fixed rating scale that would not allow for the addition of new items, researchers could consider increasing the number of raters instead of items. As Figure 1 and Table 4 demonstrate, at a certain point, gains in *g* coefficients level off—this provides valuable information for researchers as it indicates the point at which additional raters and items, which both require resources to generate, would yield little, if any, improvement in generalizability of scores. Additional facets can also be considered in GT: Lakes and Hoyt (2009) illustrated the utility of three-facet *G* (*PRIO—p*erson, *r*ater, *i*tem, *o*ccasion—analyses) and *D* studies, showing how raters, items, or additional occasions could be used to increase generalizability. GT provides researchers with the tools to weigh the feasibility, costs, and benefits of varying items, raters, and occasions, in order to construct a strong measurement design that is both feasible and cost-effective.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author acknowledges funding during the preparation of this article from the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health (NIH), for the NIH Clinical and Translational Science Awards at UCI (Grant UL1 TR000153). The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. This research is based in part on the author's doctoral dissertation, which was chaired by William T. Hoyt, University of Wisconsin, Madison. The author thanks Dr. Hoyt for helpful comments on drafts of this article.

References

Brennan, RL. Generalizability theory. New York, NY: Springer; 2001.

Cohen, J.; Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1983.

Crick, JE.; Brennan, RL. GENOVA: A generalized analysis of variance system [FORTRAN IV computer program and manual]. Dorchester: University of Massachusetts, Boston, Computer Facilities; 1982.

Crocker, L.; Algina, J. Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich; 1986.

Gulliksen, H. Theory of mental tests. Hoboken, NJ: Wiley; 1950. Effect of group heterogeneity on test reliability; p. 108-127.

Henson RK, Kogan LR, Vacha-Haase T. A reliability generalization study of the Teacher Efficacy Scale and related instru-ments. Educational and Psychological Measurement. 2001; 61:404–420.

Lakes KD. The Response to Challenge Scale (RCS): The development and construct validity of an observer-rated measure of children's self-regulation. The International Journal of Educational and Psychological Assessment. 2011; 10:83–96.

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. Applications of generalizability theory to clinical child and adolescent psychology research. Journal of Clinical Child and Adolescent Psychology. 2009; 38:144–165. [PubMed: 19130364]

Linn, RL.; Gronlund, NE. Measurement and assessment in teaching. 7th ed. Upper Saddle River, NJ: Merrill; 1995.

Pedhazur, EJ.; Schmelkin, LP. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum; 1991.

Shavelson, RJ.; Webb, NM. Generalizability theory: A primer. Newbury Park, CA: Sage; 1991.

Thompson B. Guidelines for authors. Educational and Psychological Measurement. 1994; 54:837–847.

Thorndike, RL. Personnel selection: Test and measurement techniques. New York, NY: Wiley; 1949.

Traub, RE. Reliability for the social sciences: Theory and applications. Thousand Oaks, CA: Sage; 1994.

Vacha-Haase T, Kogan LR, Thompson B. Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. Educational and Psychological Measurement. 2000; 60:509–522.

Wiberg M, Sundstrom A. A comparison of two approaches to correction of restriction of range in correlation analysis. Practical Assessment, Research, and Evaluation. 2009; 14:1–9.

Yin P, Fan X. Assessing the reliability of Beck Depression Inventory Scales: Reliability generalization across studies. Educational and Psychological Measurement. 2000; 60:201–223.

