INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

University Microfilms International
300 North Zeeb Road
Ann Arbor, Michigan 48106 USA
St. John's Road, Tyler's Green
High Wycombe, Bucks, England HP10 8HR

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

RELIABILITY THEORY FOR TEACHER EVALUATIONS: SOME PARTIAL REPLICATION METHODS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

JAMES MANUEL PRICE

Norman, Oklahoma

1977

RELIABILITY THEORY FOR TEACHER EVALUATIONS:

SOME PARTIAL REPLICATION METHODS

A DISSERTATION

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

BY

Roger L. Mellgren (On Sabbatical Leave)

DISSERTATION COMMITTEE

ACKNOWLEDGEMENTS

I would like to express my appreciation to Dr. Alan Nicewander and Dr. Larry Toothaker, co-chairmen of my committee, for supporting and encouraging me, and for giving freely of their time, knowledge, and friendship. I would also like to thank the other members of my committee, Dr. Jack Kanak, Dr. Charles Gettys, and Dr. Roger Mellgren, for their contributions to my professional development.

For their questions, their confidence in me, and their friendship, I want to thank the graduate students of the Department of Psychology.

I can never fully express my gratitude for the help, hope, and happiness given me by my parents, my brother, my sister, and all my relatives.

Above all, I want to express my love and appreciation to my wife, Paula, whose patience, understanding, and love have made life meaningful and all goals attainable.

TABLE OF CONTENTS

Abstract

Two methods are developed to deal with the problem of unequal class frequencies when the reliability of the mean rating on a single evaluation item is desired. Following the derivation of the linear models and the accompanying estimation methods, simulation results are presented, comparing the proposed methods with methods based on true replication of the evaluation procedure.

RELIABILITY THEORY FOR TEACHER EVALUATIONS: SOME PARTIAL REPLICATION METHODS

Because student evaluations of teachers may play an important part in reaching decisions on promotion, tenure, and salary increases, such evaluations should be examined for reliability and validity as measurement instruments. Evaluation reliability, the subject of this paper, deals with the determination of the extent to which ratings given an instructor reflect true abilities rather than evaluation "noise." Studies of evaluation reliability have usually been conducted in the same manner as other measurement techniques. Costin (1968, 1971) used the test-retest coefficient of stability with factor scores derived from evaluations given at two points in the same semester and on factor scores from successive semesters. Somers and Southern (1974) computed estimates of internal consistency by means of coefficient alpha and average item intercorrelations.

However, several factors distinguish the assessment of the reliability of teacher evaluations from other measurement situations. First, since the amount of time allotted for completing an evaluation is normally limited to part of one class period, it is not desirable to increase the number of items in order to increase the reliability of the entire instrument. Thus the reliability of individual items is

of critical importance. Second, instructors are usually compared with their colleagues on the basis of the mean (or median) rating received on each item, rather than on the basis of the ratings given by individual students. Third, except in special circumstances, evaluations are generally conducted only once a semester. Multiple administrations of evaluations usually require obtaining the permission of instructors and administration, which may lead to use of a biased sample. In some cases, evaluations are conducted more than once in order to give instructors "feedback" about their teaching, in which case a coefficient of stability would be inappropriate.

We are therefore interested in determining the reliability of a single average score received on an item administered only once. The classical methods of test-retest and parallel forms require more class time to be sacrificed and may be contaminated by memory or fatigue effects. Furthermore, Lord and Novick (1968, Ch. 7) have pointed out the inadequacies of these methods when class sizes are moderate or when it is not reasonable to assume parallelism of the raters.

Two theories that lend themselves to the evaluation situation as described are generalizability theory (see, e.g., Cronbach, Rajaratnam, and Gleser, 1963) and generic true score theory (Lord and Novick, 1968, Ch. 7). These equivalent theories generate estimation methods based upon the random selection of instructors, items, subject matter, etc., to serve as levels of random factors in an analysis of variance (ANOVA) design. The meaning of reliability in such a setting is our ability to generalize the results of the evaluation to larger populations from which we have sampled our effects.

For such a formulation, we can define several different "true" scores for each observed rating by defining a true score as the expected value of the observed score across certain of the populations. For instance, the true score for instructor j on item k of an evaluation is the expected value of the observed rating of instructor j on item k, with the expectation taken across the populations of student raters, subject matter, etc. That is, the universe score of generalizability theory or the true score of generic true score theory corresponds to a main effect mean or a cell mean of the ANOVA framework.

Furthermore, for each of the ways in which we can define a true score, we can define a reliability coefficient as the intraclass correlation coefficient associated with the effects of interest. The intraclass correlation coefficient, as a ratio of variance components, tells us the extent to which observed score variability is due to the variability of the effects in which we are interested, and is thus essentially equivalent to the reliability coefficient of classical true score test theory.

