Generalizability and Decision Studies: An Example Using Conversational Language Samples

Cornelia Taylor Bruckner, Paul J. Yoder and R.A. McWilliam

Journal of Early Intervention 2006 28: 139
DOI: 10.1177/105381510602800205

Journal of Early Intervention, 2006
Vol. 28, No. 2, 139–153

RESEARCH METHODS ARTICLE

Generalizability and Decision Studies: An Example Using Conversational Language Samples

CORNELIA TAYLOR BRUCKNER, PAUL J. YODER, AND R. A. McWILLIAM
Vanderbilt University

Observational data collected in ecologically valid measurement contexts are likely to be influenced by contextual factors irrelevant to the research question. Using multiple sessions and raters often improves the stability of scores for variables from such contexts. Generalizability (G) theory can be used to give researchers important information about sources and amounts of error in their observed scores. G theory also can be used to estimate the number of raters and sessions necessary to obtain reliable scores by conducting a decision (D) study. A general overview of G theory is provided in the present paper and its potential application for one type of a two-dimensional, fully crossed observational design is illustrated, using conversational language samples obtained from preschoolers with grammatical and phonological impairments. The paper includes the Internet link to an Excel™ spreadsheet, which calculates values necessary for the type of G and D studies exemplified here.

Researchers in early childhood special education place a high value on naturalistic measurement contexts to maximize the ecological validity and generalizability of findings (Odom, Favazza, Brown, & Horn, 2000). One characteristic of most naturalistic contexts is that many aspects are left to vary across participants and sessions. Measurement contexts allowed to vary are termed *unstructured*. Factors that are allowed to vary in unstructured measurement conditions and that affect a variable's observed scores cause the variable to be unstable across sessions. Therefore, variance across measurement conditions has a negative impact on the reliability of observed scores.

Reliability is a measure of the consistency of observed scores (Cronbach, Rajaratnam, & Gleser, 1963). Reduced reliability usually leads to an increased probability of Type II error (Cohen, Cohen, West, & Aiken, 2003). Type II errors occur when the null hypothesis is not rejected when it is false in the population. Reliability estimates aid interpretation of results of research studies and comparison of findings across research studies. The American Psychological Association Task Force on Statistical Inference recommends that reliability be reported for all scores reported in APA journals.

> It is important to remember that a test is not reliable or unreliable... Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric. (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 596)

Even with this specific recommendation and the risk of Type II errors, it is estimated that 75% of empirical studies do not report the reliability of scores (Vacha-Haase, Henson, & Caruso, 2002).

According to classical test theory, all observed scores are composed of the true score plus error. To the extent that error can be minimized, Type II errors can be reduced (Cohen et al., 2003). The true score is defined in classical test theory as a personal (or object of measurement) parameter that remains constant across parallel forms of measurement (Feldt & Brennan, 1989). When a variable consists of systematic observations of participants across time, different sessions are analogous to *parallel forms of measurement*. A true score cannot be perfectly quantified. Observed scores vary in the extent to which they reflect a true score. In practical terms, the best estimate of a true score is the most accurate representation of what we intend to measure. Observed scores with much error do not reflect "true" between-subject (or object of measurement) variance as accurately as observed scores with less error.

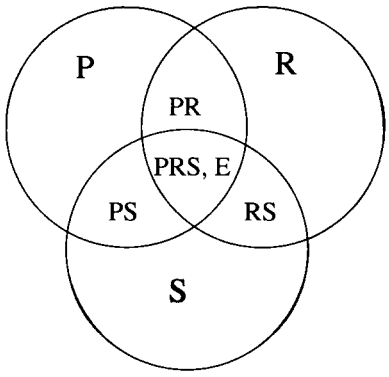
Researchers in early childhood special education often want to measure stable tendencies to act in a certain way (e.g., engage with materials) or stable abilities (e.g., use language to communicate). Using many sessions and raters to measure a variable for every participant is expensive, so methods that permit estimation of the numbers of sessions and raters required to obtain stable estimates would be useful. Because reliability is a characteristic of scores not of tests, reliability should be recalculated for each variable, measurement context, and subject population (Thompson & Vacha-Haase, 2000). G theory is a measurement approach that allows researchers to identify the number of sessions, raters, and other aspects of measurement needed to produce reliable score estimates in unstructured measurement contexts such as naturalistic contexts. G theory improves on classical test theory by (a) estimating main effects and interaction effects for all aspects of the measurement context simultaneously (Thompson, 2003), (b) comparing reliability across combinations of levels of the aspects of a measurement context, and (c) making explicit the measurement contexts to which results can be generalized (Eason, 1991; Thompson, 1991; Webb, Rowley & Shavelson, 1988).

The purpose of this article is to provide a general overview of one application of G theory to a common problem in observational research (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). For observational researchers, two critical aspects of measurement are sessions and raters: How many are needed to have scores interpretable as stable or dependable? Three differences between this tutorial and other tutorials on G theory (Cronbach & Shavelson, 2004; O'Brian, O'Brian, Packman, & Onslow, 2003; Scarsellone, 1998; Shavelson & Webb, 1991) are that this one includes (a) a presentation of the need for information from G theory when observational data are collected in unstructured contexts, (b) a link to a webpage that calculates the necessary variance components and coefficients, and (c) an applied example of a frequently used measurement context for preschoolers with disabilities: conversational-language samples.

