DOCUMENT RESUME

ED 322 186	TM 015 346
AUTHOR	Lefebvre, Daniel J.; Suen, Hoi K.
TITLE	Applying Generalizability Theory To Evaluate Treatment Effect in Single-Subject Research.
PUB DATE	Apr 90
NOTE	34p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Classroom Observation Techniques; Disabilities; Educational Research; *Error of Measurement; *Evaluation Methods; *Generalizability Theory; Interpersonal Relationship; *Outcomes of Treatment; *Preschool Children; Preschool Education; *Research Methodology; Sampling; Special Education; Statistical Inference; Statistical Significance; Student Behavior
IDENTIFIERS	*Single Subject Research Design; Type I Errors

ABSTRACT

An empirical investigation of methodological issues associated with evaluating treatment effect in single-subject research (SSR) designs is presented. This investigation: (1) conducted a generalizability (G) study to identify the sources of systematic and random measurement error (SRME); (2) used an analytic approach based on G theory to integrate measurement errors into subject sampling errors during statistical analyses of data from a multiple-baseline design; (3) compared this new approach with a conventional t-test to determine the extent to which conventional inferential statistics were inflated, increasing the probability of Type I errors in SSR; and (4) examined discrepancies among the dependability coefficients and the statistical tests of significance. The behavior of three preschool children with handicaps and six non-handicapped peer confederates, enrolled in an integrated preschool classroom in Pittsburgh (Pennsylvania), was observed during consecutive school days over the course of 5 months. A continuous observational recording procedure was used to code interactions among each of the three target children, their peer confederates, and the teacher. Examination of the grand means revealed that 10 behavior sequences did not occur. There were more interactions between the target children and peers, fewer interactions between the children and teachers, and fewer non-social utterances. Overall, 63% of the behavior sequences occurred less than 10% of the time during which reliability observations were made. Although the comparison between statistical procedures did not confirm that the G analytic approach would identify a significant number of Type I errors, the results confirm the direction of the effect proposed by H. K. Suen et al. (1990). SSR practitioners should: adopt a G approach to study the combined influence of SRME; report variance components and minimize the emphases on reliability coefficients; and investigate further applications of the G analytic approach. A 39-item list of references and six data tables are included. (RLC)

Applying Generalizability Theory to Evaluate Treatment Effect in Single-Subject Research

Daniel J. Lefebvre

University of Connecticut

Hoi K. Suen

Pennsylvania State University

Paper presentation at the Annual Meeting of the American Educational Research Association, Boston, MA., April, 1990

The conventional approach to educational and psychological research considers the stage of instrument development as distinct from procedures associated with the stage of data analysis (cf. Borg & Gall, 1989; Campbell & Stanley, 1966; Kerlinger, 1986). Typically, a method of data collection (e.g., observational procedure, test, questionnaire) is developed to measure certain behaviors or psychological constructs and the reliability and validity of the scores generated by this method of data collection are assessed. Having established that reliability is above a certain threshold, researchers then proceed with data analysis. Yet, at least two qualifications to this conventional approach can be proposed, particularly as it relates to single-subject research.

First, if, as is often the case, researchers fail to specify and investigate all of the relevant conditions of measurement, interpretation of both the estimates of reliability and the subsequent analyses of data is confounded. In principle, the reliability and validity of direct observation procedures are established through the use of a written behavior code and direct training of the observer (Baer, Wolf, & Risley, 1987). Yet, a great deal of confusion and controversy currently exists regarding what to report as evidence of the reliability of observational measures (cf. Suen, 1988). Increasingly, behavior analysts as well as psychometricians have recommended an intraclass correlation or generalizability approach for estimating the reliability of observational measures (Bakeman & Gottman, 1986; Berk, 1979; Cone, 1977; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Hartmann, 1982; Mitchell, 1979; Suen, 1988). By investigating all of the relevant sources of measurement error, the researcher is able to design a measurement procedure that minimizes error for a particular purpose (Rowley, 1989; Suen & Ary, 1989).

Secondly, measurement errors detected during the assessment of reliability typically are not considered explicitly during data analysis. Having established that reliability is above a certain threshold, data analysis proceeds under the assumption that data are *perfectly* reliable with *no* measurement error. This can lead researchers to conclude that a treatment effect exists when, in fact, the observed differences in scores may be an artifact of combined measurement error (i.e., Type I error is inflated). This problem can become particularly acute in applications of single-subject research and when multifaceted measurement procedures are used to gather data (Suen, Owen, Kehle & Campo, 1990).

Many researchers who employ observational measures within the context of single-subject research designs rely upon visual analysis of graphic presentations to analyze the effect of planned interventions on children with handicaps (Tawney & Gast, 1984). The reliability of visual inspection as a means of analysis, however, has been questioned repeatedly (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Gottman & Glass, 1978; Jones, Vaught & Weinrott, 1978; Wampold & Furlong, 1981b). Despite this criticism, few researchers have adopted statistical analyses as a supplement to, or replacement of, visual inspection. Proponents of single-subject research have argued that statistical procedures are not viable for single-subject research (Kazdin, 1980) or have dismissed such procedures as tactics which may or may not apply to direct observation (Baer, 1977; Baer, et al., 1987).