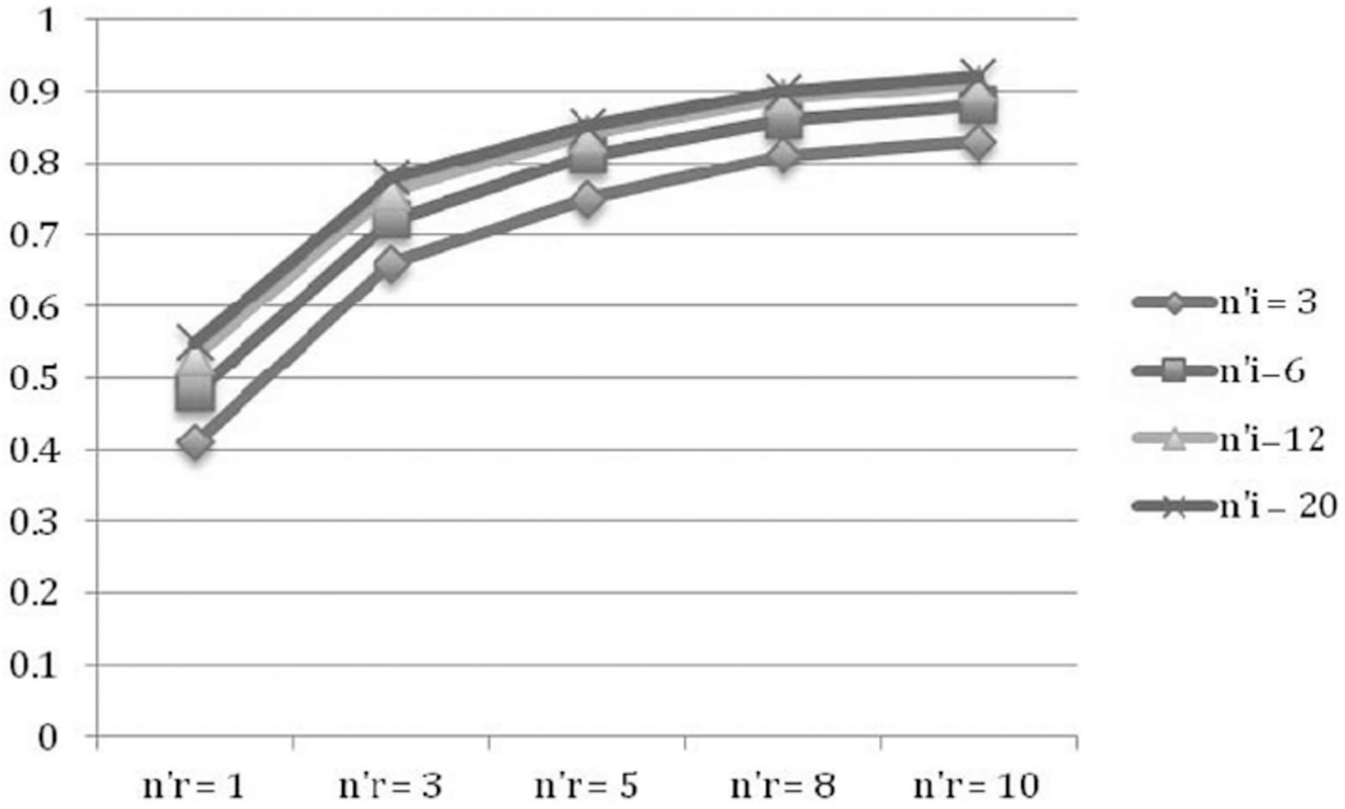

Figure 1.
Estimated *g* coefficients for different *D* study designs using the Response to Challenge Scale–Cognitive Control subscale in a sample of kindergarteners and first-grade students (see Table 4) are shown here graphically. *n'r* refers to the number of raters; *n'i* refers to the number of items.

Table 1

Comparison of PRI for the Response to Challenge Scale–Cognitive Subscale in Homogeneous and Heterogeneous Groups

	Homogeneous group					Heterogeneous group				
Var src	***df***	**Var est**	***SE***	**% Var**	***p***	***df***	**Var est**	***SE***	**% Var**	***p***
	Grades K–1					Grades K–5				
P	58	0.29	0.07	24	.00**	180	0.33	0.04	32	.00**
R	4	0.10	0.06	8	.14	4	0.06	0.04	6	.11
I	5	0.00	0.01	0	.43	5	0.00	0.01	0	.40
PR	232	0.22	0.03	18	.00**	720	0.14	0.01	14	.00**
PI	290	0.06	0.01	5	.00**	900	0.06	0.06	5	.00**
RI	20	0.04	0.01	3	.02*	20	0.04	0.04	4	.00**
PRI,e	1160	0.52	0.02	42	.00**	3600	0.42	0.42	39	.00**
	Grades 2–3					Grades K–5				
P	58	0.18	0.04	25	.00**	180	0.33	0.04	32	.00**
R	4	0.04	0.03	6	.07	4	0.06	0.04	6	.11
I	5	0.00	0.01	0	.50	5	0.00	0.01	0	.40
PR	232	0.10	0.01	14	.00**	720	0.14	0.01	14	.00**
PI	290	0.05	0.01	7	.00**	900	0.06	0.06	5	.00**
RI	20	0.06	0.02	8	.00**	20	0.04	0.04	4	.00**
PRI,e	1160	0.30	0.01	41	.00**	3600	0.42	0.42	39	.00**
	Grades 4–5					Grades K–5				
P	62	0.22	0.05	26	.00**	180	0.33	0.04	32	.00**
R	4	0.07	0.05	8	.08	4	0.06	0.04	6	.11
I	5	0.01	0.01	1	.19	5	0.00	0.01	0	.40
PR	248	0.01	0.02	1	.17	720	0.14	0.01	14	.00**
PI	310	0.06	0.01	7	.00**	900	0.06	0.06	5	.00**
RI	20	0.05	0.02	7	.00**	20	0.04	0.04	4	.00**
PRI,e	1240	0.41	0.02	49	.00**	3600	0.42	0.42	39	.00**