General Description of Generalizability Theory

G theory is a statistical method developed by Cronbach et al. (1972) to analyze the reliability of test scores. Whereas classical test theory views measurement error as a unitary quantity, G theory partitions error by its sources using the logic of analysis of variance (ANOVA; Cronbach, Rajaratnam, & Gleser, 1963). Figure 1 shows an example of partitioning variance into components including main effects of person, rater, session, the two-way interaction between person and session, person and rater, rater and session, and the three-way interaction among person, rater, and session plus other sources of variance not accounted for by the model. Variance due to person is the estimate for true score variance when person is the object of measurement.

G theory expands on classical test theory by stating that various aspects (i.e., facets) of the measurement process might affect observed scores and that the degree of error due to



Venn diagram showing the partitioning of variance in the two-facet, fully crossed G study.
P = proportion of variance in individuals' observed scores due to true score (object of measurement).
S = session facet
R = rater facet
PR = proportion of variance in individuals' observed scores due to interactions between person and rater.
PS = proportion of variance in individuals' observed scores due to interactions between person and session.
RS = proportion of variance in individuals' observed scores due to interactions between rater and session.
PRS,E = proportion of variance in individuals' observed scores due to interactions across person, rater, session, and error.

Figure 1.
Venn diagram showing the partitioning of variance in the two-facet, fully crossed G study.

different facets can be estimated (Knowles & Condon, 2000). Information from a G study can be used to conduct a decision (D) study to evaluate and compare the reliability of scores under various study designs (e.g., four vs. five raters, nested designs vs. crossed designs). This is similar to using the Spearman-Brown prophecy formula to estimate the number of items needed to obtain a reliable score on a test (Crocker & Algina, 1986). The Spearman-Brown prophecy formula treats error as a unitary term including error due to all facets of measurement (Lord & Novick, 1968). The formula is used to determine the number of items (or components) that are necessary to obtain a desired criterion of reliability. A D study improves on this method by considering the contribution of each component of error variance and adjusting reliability estimates under different combinations of levels of facets and different study designs.

The elements of a G or D study design are the number of facets, sampling method used to select individual forms of the facets (i.e., fixed vs. random), and the pairing of conditions between facets (i.e., crossed vs. nested). The elements of the G and D study can be combined to create many different designs (e.g., two facets fully crossed with random facets, three facets nested with fixed facets). G theory can be used to determine the reliability of scores used for relative or absolute decisions. For scores to be reliable for relative decisions they should rank the units of measurement in a similar order across facets and forms. Reliable ranking is important when data are used to estimate the relationship between two variables using correlational methods or the difference between two or more groups examined within the study when rankings and not absolute levels of performance are of interest. An example of a question that requires relative reliability is, "Does a positive association exist between reading and math ability?" For scores to be reliable for absolute decisions they should reliably rank participants and take on similar values across facets of measurement. A reliable absolute value of scores is important when the score is interpreted by its level (e.g., absolute number of items correct) or is being judged against a criterion external to the sample. For example, when an IQ test score is used to determine if a child has an intellectual disability, his or her score is judged against a criterion of 70. If his or her score is less than 70, the score is in the range of mental retardation. This decision is made regardless of the child's ranking among the other children in the sample. Absolute decisions require a more stringent test of reliability than do relative decisions (Shavelson & Webb, 1991).

To calculate how much influence different facets of measurement have on observed scores, one must have parallel forms of each predicted source of error. In classical test theory, parallel forms are matched and expected to have equivalent means, variances, and covariances across forms (Lord,

1955; Stanley, 1971). In G theory, parallel forms are random and assumed to be drawn from a population of possible forms. Samples of forms or conditions, drawn from the appropriate population of forms or conditions, are expected to have a mean, variance, and covariance representative of the population parameters, but there is no requirement that individual forms are equivalent (Cronbach, Rajaratnam, & Gleser, 1963). Conditions are considered part of a set if they would be acceptable substitutes for each other (Webb, Rowley, & Shavelson; 1988). For each facet of the G study, there must be more than one condition randomly selected from the population of conditions or *universe*. For example, if children were the objects of measurement and the reliability of scores across observational sessions was being investigated, the children would be observed during multiple sessions. In this example, the design is fully-crossed, session is the facet of interest, and researchers would randomly select a set of similar sessions, or conditions, across which we expect observed scores to be consistent or reliable from a universe of sessions (Brennan, 2001).

Previous Application of G theory in Observational Measurement in Early Childhood

G theory has been applied previously to observational measurement in early childhood education. McWilliam and Ware (1994) conducted a study to determine the effect of measuring variables across different sessions with different raters. The dependent variables of interest were nine types of engagement in children with disabilities. The data were collected in a series of 15-minute sessions within a childcare center. The fully crossed design used three raters and four sessions for every child. Thus, the facets were raters and sessions. Using ANOVA logic, the researchers estimated the extent to which each facet accounted for between-subject variance in the observed scores. Total error variance was partitioned into variance attributable to interactions between raters and children; sessions and children; and a three-way interaction among sessions, raters, and children plus error not accounted for by the model. The interaction of raters and sessions was not of interest because it did not influence the ranking of children on the observed scores. Error owing to the interaction between raters and children represents different raters giving different relative standings to children. Error owing to the interaction between sessions and children represents differences in the relative standings of children in different sessions. Error owing to the three-way interaction can be interpreted as variations in relative child standings as a function of both sessions *and* raters or other unsystematic or systematic sources of variation that were not measured in the study.

McWilliam and Ware calculated a type of reliability coefficient for each variable that indicated the proportion of the observed score due to true score. The reliability of the scores for the nine types of engagement in the study ranged from .38 to .83 (M = .60, SD = .16).