Since questions regarding visual inspection persist, the investigation of new methods of data analysis which take into account both measurement and subject sampling errors is warranted. While the generalizability theoretical framework (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a general conceptual mechanism for combining both sources of score variance, few researchers have investigated whether statistical extensions of the generalizability approach can be applied to analyze data from single-subject research designs. It would appear reasonable to follow the advice of Suen, et al. (1990) and carry forward the measurement error information gained during the assessment of reliability into the data analysis stage. Such a procedure may allow researchers to identify, and possibly reduce, Type I errors that occur when conventional inferential statistics are used in single-subject research. Additionally, this statistical procedure could be useful as a complement to visual inspection.

The primary goal of this study was to investigate empirically the methodological issues associated with evaluating treatment effect in single-subject research designs. This study (a) conducted a Generalizability (G) study to identify the sources of systematic and random measurement error, (b) utilized an analytic approach based upon generalizability theory (Suen, et al, 1990) to integrate measurement errors into subject (i.e., sampling of occasions) sampling errors during statistical analyses of data from a multiple-baseline AB design; (c) compared this new approach with a conventional t-test to determine the extent to which conventional inferential statistics were inflated, increasing the probability of Type I errors in single-subject research; and (d) examined discrepencies among the dependability coefficients and the statistical tests of significance. By incorporating systematic and random measurement error components into the computation of the standard error of the mean, it was anticipated that the method recommended by Suen, et al (1990) would provide a valid measure of treatmei.. effectiveness in single-subject research and call into question tie conventional distinction between reliability analysis and data analysis.

Conditions of the Investigation

This study represented a secondary analysis of observational data reported in a previous single-subject study which investigated the effects of self-monitoring on preschool children's use of social interaction strategies with their autistic peers (Sainato, Goldstein & Strain, 1989). These data were used in the current study to compare methods for analyzing single-subject data.

Specification of Subject Characteristics

Three preschool children with handicaps and six nonhandicapped peer confederates, enrolled in an integrated preschool classroom in Pittsburgh, Pennsylvania, were observed during consecutive school days over the course of five months. Each of the three children with handicaps was referred for services based upon a diagnosis of moderate autism, using the Childhood Autism Rating Scale (Schopler, Reicher, DeVellis, & Daly, 1980). These three boys, Jason, David and Bert, were 50, 56, and 43 months old, respectively, at the outset of the study. They we.e selected for participation because they exhibited low rates of interactions with peers, responded inconsistently or negatively to peer initiations, and did not direct their comments to peers. Jason often was repetitious in his interactions with peers. David used one- and two-word utterances and exhibited echolalia and a severe attention deficit. Bert displayed frequent tantrums. Three normally developing peers were nominated by the teacher as peer confederates. The two remaining nonhandicapped children in the class served as the second peer in each play group (i.e., triad).

Setting

Behavioral observations occurred during a sociodramatic play activity in which a child with handicaps and two nonhandicapped peer confederates participated for a seven minute period each day. Two teachers alternated the monitoring of play sessions. Five sociodramatic play activities (e.g., housekeeping, dress-up) were selected, and the amount of materials, their arrangement, and the scheduling (i.e. rotation) o' play activities remained constant over the course of the study.

Single-Subject Research Design

Sainato, et al. (1989) employed a multiple baseline design acrcss subjects to evaluate the effects of training in self-monitoring on the peers' use of facilitative, social interaction strategies with their autistic peers. The self-monitoring training package was implemented after an initial baseline and a second baseline that followed the teaching of the facilitative strategies. The conditions of teacher involvement specified that, during baselines and all subsequent interventions, the teacher (a) introduce the play activities and provide two or three general ideas on how to play with the available materials, (b) introduce the posters illustrating the four facilitative, social interaction strategies, and (c) monitor the activity, enforce classroom rules, and keep children in the play area.

Data Collection Techniques

A continuous observational recording procedure was used to code interactions among each of the three target children, their peer confederates, and the teacher. Live observations were conducted daily during the first five minutes of each play session following the teacher's introduction. The observational recording prccedure itself divided these five-minute samples into 30 ten-second intervals to assist reliability analyses and the analyses of sequential data. An audiotape marking the ten-second intervals cued the observers to change intervals.

Each interval consisted of five columns of subject codes, designating the target child (T), peers (P), or teacher (A) as the individual who was initiating or responding within an interaction sequence. Observers coded the occurrence/nonoccurrence of all behaviors directed to the target child, all behaviors directed by the target child to peers, and teacher verbal statements to both the target child and peer confederates in each interval. A total of thirty-three possible social interaction sequences were recorded. Interactive behaviors were coded sequentially. Data recorded through live observation were supplemented by the use of audiotapes of each play session. Observers were required to listen to the audio recordings independently before submitting the final coding for each sample.