Psychol Assess. Author manuscript; available in PMC 2014 June 01.

Note. *P* = person; *R* = rater; *I* = item; Var src = source of variance; Var est = variance estimate; *SE* = standard error; % Var = percentage of total variance; *PRI,e* = three-way interaction of person, rater, and item. Heterogeneous group results were reported in Lakes and Hoyt (2009).

* $p < .05$.

** $p < .01$.

Table 2

Comparison of PRI for the Response to Challenge Scale-Affective Subscale in Homogeneous and Heterogeneous Groups

	Homogeneous group					Heterogeneous group				
Var src	***df***	**Var est**	***SE***	**% Var**	***p***	***df***	**Var est**	***SE***	**% Var**	***p***
			Grades K-l					Grades K-5		
P	58	0.54	0.11	37	.00**	180	0.64	0.07	46	.00**
R	4	0.04	0.03	3	.21	4	0.04	0.03	3	.19
I	6	0.09	0.05	6	.14	6	0.06	0.04	5	.13
PR	232	0.27	0.03	18	.00**	720	0.19	0.01	14	.00**
PI	348	0.04	0.01	4	.00**	1080	0.04	0.01	3	.00**
RI	24	0.06	0.02	4	.02*	24	0.06	0.02	4	.01**
PRI,e	1392	0.42	0.02	29	.00**	4320	0.35	0.01	26	.00**
			Grades 2–3					Grades K-5		
P	58	0.43	0.09	42	.00**	180	0.64	0.07	46	.00**
R	4	0.03	0.02	3	.15	4	0.04	0.03	3	.19
I	6	0.05	0.03	5	.09	6	0.06	0.04	5	.13
PR	232	0.13	0.02	12	.00**	720	0.19	0.01	14	.00**
PI	348	0.03	0.01	3	.00**	1080	0.04	0.01	3	.00**
RI	24	0.06	0.02	6	.00**	24	0.06	0.02	4	.01**
PRI,e	1392	0.28	0.01	28	.00**	4320	0.35	0.01	26	.00**
			Grades 4–5					Grades K-5		
P	62	0.49	0.09	41	.00**	180	0.64	0.07	46	.00**
R	4	0.04	0.03	4	.16	4	0.04	0.03	3	.19
I	6	0.05	0.03	5	.11	6	0.06	0.04	5	.13
PR	248	0.16	0.02	13	.00**	720	0.19	0.01	14	.00**
PI	372	0.04	0.01	4	.00**	1080	0.04	0.01	3	.00**
RI	24	0.06	0.02	5	.00**	24	0.06	0.02	4	.01**
PRI.e	62	0.49	0.09	41	.00**	4320	0.35	0.01	26	.00**

Psychol Assess. Author manuscript; available in PMC 2014 June 01.

Note. *P* = person; *R* = rater; *I* = item; Var src = source of variance; Var est = variance estimate; *SE* = standard error; % Var = percentage of total variance; *PRI,e* = three-way interaction of person, rater, and item. Heterogeneous group results were reported in Lakes and Hoyt (2009).

* $p < .05$.