Of interest from this study was the high proportion of variance in the observed scores attributable to a Sessions × Children interaction (see Figure 2). A Sessions × Children interaction in the context of this type of reliability study indicates that the relative ranking of children differed across sessions. For example, a child ranked highest in one of the types of engagement in the first session might be ranked in the middle in the second session and at the bottom in the third sesssion. This results in any one session producing inaccurate rankings of the children on the variable of interest. In the McWilliam and Ware study, 50% of the variance in observed scores for the coding categories *engagement with materials* and *nonengaged* was due to the interaction between sessions and children. The reliability of these scores was very low and any statistical analyses performed on these scores would have high Type II error and low statistical power. Through the use of D study analyses, McWilliam and Ware demonstrated that averaging across many sessions and

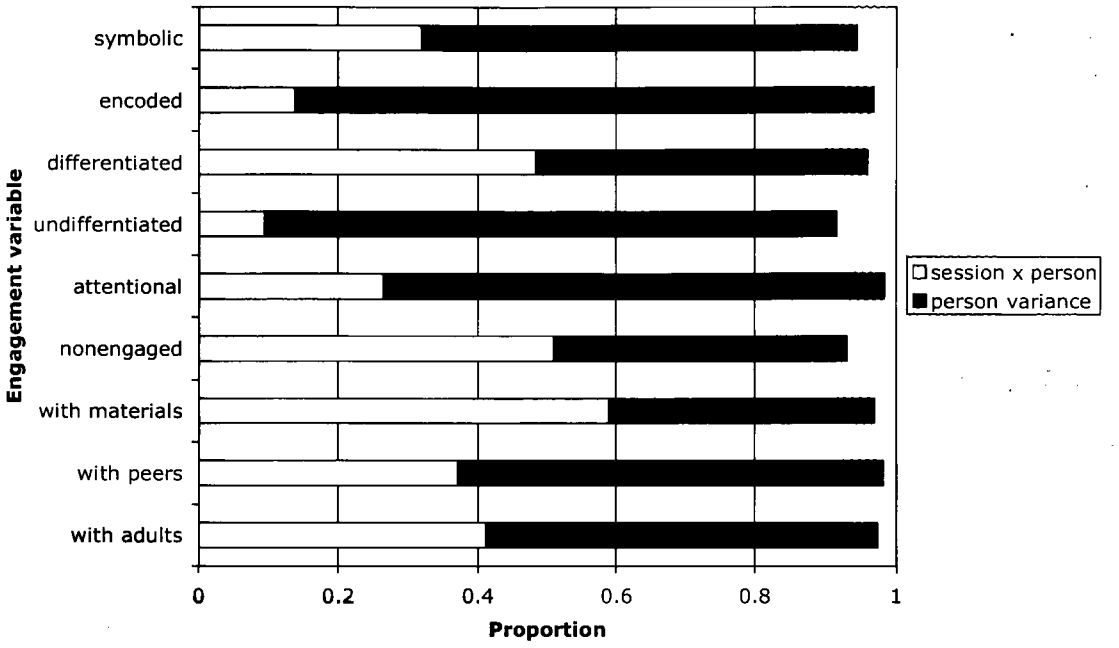


Figure 2.
Proportion of variance in observed engagement scores due to true score and session when scores are averaged across four sessions and four raters. Note the variance due to the interaction between rater and person is not included so not all bars extend to one.

raters would increase the stability (i.e., reliability) of the scores.

Similar to the McWilliam and Ware design, the example in the present paper focuses on a fully crossed design with raters and sessions as facets of measurement with observed scores used to make relative decisions. Many other measurement conditions could be considered facets. A G study design that includes a score for every person on every level of every facet is considered "fully crossed." This paper will deal only with the fully crossed model, assuming sessions and raters are the facets of interest and relative decisions are to be made. For information on the application of other models (e.g., nested designs), see Brennan (2001) or Shavelson and Webb (1991).

Calculation of the g Coefficient

The G study described above would use the mean squares and *N* values from an ANOVA performed on the fully crossed scores to estimate the reliability of obtained between-subject variance on the continuous measure of interest (see Table 1). Using calculators or available software to compute these variance components and subsequent *g* coefficients can be extremely tedious. Therefore, we provide the Internet link to a website that calculates the *g* coefficients and their component values (i.e., variance components) for fully crossed two-facet designs that will be used to make relative decisions (http://kc.vanderbilt.edu/quant/gcalc/). Knowing how these values are calculated, however, illustrates the logic behind them.

The *g* coefficient is the proportion of the observed between-subject variance due to the best estimate of the true score. The best estimate of the true score, when sessions and raters are the facets of interest, is the average of scores across sessions and raters within subject. The estimate of observed between-subject variance as shown in equation 1 is the sum of estimated true score variance plus all sources of error (see Table 1 for definitions of variance components). Because participants are the objects of measurement, it is expected that this variance component will be large;

Table 1
Equations for the Variance Components of the G Study

Source of Variation	Variance component	Equation for calculation from mean squares (MS)
Persons (p)	σ^2p	(MSp − MSps − MSpr + MSprs,e)/(nr*ns)
Person × rater (pr)	σ^2pr	(MSpr − MSprs,e)/ns
Person × session (ps)	σ^2ps	(MSps − MSprs,e)/nr
Person × rater × session, error (prs,e)	σ^2prs,e	MSprs,e

Note. Adapted from Shavelson and Webb (1991); MSp = mean square person; MSps = mean square person × session; MSpsr,e = mean square person × session × rater + error; nr = number of raters; ns = number of sessions.