Research Design

The observational data recorded by two observers during the reliability sessions in the Sainato et al. (1989) study were reanalyzed. Those scores were interpreted within a criterion-referenced framework to estimate the frequency of peer confederates' use of facilitative social interaction strategies, the frequency of children's social interactions, and the frequency of teacher prompts directed at peer confederates and target children. The object of measurement in these analyses, then, was defined as behavior in time (days). The observation schedule resulted in 16, 19 and 16 reliability sessions, respectively, for the three target children (Jason, David, and Bert) and their play groups. These reliability sessions represented 39% of the observational sessions conducted by Sainato, et al. (1989).

A number of decisions were made to control variables extraneous to this investigation. First, the reliability sessions associated with the second baseline phase were excluded from the analyses. The exact mechanism for integrating measurement error into data analyses for more complex designs (i.e., ANOVA for multiple phase periods) has yet to be developed, since it may require the pooling of variance estimates. In effect, this investigation focused on analyzing data from an AB multiple baseline design across three target children and their peers. Second, to reduce the possibility of treatment confounding the estimates of the variance components, the variance components associated with the baseline and treatment phases were computed separately. These estimates were based upon a sample of eight baseline days and five treatment days for Jason, seven baseline days and five treatment days for David, and seven baseline days and three treatment days for Bert. Third, to address the methodological issues related to data analysis, it was necessary to balance the number of observation sessions across phases. To achieve this balance, reliability sessions were randomly selected from the baseline phase. Calculation of the standard error of the mean terms and the test statistics (i.e., t-tests) were based upon balanced samples of five days for Jason and David, and three days for Bret, across the two phases.

Methods of Data Analyses

Variance components were computed based upon how Sainato et al. (1989) intended to summarize (i.e., molecular and molar behavior sequences) and report the data. In effect, Generalizability (G) studies were conducted for 19 behavior sequences that were associated with each target child and his play group. Given that the variance components and dependability coefficients (Φ) were calculated separately for baseline and treatment phases, a total of 114 scenarios was investigated. The variance components and Φ coefficients were generated via GENOVA, a specialized computer program for generalizability analyses (Crick & Brennan, 1983). Each Generalizability (G) study employed a 2-facet design, interval nested within days crossed with observers (D x I:D x O), to estimate the systematic error associated with intervals and with observers and to estimate random error. The generalizability phi coefficient, Φ, was defined by Brennan and Kane (1977a):

$$\Phi = \frac{\sigma_d^2}{\sigma_d^2 + \sigma^2(\Delta)}, \qquad (1)$$

where σ_d^2 is the true variance and $\sigma^2(\Delta)$ is the absolute error variance (Brennan, 1983; Suen & Ary, 1989). For the Decision (D) studies in these analyses, then,

$$\sigma^2(\Delta) = \sigma_o^2 + \sigma_{do}^2 + \sigma_{i:d}^2/30 + \sigma_{io:d}^2/30, \qquad (2)$$

where σ_{o}^2 is the observer main effect variance (systematic observer bias), σ_{do}^2 is the interaction between observers and days, $\sigma_{i:d}^2$ is the main effect variance associated with intervals nested within days, and $\sigma_{io.d}^2$ is the interaction between observers and intervals nested within days. The variance components associated with intervals were divided by the number of intervals (i.e., 30 intervals) to reflect the absolute error variance associated with the *average* interval score within sessions (i.e., days). The number of observers was defined as one in the D-study scenarios.

A conventional t-test was compared to an analytic approach based upon generalizability theory (Suen, et al., 1990). The descriptive statistics (i.e., phase means, standard error of the means, and variances) for the conventional t-test were computed via the SPSSx (3rd Edition, 1989) procedure CONDESCRIPTIVE. The standard error term for the generalizability analytic approach was based on information estimated via GENOVA (Crick & Brennan, 1983). The standard error terms for both approaches were computed based upon mean session scores, using only baseline data.

By convention, the standard error of the mean term that is used in computing the t-test statistic is based on data from only one observer. Usually, a primary observer is designated at the beginning of the study and the second observer is used for reliability checks. Yet, since one observer can be expected to show more (or less) consistency across days, the choice of observer can influence dramatically the test statistic. For example, if a researcher happens to use the observer with a higher observed standard error term, fewer significant results would be reported. The standard error term in GENOVA, however, is based upon data from two observers. To allow for a comparison between the two approaches, descriptive statistics were computed for both observers and the conventional t-test statistic was computed based upon an averaged standard error term.