** $p < .01$.

Table 3

Comparison of PRI for the Response to Challenge Scale–Motor Subscale in Homogeneous and Heterogeneous Groups

	Homogeneous group					Heterogeneous group				
Var src	***df***	**Var est**	***SE***	**% Var**	***p***	***df***	**Var est**	***SE***	**% Var**	***p***
			Grades K-l					Grades K-5		
P	58	0.51	0.11	39	.00**	180	0.66	0.08	47	.00**
R	4	0.06	0.05	5	.18	4	0.07	0.04	5	.18
I	2	0.00	0.01	0	.48	2	0.00	0.00	0	.50
PR	232	0.36	0.04	27	.00**	720	0.27	0.02	19	.00**
PI	116	0.02	0.02	2	.11	360	0.04	0.01	3	.00**
RI	8	0.03	0.01	2	.10	8	0.02	0.01	2	.09
PRI,e	464	0.35	0.02	26	.00**	1440	0.35	0.01	25	.00**
			Grades 2–3					Grades K-5		
P	58	0.63	0.13	51	.00**	180	0.66	0.08	47	.00**
R	4	0.04	0.03	4	.16	4	0.07	0.04	5	.18
I	2	0.00	0.00	0	.50	2	0.00	0.00	0	.50
PR	232	0.24	0.03	20	.00**	720	0.27	0.02	19	.00**
PI	116	0.01	0.01	1	.22	360	0.04	0.01	3	.00**
RI	8	0.01	0.01	1	.12	8	0.02	0.01	2	.09
PRI,e	464	0.29	0.02	24	.00**	1440	0.35	0.01	25	.00**
			Grades 4–5					Grades K-5		
P	62	0.62	0.12	43	.00**	180	0.66	0.08	47	.00**
R	4	0.12	0.08	8	.17	4	0.07	0.04	5	.18
I	2	0.00	0.00	0	.50	2	0.00	0.00	0	.50
PR	248	0.17	0.03	12	.00**	720	0.27	0.02	19	.00**
PI	124	0.07	0.02	5	.01**	360	0.04	0.01	3	.00**
RI	8	0.03	0.02	2	.12	8	0.02	0.01	2	.09
PRI,e	496	0.42	0.03	29	.00**	1440	0.35	0.01	25	.00**

Note. P = person; *R* = rater; *I* = item; Var src = source of variance; Var est = variance estimate; *SE* = standard error; % Var = percentage of total variance; *PRI,e* = three-way interaction of person, rater, and item. Heterogeneous group results were reported in Lakes and Hoyt (2009).

** $p < .01$

Table 4

Expected g Coefficients for Various Decision (D) Study Designs in Fully Crossed Homogeneous Samples (PRI)

No. of Raters × No. of Items		RCS Cognitive *g*			RCS Affective *g*			RCS Motor *g*		
n'r	*n'i*	Grades K-l	Grades 2–3	Grades 4–5	Grades K-l	Grades 2–3	Grades 4–5	Grades K-l	Grades 2–3	Grades 4–5
1	3	.41	.44	.54	.55	.64	.58	.51	.65	.64
	6	.48	.53	.69	.60	.70	.66	.55	.68	.71
	12	.53	.58	.80	.63	.73	.70	.57	.70	.74
	20	.55	.61	.85	.65	.75	.72	.58	.71	.76
3	3	.66	.68	.74	.78	.83	.79	.75	.84	.82
	6	.72	.75	.85	.81	.87	.84	.78	.87	.87
	12	.76	.80	.91	.84	.89	.87	.80	.88	.89
	20	.78	.82	.94	.84	.90	.88	.80	.88	.90
5	3	.75	.76	.80	.84	.88	.86	.83	.90	.87
	6	.81	.82	.89	.88	.91	.90	.85	.91	.91
	12	.84	.86	.94	.89	.93	.92	.87	.92	.93
	20	.85	.88	.96	.90	.93	.93	.87	.93	.94
8	3	.81	.81	.84	.89	.91	.90	.88	.93	.91
	6	.86	.87	.91	.91	.94	.93	.90	.94	.94
	12	.89	.90	.95	.93	.95	.94	.91	.95	.95
	20	.90	.92	.97	.93	.96	.95	.91	.95	.96
10	3	.83	.83	.86	.91	.93	.91	.90	.95	.92
	6	.88	.89	.92	.92	.95	.94	.92	.95	.94
	12	.91	.92	.96	.94	.96	.95	.93	.96	.96
	20	.92	.93	.97	.95	.96	.96	.93	.96	.96

Note. g = predicted generalizability coefficient for a *D* study with the specified number of items (*I*) and raters (*R*), when items and raters are crossed with persons (*P*; i.e., same group of raters for each ratee); RCS = Response to Challenge Scale; *n'i* = number of items; *n'r* = number of raters. Results are based on the same *Ns* reported in Tables 1–3 for homogeneous groups.