we expect people to differ in their ability or aptitude on the variable of interest. The *g* coefficient calculated in Equation 1 represents the reliability of scores from one random rater and one random session. As will be demonstrated later in the paper, *g* coefficients can be calculated for alternative combinations of number of raters and sessions. The σ^2 symbols in the formula are variance components, where p = person (object of measurement), r = raters, s = session, and e = residual error.

g coefficient:

$$= \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pr} + \sigma^2_{ps} + \sigma^2_{prs,e}} \quad (1)$$

The variance components are derived from the mean squares and relevant n values obtained from the ANOVAs. To derive the mean squares from the ANOVA, data should be entered into a spreadsheet so that scores are organized into columns. Figure 3 illustrates a spreadsheet using SPSS format for a single dependent variable. Separate ANOVAs are necessary for each dependent variable. It is useful to note that, unlike between-subjects ANOVA, the participant ID is listed more than once in the first column. Two more columns contain a number representing particular sessions and raters. Finally, a fourth column contains the dependent variable scores produced for each case (i.e., subject/session/rater unit). The number of cases (i.e., the number of rows in the spreadsheet) is the number of participants × number of sessions × number of raters. In this way, each participant will have several scores (i.e., number of sessions × number of raters) entered into the same column for the dependent variable. Using these data, a two-way ANOVA is conducted with sessions and raters as the factors (columns) and people as the objects of measurement (rows) and observed scores as the dependent variable (see Figure 3). Two-way ANOVA procedures are available through most statistical software packages including SPSS and SAS. SPSS also includes an option that calculates variance components. This option can be used to derive the estimated variance of the object of measurement, facets of the measurement context, and the two- and three-way interactions.

To calculate the estimated variance components and the *g* coefficient (i.e., the reliability coefficient) for the dependent variable from one session and one rater the following results of the ANOVA should be entered into the first row of the Excel™ spreadsheet: (a) the mean square for person (i.e., subject), (b) the mean square for the Person × Session interaction, (c) mean square for the Person × Rater interaction, (d) mean square for the Person × Rater × Session interaction, (e) the number of participants in the G study, (f) the number of fully crossed raters, and (g) the number of fully crossed sessions.

The calculation of variance components using the ANOVA model assumes that scores are normally distributed across the facets. An

	child	rater	session	dv
1	2	1	1	56
2	2	1	2	31
3	2	2	1	47
4	2	2	2	31
5	5	1	1	56
6	5	1	2	62
7	5	2	1	61
8	5	2	2	58
9	9	1	1	86
10	9	1	2	40
11	9	2	1	80
12	9	2	2	64

Tests of Between-Subjects Effects

Dependent Variable: DV

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26173.958[a]	95	275.515	.	.
Intercept	200020.042	1	200020.042	.	.
CHILD	15604.958	23	678.476	.	.
RATER	.667	1	.667	.	.
SESSION	57.042	1	57.042	.	.
CHILD * RATER	3780.333	23	164.362	.	.
CHILD * SESSION	5915.958	23	257.216	.	.
RATER * SESSION	6.000	1	6.000	.	.
CHILD * RATER * SESSION	809.000	23	35.174	.	.
Error	.000	0	.		
Total	226194.000	96			
Corrected Total	26173.958	95			

a. R Squared = 1.000 (Adjusted R Squared = .)

Figure 3.
Example SPSS™ data entry and output for analysis of variance to generate variance component estimates.

alternative hypothesis is that data obtained from the observation of behavior are better modeled as a time-dependent distribution like the Poisson distribution (Rogosa & Ghandour, 1991). If observational data are better represented by a Poisson distribution, problems arise in the interpretation of the Person × Session interaction and the three-way interaction error term. When deviations from normality exist in data obtained from the observation of behavior, data might be better represented by a time-dependent distribution. When data are better represented by a time-dependent distribution, one should calculate variance components using cumulative process formulation (see Rogosa & Ghandour, 1991 for a full description of this method). The Kolmogorov-Smirnov statistic and the Shapiro-Wilk statistic are available in SPSS to test the null hypothesis that data follow a normal distribution (see SPSS help menu for additional information).

Using the Spreadsheet to Conduct Decision Studies

The variance estimates from the G study can be used to make decisions about the number of sessions and raters across which an investigator would need to average scores to obtain a reliable estimate of participants' true scores. This application of G theory is called a decision (D) study. A D study uses obtained variance components to compute the *g* coefficient for different scenarios involving hypothesized numbers of facets, such as sessions and raters. The crux of a D study for the fully crossed, two-facet design described in the present paper is a quantity called the "relative error variance" which is calculated using Equation 2. As before, the σ^2 symbols in the formula are variance components, p = person, r = rater, s = session, and e = residual error.