In contrast to the conventional approach, the generalizability analytic approach recommended by Suen, et al., (1990) integrates measurement errors into subject (i.e., sampling of occasions) sampling errors during statistical analyses of data. The standard error of the mean term, $\sigma(X)$, can be derived from the error variance associated with the observed mean score. For the purpose of this study, the value of $\sigma(X)$ for the observed mean session score was derived from the following equation (cf. Brennan, 1983; Suen, et al., 1990):

$$\sigma(X) = N_i \left[\frac{\sigma_d^2}{N_d} + \frac{\sigma_{i.d}^2}{N_i N_d} + \frac{\sigma_o^2}{N_o} + \frac{\sigma_{do}^2}{N_d N_o} + \frac{\sigma_{io:d}^2}{N_i N_o N_d} \right]^{½} \qquad (3)$$

where σ_d^2 is the main effect variance associated with the object of measurement (days), $\sigma_{i:d}^2$ is the main effect variance associated with the nested term, intervals nested within days, σ_o^2 is the observer main effect variance, σ_{do}^2 is the interaction between days and observers, and $\sigma_{io:d}^2$ is the interaction between observers and intervals nested within days. The mean error variance, $\sigma^2(X)$, the term located within the brackets above, was estimated directly via GENOVA by specifying a D-study with the object of measurement (days) identical to the sample examined within the G-study (e.g., 5 days for Jason) and the observer facet equal to one. This standard error of the mean term, $\sigma(X)$, was then used as the denomi!ator to compute a t-test statistic.

This generalizability test statistic was compared with the conventional t-test statistic to determine whether there were significant differences among the results after measurement error was taken into account. The degrees of freedom associated with the significance tests were calculated using the number of days in baseline (N_1) and treatment (N_2) phases ($df = N_1 + N_2 - 2$). Given the directional alternative hypotheses proposed by Sainato, et al. (1989), the critical values for a one-tailed test of significance were applied to the test statistics. This comparison served to estimate the extent to which conventional inferential statistics were inflated, increasing the probability of Type I errors in single-subject research. Finally, these statistical tests of significance and the dependability (Φ) coefficients were examined to determine whether any discrepencies existed among the results.

Results

Examination of the grand means revealed that a total of ten behavior sequences did not occur at all during either baseline or treatment phases. Examination of the grand means also described two important conditions of this investigation. First, in most cases, the behavior sequences targeted for intervention changed in a manner anticipated by Sainato, et al. (1989). Specifically, during the treatment phase, there were more interactions between the target children and peers, fewer interactions between the children and the teachers, and fewer nonsocial utterances. Second, the majority of behavior sequences occurred at an extremely low rate. Overall, 63% of the behavior sequences occurred less than ten percent of the time during which reliability observations were being conducted. This condition was even more pronounced during the baseline phase when 68% of the behavior sequences occurred less than ten percent of the time. Of the 18 behavior sequences which occurred more frequently (i.e., $.10 \leq X < .45$), six involved nonsocial utterances on the part of the children and five involved interactions with teachers.

Examining the Variance Components

Examination of the variance components provided evidence of considerable measurement error in observers' scores, overall. By providing direct estimates of the sources of score variation, the Generalizability (G) Studies determined that, in most cases, (a) systematic error associated with intervals nested in days was substantial, (b) observer bias was negligible, (c) random measurement error associated with observers' performance across days (DO) was minimal, and (d) random measurement error (IO:D) was substantial. Additionally, true score variance (i.e., behavior across days) was sufficiently large relative to the facets, intervals in days (I:D) and observers (O), and their interactions in only 14% of the scenarios. The Φ coefficients expressed the relative magnitude of this relationship between the true score and error measurement terms, indicating that the scores of *any* one observer, using this data collection design and receiving similar training, would be highly dependable ($\Phi > .75$) for 16 of the 114 behavior sequences. Observers' scores were characterized by less measurement error during the treatment phase than during the baseline phase.

Comparing Methods for Evaluating Treatment Effect

Tables 1, 2, and 3 present the standard error of the mean terms associated with the behavior sequences observed during play activities with Jason, David, and Bert, respectively. The standard error terms reflect measurement error associated with the mean session scores during the baseline phase. The standard error of the mean estimated via a generalizability approach, the conventional standard error terms associated with the observations of observer 1 and observer 2, and an averaged conventional standard error of the mean term are reported for comparison.

Insert Tables 1, 2, & 3 about here

Tables 4, 5, and 6 present the mean session scores and the t-test statistics associated with the behavior sequences observed during play activities with Jason, David, and Bert, respectively. The mean session scores were based upon five sessions, five sessions, and three sessions, respectively, for Jason, David, and Bert. The conventional t-test statistics were computed using the averaged conventional standard error of the mean terms. The generalizability approach used the standard error terms estimated via GENOVA to compute the test

Insert Tables 4, 5, & 6 about here

statistics. The critical values ($df = 8$) for a one-tailed test of significance were 1.860 ($p < .05$), 2.896 ($p < .01$), and 3.355 ($p < .005$) for the test statistics associated with the behavior sequences cf Jason and David. The critical values ($df = 4$) for a one-tailed test of significance were 2.132 ($p < .05$), 3.747 ($p < .01$), and 4.604 ($p < .005$) for the test statistics associated with the behavior sequences of Bert.