For the situation under discussion, that of determining the reliability of a single item, a one-way random ANOVA would be used with instructors as the single main effect and student raters nested within instructor "conditions". Implementation of this design would require random selection of a number of instructors, such that no instructor is selected more than once and, ideally, such that no student is in more than one of the associated classes. The mean square between instructors and the mean square within instructors are used to form estimates of true score variance and observed score variance, and the

ratio of these estimates is used as an estimate of the intraclass correlation, or reliability, coefficient.

From the computational point of view, the simplest design is one with equal numbers of students rating each instructor. This condition will rarely be met in practice if instructor-class combinations are indeed randomly sampled for the estimation procedure. One instructor may be evaluated by a five-person graduate seminar, while another is evaluated by a 500-student survey course. Even if instructors are sampled from categories such as "large lecture," "small lecture," or "laboratory course," and a reliability coefficient is computed for each category, there will still be some disparity in the number of students rating each instructor. Class mean ratings will therefore be computed on different numbers of ratings, and the derivations of the estimates of variance components is no longer the simple procedure outlined in standard texts (e.g., Scheffé, 1959, p.228; Winer, 1971, p. 286).

Suppose four instructors are selected for the purposes of estimating item reliability. If their classes are included in a category such as "small lecture," the number of students in the classes will probably be comparable but not necessarily equal. Class sizes might be 10, 24, 30, and 40. Generalizability theory and generic true score theory have been developed, for the most part, on the basis of equal class sizes. Should we therefore discard ratings from the three larger classes in order to have 10 ratings per instructor? To do so means ignoring 62% of the information gathered; it would be preferable to find some way to use all the data.

Two recent studies attempted to use generalizability theory

to assess evaluation reliability. Doyle and Whitely (1974), in an equivocal analysis, seem to have circumvented the problem of unequal class sizes by not reporting them, while Kane, Gillmore, and Crooks (1976) discarded data from the sampled classes in order to balance the design. Neither of these approaches may be viewed as exemplary of the use of the generic true score-generalizability methodology.

While directing their considerations mainly to the balanced ANOVA designs, Cronbach, Gleser, Nanda, and Rajaratnam (1972, pp. 207-208) acknowledge the problem of unequal class frequencies and suggest that, rather than discard information, an unbalanced ANOVA design be used, following the work of Graybill (1961). In making this recommendation, Cronbach et al. do not indicate that the resulting intraclass correlation estimate is based on a single rating and should be adjusted by the Spearman-Brown formula to reflect the reliability of the class mean rating. Nor do they give the form of the population reliability coefficient being estimated. The present paper will determine both the form of the population coefficient and the method for adjusting the reliability of a single rating to reflect the reliability of the class mean rating.

In addition, we will investigate the adequacy of another method of dealing with the problem of unequal class sizes. Suppose we balance the ANOVA design by randomly dividing each class in half, compute the mean rating for each class-half, and then use the class-half mean ratings as the observations in the reliability estimation process. Again, the Spearman-Brown formula would be used to reflect the reliability of the mean rating for the whole class. Such a method would leave the

mean rating for each instructor unchanged, provide an equal number of ratings for each instructor, and base the reliability not on single ratings, but on mean ratings for each half of each class.

In each of these methods we are estimating the performance of actual replication of the mean rating by using either a mean based on half as many ratings or by a mean based on only one rating. Use of the class-half method is the closer approximation, intuitively, and will be referred to as the "first-order partial replication of the mean," since each half of each class serves as a partial replicate of the performance of the entire class. The method using the reliability of a single rating is less closely related to the performance of the whole class and will be referred to as the "second-order partial replication of the mean."

Two differences between the partial replication methods are immediately obvious. Based on more information, the first-order method uses "observations" with a smaller error variance than those used in the second-order method. However, the first-order method uses fewer "observations" per instructor in the ANOVA-based estimation process than does the second-order method. The present paper will examine the effects of these differences on the performance of the two methods of estimation.

Linear Models

Suppose that J instructors are selected for the purpose of estimating the reliability of a given item on an evaluation instrument. Suppose that instructor j is rated by n_j students, and let X_{ij} be the rating given instructor j by student i. Let ζ_j be the mean rating

received by instructor j over the population of student raters, i.e., $\zeta_j = E_i(X_{ij})$. The quantity ζ_j, called the universe score by Cronbach et al. (1963) or the generic true score by Lord and Novick (1968), is a random variable over instructors and is closely related to the true score of classical test theory. The observed rating X_{ij} can be expressed as

$$X_{ij} = \zeta_j + \varepsilon_{ij} \tag{1}$$

where $\varepsilon_{ij} = X_{ij} - \zeta_j$ is the residual or "generic error of measurement." The corresponding linear model for the one-way random ANOVA is