Relative error variance:

$$\sigma^2_{Rel} = \frac{\sigma^2_{pr}}{n'_r} + \frac{\sigma^2_{ps}}{n'_s} + \frac{\sigma^2_{prs,e}}{n'_r n'_s} \qquad (2)$$

The *g* coefficients for the D-study are calculated by dividing the variance for persons (σ^2_p) by $\sigma^2_p + \sigma^2_{Rel}$. Values for *n*, as implied by the relative error formula, affect relative error variance and thus the resulting *g* coefficient in the D study. To use the Excel™ spreadsheet to estimate the *g* coefficient under hypothesized numbers of sessions and raters, copy the mean squares obtained in the G study from the first row in the Excel™ spreadsheet to all rows that will be included in the calculation of the *g* coefficients. Note that the variance component for the Session × Rater × Person interaction is the mean square for that term. Enter the hypothesized *n* values for sessions and raters (rows 2 – 16 in Figure 4). The decision of how many raters and sessions to use in future studies should be based on what it takes to attain *g* coefficients greater than the criterion level G (e.g., .70). In selecting among the hypothetical scenarios producing *g* coefficients greater than the selected criterion, the investigator should also consider the relative feasibility of adding raters versus sessions and associated costs. If using the variance components option in SPSS, the estimated variance for the object of measurement, facets of the measurement context, and interactions are calculated by SPSS. Equations 1 and 2 can be used to calculate the reliability under different numbers of sessions and raters from the estimated variance components provided by the SPSS *variance components* option. Other computer programs that can identify the number of levels within each facet that maximizes reliability of scores while minimizing costs are available (Parkes, 2000).

Generalizability Theory Applied to Language Sampling

To demonstrate the values of G and D studies and the Excel™ spreadsheet for unstructured measurement contexts, we examined selected speech and language variables from conversational language samples in preschoolers with grammatical and phonological impairments. Because the most frequent communicative context for preschoolers is conversation, it is considered the most ecologically valid. Conversational

H	I	J	K	L	M	N
VARp	VARps	VARpr	VARrel	G	hypNr	hypNs
0.392675	0.3141	0.0768	0.6153	0.38956819	1	1
0.392675	0.3141	0.0768	0.34605	0.53155775	1	2
0.392675	0.3141	0.0768	0.2563	0.60506953	1	3
0.392675	0.3141	0.0768	0.211425	0.65001655	1	4
0.392675	0.3141	0.0768	0.4647	0.45799679	2	1
0.392675	0.3141	0.0768	0.25155	0.60953083	2	2
0.392675	0.3141	0.0768	0.1805	0.68508745	2	3
0.392675	0.3141	0.0768	0.144975	0.73035432	2	4
0.392675	0.3141	0.0768	0.4145	0.48648063	3	1
0.392675	0.3141	0.0768	0.22005	0.64086662	3	2
0.392675	0.3141	0.0768	0.15523333	0.7166801	3	3
0.392675	0.3141	0.0768	0.122825	0.76173618	3	4
0.392675	0.3141	0.0768	0.3894	0.50209379	4	1
0.392675	0.3141	0.0768	0.2043	0.65777461	4	2
0.392675	0.3141	0.0768	0.1426	0.73359488	4	3
0.392675	0.3141	0.0768	0.11175	0.77846062	4	4

Note. VARp = variance person; VARps = variance person x session; VARpr = variance person x rater; VARrel = relative error variance; G = generalizability coefficient; hypNr = hypothetical number of raters; hypNs = hypothetical number of sessions.

Figure 4.
D study estimates calculated from the spreadsheet formulas.

sampling is the measurement context of choice for many speech and language variables in preschoolers with disabilities (Miller, 1981). In conversations, communication partners do not completely control children's topics or what they say about the chosen topics. The variability introduced by allowing topics and content to vary might affect stability across raters and language samples for many speech and language variables.

This issue might be particularly salient in children with phonological impairments, because the stability of speech and language variables might be poor in children with impairments owing to the cognitive demands of the conversation. When discussing topics that vary in familiarity and routinization, performance of difficult speech and language skills is likely to vary widely (Evans, 2001).

To examine the stability of scores derived from unstructured language samples, we measured three primary variables: the mean length of utterance (MLU), the number of different word roots, and the percentage of fully intelligible utterance attempts. These were selected for their functional importance to children with grammatical and phonological impairments. Because the intelligibility variable is uncommonly measured due to concerns over measurement issues (Kent, 1993), we investigated reliability of the total number of utterance attempts and the total number of fully intelligible utterances. We examined the relative reliability of single versus multiword fully intelligible utterances (i.e., analysis of the numerator of the intelligibility proportion). Multiword utterances are relevant because research indicates that multiword utterances provide more support for the content and thus greater probability of agreement in transcription than do single word utterances (Kent, 1993). Finally, we conducted a D study that provided the basis for recommendations about the numbers of raters and sessions needed to produce acceptably reliable estimates of these speech and language variables and their component values.

METHOD

Participants

Twenty-four children with grammatical and phonological impairments participated in the research (see Table 2). Participants were recruited through a larger study involving

Table 2
Means and Standard Deviations for Participant (N = 24) Descriptor Variables

Variable	*M*	*SD*
Chronological age (years)	4.2	2.4
Leiter IQ	105.0	16.1
MLU in morphemes from analysis set	2.1	.6
Percentage of utterances fully intelligible	12.8	3.6
Arizona Articulation Proficiency Scale percentile ranking	2.6	2.4
PLS-3 receptive percentile rank	22.4	25.4
PLS-3 expressive percentile rank	4.3	2.6

Note. MLU = mean length of utterance; PLS-3 = Preschool Language Scale (3rd ed.).

preschoolers with grammatical and phonological impairments (Yoder, Gardner, & Camarata, 2005). Grammatical impairment was defined as MLU at least 1.3 *SD* below that expected for the child's chronological age or a ranking of 10th percentile or below on the expressive scale of the *Preschool Language Scale-3rd edition* (PLS-3; Zimmerman, Steiner, & Pond, 1992). Additionally, phonological impairment was defined as performance below the 10th percentile on the *Arizona Articulation Proficiency Scale* (AAPS; Fudala & Reynolds, 1986). All participants had nonverbal intelligence quotients above 80 and passed a speech-reception threshold screening.