Discussion

To determine whether there were significant differences between the methods of data analysis after measurement error was taken into account, a generalizability analytic approach (Suen et al., 1990) was compared with a conventional t-test. The basis for this comparison was in the calculation of the standard error of the mean terms. Tables 1, 2, and 3 indicated considerable differences between the standard error terms associated with scores from observer 1 and observer 2. In 35% of the scenarios, the standard error of the mean for observer 1 was larger than the standard error of the mean for observer 2. In 30% of the scenarios, the standard error terms for observer 2 were larger. In 35% of the scenarios, the standard error terms for each observer were the same. Had the researcher adopted the scores and standard error terms of observer 2 for data analysis, for example, rather than those of observer 1, different conclusions regarding the effectiveness of treatment may have resulted. It was apparent that each of the conventional standard error terms represented less stable estimates because they were based upon less data.

Since the generalizability standard error term accounted for all measurement error in the data collection design, this term would be expected to be larger or equal to the averaged conventional standard error term. This expectation was confirmed in all but seven scenarios. The discrepancies in these seven scenarios, however, could be explained by rounding errors. By taking into account absolute (i.e., systematic) error, the generalizability approach provided evidence of the *accuracy* of the *exact* differences between phase means. In contrast, the conventional t-test provided evidence of the *relative* differences between phase means. These results provided evidence that the generalizability standard error term provided a more stable estimate of error around the mean.

The test statistics in Tables 4, 5, and 6 indicated that significant treatment effects were realized for 25 behavior sequences (44%) when the conventional t-test was used to analyze data and for 22 behavior sequences (39%) when the generalizability analytic approach was used to analyze data. The generalizability analyt'c approach identified only three scenarios where Type I error rates (i.e., rejecting a true null hypothesis) were inflated by the use of the averaged conventional standard error of the mean term. In three other scenarios, inclusion of measurement error resulted in lower levels of significance.

These results failed to confirm one of the proposed advantages of the generalizability analytic approach over a conventional t-test approach. Suen, et al. (1990) had anticipated that the inclusion of measurement error during the data analysis stage would have resulted in a significant number of discrepancies between the two approach. Yet, discrepancies were observed in onlv six of fifty-seven scenarios. In effect, the magnitudes of the treatment effects reported by Sainato, et al. (1989) were substantial *relative to combined measurement error*. One likely explanation for the failure to confirm the advantages of the generalizability analytic approach is that the main effect observer variances were uniformly extremely small in all cases.

The finding that treatment was still judged to be effective, *even in the presence of considerable measurement error*, was corroborated further when the discrepancies among the dependability coefficients and statistical tests of significance were examined. Only 14%, 21%, and 29% of the Φ coefficients that were associated with significant treatment effects met a 0.75 standard of reliability across the play groups of Jason, David, and Bert, respectively. Such discrepancies may account for some of the difficulties associated with visual analysis.

Yet, such discrepencies may also pose serious questions regarding the generalizability of scores (i.e., treatment effect). The persistent presence and magnitude of random observation error in the Sainato, et al. (1989) study did call attention to the adequacy of their data collection design. A number of potential sources of score variability were not accounted for within this data collection design. Specifying the variance across two teachers, the variance across six play activities, and/or the variance across peer pairings as facets may have further clarified the interpretation and generalizability of the scores (cf. Cone, 1977; McGaw, et al., 1972).

Statistically significant results were reviewed in light of their corresponding dependability coefficients in order to investigate whether the conventional distinction between reliability analysis and data analysis was justified. Clearly, the practical advantages of variance components for interpreting measurement error, in effect, minimize the utility of reliability coefficients. Additionally, by providing a mechanism for integrating measurement error into the analysis of data, the generalizability analytic approach allowed for the judgement of treatment effectiveness independent of sampling and measurement fluctuations. When statistical significance is found with the generalizability approach, one can ascertain that the treatment is effective, in spite of sampling and measurement errors.

Conclusions

Although the comparison between statistical procedures failed to confirm that the generalizability analytic approach would identify a significant number of Type I errors, results did confirm the direction of the effect proposed by Suen, et al. (1990). Given that the generalizability analytic approach effectively integrated measurement error into data analysis, it would appear to represent a more appropriate application of inferential statistics to single-subject research. In other situations, where the magnitude of the treatment effect is less substantial and systematic error variance is larger, the theoretical advantages of the generalizability analytic approach may be further substantiated.

Yet, a number of qualifications should be made regarding the findings of this investigation. The first qualification relates to a limitation inherent to statistical tests of significance. Statistical tests of significance on the differences between phase means do not allow for the assessment of trends in the data. As such, it would be inappropriate to consider them as substitutes for visual analysis. Yet, statistical extensions of the generalizability approach may serve as an effective complement to visual analysis techniques by providing a measure of the accuracy of exact differences between phase means. Such procedures may provide further support for conclusions in single-subject research.