Language Samples

The two, 20-min language samples were measured during conversation centered around an examiner and child playing with toys. Three different sets of developmentally appropriate toys counter-balanced across children were used. The number of toy sets was not used as a facet in the design of this G study because it was not fully crossed and the counterbalance design should randomize the effect of the toy sets on observed scores. The adult examiner was told to follow the child's attentional and play lead as long as the child continued to talk about the toy or activity. If the child did not talk about the toy or activity, the adult examiner was told to ask the child "wh"-questions (e.g. what, why, when) about the toys and activities. Additionally, the adult examiner talked about what the child was doing and about her own actions. Apart from these directives, the topics and content of language samples were allowed to vary across participants.

Two raters unfamiliar with the child transcribed both language samples independently for all participants. The software program used for the transcription was *Systematic Analysis of Language Transcripts* (SALT; Miller & Chapman, 1993) and the transcription format was that prescribed by SALT. Vocalizations were coded as words if they were in an English dictionary and had immediate nonlinguistic support (e.g., they referred to a toy to which the child was attending). Additionally, approximations to words had to meet the following criteria to be coded as a word: (a) accurate syllable structure and initial phoneme, or (b) accurate consonant-vowel or vowel-consonant combination. MLU was the average number of smallest meaningful language units (morphemes) per utterance derived from complete and fully intelligible utterances. The percentage of utterances understood was the number of utterances in which all words in the utterance has conventional shared meanings divided by the total number of utterance attempts. The number of different word roots was the number of unique words not including the same word with a different suffix, prefix, or tense.

Analyses

Variance components were calculated for six dependent variables: MLU, number of different word roots, percentage of utterance attempts understood, number of single-word

fully intelligible utterances, number of multiword fully intelligible utterances, and total number of utterance attempts. The mean squares and *ns* were extracted from the two-way ANOVA source table output in SPSS and entered into the G calculator (http://kc.vanderbilt.edu/quant/gcalc/). The variance components from the G study were subsequently used in a D study where an estimated *g* coefficient was calculated for all permutations of one to seven raters and one to seven sessions. The selection of a criterion *g* coefficient should be determined by the expected homogeneity of the scores across sessions and raters and the effect size of the difference or relation being tested in the broader research study (Cronbach et al., 1963). Because we used very similar language sampling procedures and identically trained transcribers, we expected high homogeneity of scores. We addressed this homogeneity in the context of the larger study and expected a moderate effect size ($r = .50$). Therefore, we selected what is considered a relatively large *g* coefficient of .70. The criterion level *g* coefficient used in other studies, however, might need to be larger for variables expected to have higher stability across raters and sessions and in studies with smaller effect sizes. High *g* coefficients indicate a low proportion of variance attributed to facets of the measurement context (e.g., error due to raters or sessions).

RESULTS

Results showed that scores for three of the primary variables (MLU, number of different word roots, and percentage of fully intelligible utterances) were reliable with only one rater and one session (see Table 3). Detailed analysis of the numerator and denominator of the intelligibility variable indicated, however, that at least five sessions were needed to produce a *g* coefficient over the criterion of .70 for the number of fully intelligible utterances if only one rater was used. Additionally, the D study indicated it would take six raters and seven sessions before scores were reliable for the number of understood single-word utterances. Figure 5 shows that the Person × Session variance (Var PS) component is quite large for the number of fully intelligible single-word utterances. In contrast, scores for the number of fully intelligible multiword utterances would be reliable with only one rater and two sessions (see Table 3). Because the total number of fully intelligible utterances was composed of a higher proportion of single-word than multiword utterances (t (23) = 2.04, $p < .05$, $d = .42$), the low score reliability for the number of fully intelligible utterances is most likely attributable to the low reliability for the number of intelligible single-word utterances.

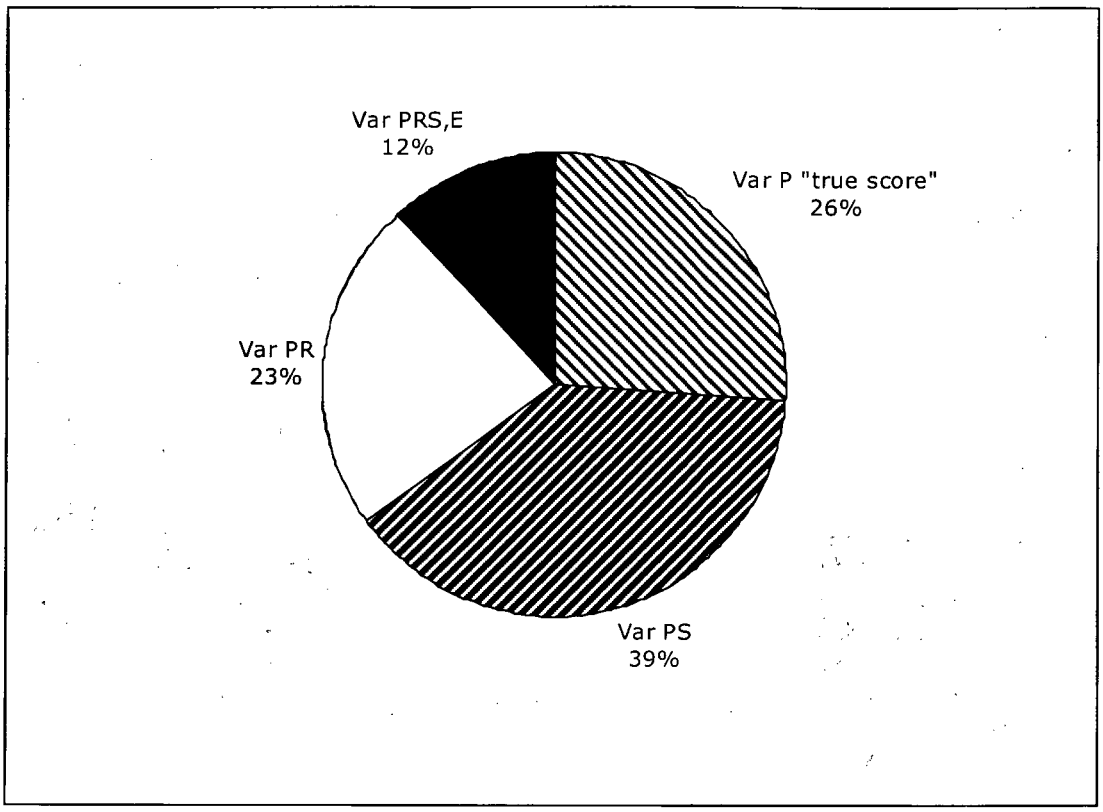
Table 3
D Study Results: Number of Sessions and Raters to Achieve g Coefficient ≥ .70