A second qualification relates to the issue of serial dependency. For the purpose of this study, the issue of serial dependency was not addressed. In this investigation, only scores from two observers during the reliability sessions were used for evaluating treatment effect. Given the small number of observation sessions within phases (i.e., five sessions), the existence of serial dependence cannot be assessed meaningfully due to a lack of power. The relatively random distribution of reliability checks within phases suggested that it was unlikely that serial dependency adversely affected the estimates of variance (Rowley, 1989). Such reliability checks more accurately represented probes (Bakeman & Dorval, 1989). By conducting more reliability checks and distributing them proportionally within and across treatment phases, researchers may be able to generate scores for evaluating treatment effect that are more dependable and less subject to bias. Yet, the nature and implications of serial dependency in behavioral observation data remain unclear (Bakeman & Dorval, 1988; Huitema, 1985; Rowley, 1989; Suen, 1987).

A third qualification relates to the need for a criterion for behavior stability within phases. To date, there has been little consensus on a criterion for behavior stability within phases. Given the instability in the conventional standard error terms, it would appear more appropriate to consider the generalizability standard error term as a criterion for behavior stability within phases. By considering the distribution of error around the mean within each phase, researchers can more accurately identify and interpret outliers in the data.

The results of this investigation support the recommendations that single-subject researchers should (a) adopt a generalizability approach to investigate the combined influence of systematic and random measurement error, (b) report variance components and minimize the emphases on reliability coefficients, and (c) investigate further applications of the generalizability analytic approach. It would appear that the methodological problems associated with reliability and data analysis in single-subject research will only be exaggerated as more independent variables are compared simultaneously (i.e., treatment packages) and multielement designs (i.e., alternating treatments or simultaneous treatments) are used. The desirable effects of such multielement designs (cf. Higgins-Hains & Baer, 1989) could be confounded unless issues related to measurement error are addressed.

References

Baer, D. M. (1977a). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis, 10,* 117-119.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 20*(4), 313-327.

Bakeman, R., & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis.* London: Cambridge University Press.

Bakeman, R., & Dorval, B. (1989). The distinction between sampling independence and empirical independence in sequential analysis. *Behavioral Assessment, 11,* 31-37.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*(5), 460-472.

Berk, R. A. (1984). Selecting the index of reliability. In R.A. Berk (Ed.), *A guide to criterion-referenced test construction* Baltimore, MD: The John Hopkins University Press.

Borg, W. R., & Gall, M. D. (1989). *Educational research* (5th ed.). New York: Longman.

Brennan, R. L. (1980). Estimating the dependability of the scores. In R.A. Berk (Ed.), *A guide to criterion-referenced test construction.* Baltimore, MD: The John Hopkins University Press.

Brennan, R. L. (1980). Applications of generalizability theory. In R.A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The John Hopkins University Press.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: ACT Publications.

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277-289.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Boston, MA: Houghton Mifflin.

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

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (ACT Technical Bulletin No. 43). Iowa City, IA: The American College Testing Program.

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

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intersubject data. *Journal of Applied Behavior Analysis, 12,* 573-579.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of observational measures. *Review of Educational Research, 48,* 157-184.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15,* 415-421.

Hartmann, D. P. (1982). Assessing the dependability of observational data. In D.P. Hartmann (Ed.), *Using observers to study behavior.* San Francisco: Jossey-Bass.

Herbert, J., & Attridge, C. (1975). A guide for developers and users of observations systems and manuals. *American Educational Research Journal, 12*(1), 1-20.

Higgins-Hains, A., & Baer, D. M. (1989). Interaction effects in multielement designs: Inevitable, desirable, and ignorable. *Journal of Applied Behavior Analysis, 22*(1), 57-69.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*(2), 107-118.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 1,* 105-126.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics, 5,* 253-260.

Kerlinger, F. N. (1986). *Foundations of behavior research* (3rd ed.). New York: Holt, Rinehart, and Winston.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observational schemes: Where are the errors? *American Educational Research Journal, 9,* 13-27.

Mitchell, S. K. (1979). Interobserver agreement, reliability and generalizability of data collected in observational studies. *Psychological Bulletin, 86*(2), 376-390.

Rowley, G. L. (1986, April). *Application of generalizability theory to observational studies: Limitations.* Paper presented at the annual meeting of the American Eduational Research Association, San Francisco, CA.

Rowley, G. L. (1989). Assessing error in behavioral data: Problems of sequencing. *Journal of Educational Measurement, 26*(3), 273-284.

Sainato, D. M., Goldstein, H., & Strain, P. S. (1989). *Effects of self-monitoring on preschool children's use of social interaction strategies with their autistic peers.* Manuscript submitted for publication.

Schopler, E., Reicher, R. J., DeVellis, R. F., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood autism rating scale. *Journal of Autism and Developmental Disorders, 10,* 91-103.

SPSS^{x}, Inc. (1989). *SPSS^{x} user's guide* (3rd ed.). Chicago, IL: Author.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavioral analysis. *Behavioral Assessment, 9*(2), 113-124.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10*(4), 1-22.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Suen, H. K., Owen, S. V., Kehle, T., & Campo, T. (1990). *Integrating measurement errors in data analysis.* Manuscript submitted for publication.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education.* Columbus, OH: Charles E. Merrill.