Language Measures	Sessions	Raters
MLU	1	1
Number of different word roots	1	1
Total utterance	1	1
Intelligible utterances	5	1
Single word	7	6
Multi word	2	1

Note. MLU = mean length of utterance.

DISCUSSION

This paper provided a general overview of generalizability theory and discussed its relevance for research in early childhood special education. We used a targeted example to demonstrate the value of G and D studies in identifying the number of sessions and raters needed to measure stable variables in an unstructured measurement context. We also provided a link to an Excel™ spreadsheet that facilitates the computation of G study variance components and the relative error variance components needed to compute the *g* coefficients in a D study for various hypothetical scenarios in which the numbers of raters and sessions vary. The Excel™ application is limited, however, to



Note. VAR P = variance person; VAR PS = variance person x session; VAR PR = variance person x rater; VAR PRS,E = variance person x rater x session, and error.

Figure 5.
Proportion each variance component contributed to observed score for intelligibility variable: Number of single word utterances.

fully crossed D studies with two facets of measurement.

In the context of our example application, the findings demonstrated that one 20-min session and one rater were sufficient to derive reliable score estimates of MLU, number of different word roots, and percentage of intelligible utterances in preschoolers with grammatical and phonological impairments. In the context of the D studies, one rater means that the score of a primary rater is sufficient to represent accurately the relative ranking of children in a sample as opposed to a score averaged across several raters. The D study showed many additional sessions and raters would be necessary to obtain reliable score estimates of the number of single-word utterances in preschoolers with severe grammatical and phonological impairments.

The accuracy of a D study is enhanced when a fully crossed design is used with many *representative* levels of each facet (Brennan, 2001). By representative, we mean the selected levels reflect the characteristics of the relevant universe of the facet. Sampling theory tells us that samples are most representative of the universe from which they have been selected when they are sampled randomly. Another way to maximize the probability that levels of facets are representative of the relevant universe is to select all levels of the relevant universe (Brennan, 2001). This is possible when the relevant universe is limited. An example from early childhood would be

measuring the use of independent eating from a spoon during snack time and lunch, which are the only times during the "preschool day" that the behavior occurs (i.e., these are the *exhaustive* times for the behavior). Either random or exhaustive selection of the levels of a facet is an ideal situation that is rarely, if ever, achieved.

D studies conducted on samples and measurement tools similar to those used in the present study do not substitute for the need to examine the reliability and validity of scores for dependent measures using data obtained from the study sample (Thompson & Vacha-Haase, 2000). Instead, D studies are a heuristic to help design studies and interpret findings. Like other heuristics that are widely used in planning studies (i.e., statistical power analysis), the accuracy of D studies depends on assumptions that are reasonable, but not necessarily accurate, for a particular situation. These heuristics are useful to the extent the data that produced the estimates are representative of the future study. In statistical power analysis, the key information is the effect size. In D studies, the key information is the relative error variance term for relative decisions and the absolute error term for absolute decisions.

As an example of the applied value of G studies, the findings demonstrate that the low proportion of true score variance in the variable *single word utterances understood* was mostly the result of the lack of stability of children's scores across sessions (i.e., Person × Session error variance). Because many sessions and raters were necessary to achieve acceptable score reliability on this variable, researchers might conclude that it is generally not feasible to measure the number of fully intelligible single-word utterances in an ecologically valid unstructured measurement context, such as a conversational language sample, in preschoolers with severe grammatical and phonological impairments. In variables with a large proportion of Person × Session variance, ecologically valid measurement contexts make measurement of a stable within-child characteristic logistically impossible. In contrast, scores on variables with high person variance and low Person × Session variance, like MLU, can be measured reliably in ecologically valid measurement contexts.