Wampold, B. E., & Furlong, M. J. (1981b). The heuristics of visual inspection. *Behavioral Assessment, 3,* 79-92.

Table 1

Comparison of Standard Error of the Mean Terms for Baseline Data

Target	Behavior	Estimated	Conventional Terms		
Child	Sequence	via GENOVA[a]	Observer 1	Observer2	Averaged
Jason	TInit - P_1	.315	.200	.400	.300
	TResp - P_1	.329	.245	.400	.322
	TInit - P_2	.355	.245	.245	.245
	TResp - P_2	.474	.400	.447	.423
	T Nonsocial	1.893	2.015	1.749	1.882
	P_1AG - T	.190	.200	.200	.200
	P_1PO - T	.190	.200	.200	.200
	P_1S - T	.000	.000	.000	.000
	P_1R - T	.190	.200	.200	.200
	P_1O - T	.391	.374	.374	.374
	P_1 Nonsocial	.585	.510	.632	.571
	P_1 - P_2	2.353	2.530	2.000	2.265
	P_2Init - T	.585	.583	.583	.583
	P_2Resp - T	.134	.200	.000	.100
	P_2 - P_1	2.720	2.478	2.088	2.283
	P_2 Nonsocial	1.138	1.000	1.068	1.034
	P_1 - Adult	.402	.400	.400	.400
	P_2 - Adult	1.061	1.049	1.049	1.049
	Adult Init	1.106	.316	.678	.497

[a] A GENeralized Analysis Of VAriance System (Crick & Brennan, 1983).

Table 2

Comparison of Standard Error of the Mean Terms for Baseline Data

Target	Behavior	Estimated	Conventional Terms		
Child	Sequence	via GENOVA[a]	Observer 1	Observer2	Averaged
David	TInit - P_1	.329	.245	.400	.322
	TResp - P_1	.958	.316	1.319	.817
	TInit - P_2	.000	.000	.000	.000
	TResp - P_2	.895	1.200	.400	.800
	T Nonsocial	2.902	3.010	2.786	2.898
	P_1AG - T	.402	.200	.200	.200
	P_1PO - T	1.613	.927	2.083	1.505
	P_1S - T	.895	.490	.860	.675
	P_1R - T	.285	.245	.200	.222
	P_1O - T	.134	.000	.200	.100
	P_1 Nonsocial	1.026	1.122	.917	1.019
	P_1 - P_2	2.257	2.387	2.059	2.223
	P_2Init - T	2.345	2.804	1.530	2.167
	P_2Resp - T	.251	.245	.200	.222
	P_2 - P_1	2.627	2.728	2.249	2.488
	P_2 Nonsocial	1.464	1.772	1.068	1.420
	P_1 - Adult	.134	.000	.200	.100
	P_2 - Adult	1.004	1.049	.894	.971
	Adult Init	1.689	1.625	1.720	1.672

[a] A GENeralized Analysis Of VAriance System (Crick & Brennan, 1983).

Table 3

Comparison of Standard Error of the Mean Terms for Baseline Data

Target	Behavior	Estimated	Conventional Terms		
Child	Sequence	via GENOVA[a]	Observer 1	Observer2	Averaged
Bert	TInit - P_1	.329	.333	.333	.333
	TResp - P_1	1.509	1.333	1.667	1.500
	TInit - P_2	.329	.333	.333	.333
	TResp - P_2	.000	.000	.000	.000
	T Nonsocial	1.269	1.202	1.333	1.267
	P_1AG - T	.664	.667	.667	.667
	P_1PO - T	.765	.577	.667	.622
	P_1S - T	1.375	1.000	1.667	1.333
	P_1R - T	.329	.333	.333	.333
	P_1O - T	.000	.000	.000	.000
	P_1 Nonsocial	2.687	2.028	1.000	1.514
	P_1 - P_2	1.607	2.028	1.000	1.514
	P_2Init - T	.944	.882	.882	.882
	P_2Resp - T	.000	.000	.000	.000
	P_2 - P_1	1.287	1.333	.882	1.107
	P_2 Nonsocial	1.981	1.453	1.000	1.226
	P_1 - Adult	.577	.577	.577	.577
	P_2 - Adult	1.394	1.528	1.202	1.365
	Adult Init	1.547	1.732	1.202	1.467

[a] A GENeralized Analysis Of VAriance System (Crick & Brennan, 1983).

Table 4

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
Jason	TInit - P_1	.200	1.200	3.333**	3.175**
	TResp - P_1	.400	9.000	26.708***	26.140***
	TInit - P_2	.600	.600	.000	.000
	TResp - P_2	.600	1.800	2.837*	2.532*
	T Nonsocial	5.400	7.000	.850	.845
	P_1AG - T	.200	1.800	8.000***	8.421***
	P_1PO - T	.200	2.200	10.000***	10.526***
	P_1S - T	.000	10.400	----[b]	----[b]
	P_1R - T	.200	.800	3.000**	3.158**
	P_1O - T	.800	1.200	1.069	1.023
	P_1 Nonsocial	1.600	1.400	-.350	-.342
	P_1 - P_2	9.000	4.000	-2.207*	-2.125*
	P_2Init - T	1.200	3.600	4.117***	4.103***
	P_2Resp - T	.200	1.200	10.000***	7.463***
	P_2 - P_1	8.800	4.600	-1.840	-1.544
	P_2 Nonsocial	3.000	3.600	.580	.527

(table continues)

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
	P_1 - Adult	.600	.000	-1.500	-1.493
	P_2 - Adult	2.000	.000	-1.907*	-1.885*
	Adult Init	8.000	4.400	-7.243***	-3.255**

Note. Table reports mean session scores for five sessions (i.e., days). GENOVA is the GENeralized Analysis Of VAriance System developed by Crick & Brennan, (1983). The degrees of freedom associated with the significance tests were calculated using the number of sessions (days) in baseline (N_1) and treatment (N_2) phases ($df = N_1 + N_2 - 2$).

[a] Computed using Expected (Averaged) Standard Error of the Mean for one observer. [b] Since the behavior sequence did not occur during the baseline phase, standard error terms and test statistics could not be computed.

*$p < .05$. **$p < .01$. ***$p < .005$.

Table 5

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
David	TInit - P_1	.400	.600	.621	.607
	TResp - P_1	1.000	2.000	1.224	1.044
	TInit - P_2	.000	.000	.000	.000
	TResp - P_2	1.200	.200	-1.250	-1.117
	T Nonsocial	5.600	11.800	2.139*	2.136*
	P_1AG - T	.800	2.400	8.000***	3.980***
	P_1PO - T	1.600	4.800	2.126*	1.984*
	P_1S - T	1.200	3.600	3.556***	2.682*
	P_1R - T	.600	.200	-1.802	-1.403
	P_1O - T	.000	.200	2.000*	1.492
	P_1 Nonsocial	2.600	7.400	4.710***	4.678***
	P_1 - P_2	8.000	7.600	-0.180	-0.177
	P_2Init - T	4.600	2.000	-1.200	-1.109
	P_2Resp - T	.400	.000	-1.802	-1.594
	P_2 - P_1	9.200	12.400	1.286	1.218
	P_2 Nonsocial	4.200	2.000	-1.549	-0.736

(table continues)

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
	P_1 - Adult	.000	.200	2.000*	1.492
	P_2 - Adult	4.000	1.600	-1.236	-1.195
	Adult Init	10.200	10.000	-0.120	-0.118

Note. Table reports mean session scores for five sessions (i.e., days). GENOVA is the GENeralized Analysis Of VAriance System developed by Crick & Brennan, (1983). The degrees of freedom associated with the significance tests were calculated using the number of sessions (days) in baseline (N_1) and treatment (N_2) phases ($df = N_1 + N_2 - 2$).

[a] Computed using Expected (Averaged) Standard Error of the Mean for one observer.

*p < .05. **p < .01. ***p < .005.

Table 6

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
Bert	TInit - P_1	.333	.333	.000	.000
	TResp - P_1	1.333	5.333	2.667*	2.651*
	TInit - P_2	.333	.000	-1.000	-1.003
	TResp - P_2	.000	.000	.000	.000
	T Nonsocial	1.667	6.000	3.420*	3.414*
	P_1AG - T	.667	3.000	3.498*	3.513*
	P_1PO - T	2.000	7.333	8.574***	6.971***
	P_1S - T	1.000	8.000	5.251***	5.091***
	P_1R - T	.333	.333	.000	.000
	P_1O - T	.000	.333	----[b]	----[b]
	P_1 Nonsocial	5.333	2.333	-1.981	-1.116
	P_1 - P_2	8.667	9.667	.660	.622
	P_2Init - T	1.333	.667	-0.775	-0.705
	P_2Resp - T	.000	.000	.000	.000
	P_2 - P_1	5.667	10.000	3.914**	3.367*
	P_2 Nonsocial	5.333	2.333	-2.447*	-1.514

(table continues)

Comparison of Conventional T-Test with Generalizability Approach

Target	Behavior	Mean Session Scores		Test Statistics	
Child	Sequence	Baseline	Treatment	Conventional[a]	GENOVA
	P_1 - Adult	1.000	.667	-0.423	-0.423
	P_2 - Adult	3.000	1.000	-1.465	-1.435
	Adult Init	9.000	7.333	-1.136	-1.078

Note. Table reports mean session scores for three sessions (i.e., days). GENOVA is the GENeralized Analysis Of VAriance System developed by Crick & Brennan, (1983). The degrees of freedom associated with the significance tests were calculated using the number of sessions (days) in baseline (N_1) and treatment (N_2) phases ($df = N_1 + N_2 - 2$).

[a] Computed using Expected (Averaged) Standard Error of the Mean for one observer. [b] Since the behavior sequence did not occur during the baseline phase, standard error terms and test statistics could not be computed.

*$p < .05$. **$p < .01$. ***$p < .005$.