Within all children, a continuum of behavioral mastery exists such that mastered behaviors can be produced independently, emerging behaviors are produced in collaboration with social or environmental supports, and some behaviors cannot be produced (Vygotsky, 1978). The extremes of this range of mastery can be labeled context independent (mastered behaviors) and context dependent (emerging behaviors). We can hypothesize that, for each child, the dependent variable we are measuring lies somewhere along this continuum. Often, our research questions require that we measure stable tendencies in child behavior so we can examine individual differences between children. Without controlling the environmental supports required to elicit an emerging behavior, however, the observed score will be a product of the environmental supports available in the context and the child's mastery of the skill (Person × Session interaction). Our results support the notion that as language develops from single-word utterances to multiword utterances there is a reduction in the proportion of observed score variance due to Person × Session interactions. One interpretation of this finding is that as children begin to master language they depend less on contextual cues to produce accurate utterances. The measurement conundrum is how to measure context dependent variables in ecologically valid unstructured contexts where the features allowed to vary across sessions affect the rate and level of the dependent variable. This Personal × Session interaction would reduce the g coefficient for the dependent variable, indicating poor score reliability (see Table 4).

Results of the present study suggest that to obtain a stable measurement of context-dependent variables, a structured measurement condition might be necessary if reliability of scores is important. If we choose to measure context-dependent variables in naturalistic and thus unstructured conditions, we are probably more accurate limiting

Table 4
Expected g Coefficient: Levels of Structure in Measurement Context and Context Dependence of Behavior

		Behavior	
		Context dependent	Context independent
Measurement context	Unstructured	Low g	High g
	Structured	High g	High g

interpretation of the variable as dependent on particular measurement contexts, instead of primarily as stable characteristics of children. In contrast, we might measure stable communication abilities in unstructured measurement contexts once children's communication behaviors become less context-dependent.

Generalizability theory is a powerful tool for analyzing score reliability in ecologically valid measurement contexts. Increased reliability of scores will increase power to detect functional relationships that exist between variables of interest. G theory improves on classical test theory models of reliability by simultaneously estimating error attributable to different facets of the measurement context. Knowing the proportion of variance in observed scores that is due to the object of measurement and the proportion due to facets of the measurement context allows a researcher to plan studies that reduce variance in observed scores due to error associated with various facets. Two methods available for reducing error variance in context-dependent variables are (a) averaging across several levels of the facet or facets that are contributing the largest proportion of error variance and (b) structuring the measurement context so that opportunities and supports are similar across the objects of measurement. When ecological validity is of paramount importance, researchers should use the first method for increasing score reliability. When resources are limited or ecological validity is not required, adding structure to the measurement context can be an affordable mechanism for increasing the reliability of scores.

REFERENCES

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: John Wiley.

Cronbach, L. J., Rajaratnam, N. R., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology, 16*, 137–163.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391–418.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Evans, J. L. (2001). An emergent account of language impairments in children with SLI: Implications for assessment and intervention. *Journal of Communication Disorders, 34*, 39–54.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd. ed.*, pp. 105–147). New York: Macmillan.

Fudala, J., & Reynolds, W. (1986). *Arizona Articulation Proficiency Scale (2nd ed.)*. Los Angeles: Western Psychological Services.

Kent, R. (1993). Speech-intelligibility and communication competence in children. In A. Kaiser & D. Gray (Eds.), *Enhancing children's communication: Research foundations for intervention* (pp. 223–242). Baltimore: Brookes.

Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12*, 245–252.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325–336.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention, 18*, 34–47.

Miller, J. (1981). *Assessing language production in children: Experimental procedures*. Boston: Allyn and Bacon.

Miller, J., & Chapman, R. (1993). *SALT: Systematic analysis of language transcripts*. Madison, WI: University of Wisconsin.

O'Brian, N., O'Brian, S., Packman, A., & Onslow, M. (2003). Generalizability theory: Assessing reliability of observational data in the communication sciences. *Journal of Speech Language and Hearing Research, 46*, 711–7.

Odom, S. L., Favzza, P. C., Brown, W. H., & Horn, E. M. (2000). Approaches to understanding the ecology of early childhood environments for children with disabilities. In T. Thompson, D. Felce, & F. J. Symons (Eds.), *Behavioral observation: Technology and application in developmental disabilities* (pp. 193–214). Baltimore: Brookes.

Parkes, J. (2000). The relationship between the reliability and cost of performance assessments. *Educational Policy Analysis Archives, 8*, 16, Retrieved December 2, 2005, from http://epaa.asu.edu/epaa/v8n16/

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics, 16*, 157–252.

Scarsellone, J. M. (1998). Analysis of observational data in speech language research using Generalizability theory. *Journal of Speech Language and Hearing Research, 41*, 1341–1347.

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

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 356–442). Washington, D. C.: American Council on Education.

Thompson, B. (1991). Review of *Generalizability theory: A primer* by R. J. Shavelson & N. M. Webb. *Education and Psychological Measurement, 51*, 1069–1075.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Newbury Park, CA: Sage.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174–195.

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving towards improved understanding and use of score reliability. *Educational and Psychological Measurement, 62*, 562–569.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21*, 81–90.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Yoder, P. J., Gardner, E., & Camarata, S. (2005). Treatment effect on and predictors of speech intelligibility and length of utterance in children with specific language and intelligibility impairments. *Journal of Early Intervention, 28*, 34–49.

Zimmerman, I., Steiner, B., & Pond, R. (1992). *Preschool Language Scale (3rd ed.)*. San Antonio: Psychological Corporation.

Correspondence concerning this article should be addressed to **Cornelia Taylor Bruckner,** Department of Special Education, Peabody Box 328, Nashville, Tennessee 37203. *E-mail: cornelia.taylor@vanderbilt.edu*

The authors would like to acknowledge Dr. Andrew Tomarken for his assistance in deriving the equations for the Excel spreadsheet and Dr. Patricia Snyder for her technical review. Research supported by the National Institute of Deafness and Communication Disorders (P50DC03282).