$$X_{ij} = \mu + a_j + e_{ij} \tag{2}$$

where μ is the expected rating over raters and instructors, a_j is the deviation $\mu_j - \mu$ of the expected rating of instructor j from the overall mean, and e_{ij} is the ANOVA residual or error term. For fixed j and the usual ANOVA assumption of zero expectation of the errors, the true score ζ_j is given in ANOVA terms as

$$\zeta_j = E_i(X_{ij}) = E_i(\mu + a_j + e_{ij}) = \mu + a_j = \mu_j. \tag{3}$$

Thus ζ_j of the true score model equals μ_j of the ANOVA model, and hence $\varepsilon_{ij} = e_{ij}$. In both models, the error term includes all discrepancies between the observed rating and its expected value, including any interactions between instructor and item, rater and instructor, etc.

The coefficient of reliability (or generalizability or generic reliability) for a single rating, $\rho^2(X,\zeta)$, is the proportion of the observed rating variance $\sigma^2(X)$ that is linearly predictable from the true scores, i.e.,

$$\rho^2(X,\zeta) = \sigma^2(\zeta)/\sigma^2(X) \qquad (4)$$

where $\sigma^2(\zeta)$ is the variance of the true scores ζ_j. For true scores and error scores uncorrelated, (4) can be written as

$$\rho^2(X,\zeta) = \sigma^2(\zeta)/\{\sigma^2(\zeta) + \sigma^2(\varepsilon)\}, \qquad (5)$$

where $\sigma^2(\varepsilon)$ is the variance of the generic errors, ε_{ij}.

Using the ANOVA model for X_{ij}, we can express the reliability $\rho^2(X,\zeta)$ as the intraclass correlation coefficient for the random instructor effect, i.e.,

$$\rho^2(X,\zeta) = \sigma^2(a)/\{\sigma^2(a) + \sigma^2(e)\} \qquad (6)$$

where $\sigma^2(a)$ is the variance of the instructor effects a_j, and $\sigma^2(e)$ is the variance of the residuals e_{ij}.

The formula for the intraclass correlation coefficient given in (6) is the reliability for a single rating if the errors e_{ij} have equal variances. The reliability of the class mean rating, $\rho^2(\bar{X},\zeta)$, is found by replacing the variance of e_{ij} in (6) with the variance of the mean error for a class, $\bar{e}_{\cdot j}$, i.e., by replacing $\sigma^2(e)$ by $\sigma^2(\bar{e}_j) = \sigma^2(e)/n_j$ to give

$$\rho^2(\bar{X},\zeta) = \sigma^2(a)/\{\sigma^2(a) + \sigma^2(\bar{e}_j)\} \qquad (7)$$

$$= \sigma^2(a)/\{\sigma^2(a) + \sigma^2(e)/n_j\} \; . \qquad (8)$$

If $n_1 = n_2 = \cdots = n_J = n$, the result is one reliability coefficient for the item; however, if the n_j vary from class to class, then we can derive a separate coefficient for each class size. This may be desir-

able if evaluations can be tailored to fit different classes. In general, however, we will be interested in an item's reliability across a range of class situations. Since true score variance is constant, we are most likely interested in the proportion of the average observed rating variance that is linearly predictable from the true scores. This proportion is given by

$$\rho^2(\bar{X},\zeta) = \sigma^2(a)/\overline{\sigma^2(\bar{X})} \tag{9}$$

$$= \sigma^2(a)/\{\overline{\sigma^2(a) + \sigma^2(e)/n_j}\} \tag{10}$$

$$= \sigma^2(a)/\{\sigma^2(a) + \overline{\sigma^2(e)/n_j}\}. \tag{11}$$

Now

$$\overline{\sigma^2(e)/n_j} = (1/J)\sum_j\{\sigma^2(e)/n_j\} \tag{12}$$

$$=\{\sigma^2(e)/J\}\ \sum_j(1/n_j) \tag{13}$$

$$= \sigma^2(e)/\{J/\sum_j(1/n_j)\} \tag{14}$$

$$= \sigma^2(e)/\tilde{n}, \tag{15}$$

where $\tilde{n}$ is the harmonic mean of the n_j. Thus,

$$\rho^2(\bar{X},\zeta) = \sigma^2(a)/\{\sigma^2(a) + \sigma^2(e)/\tilde{n}\}\ . \tag{16}$$

The coefficient given in (16) is the second-order partial replication coefficient of reliability for the situation of unequal class sizes.

The usual methods of estimating intraclass correlations from sample data involve combining the mean square between instructors and the mean square within instructors in accordance with their expected values to obtain unbiased estimates of $\sigma^2(a)$ and $\sigma^2(e)$. For the unbalanced one-way ANOVA being considered, Graybill (1961) gives the

expected values for the mean square between instructors (MS_B) and the mean square within instructors (MS_W) as

$$E(MS_B) = \sigma^2(e) + K\sigma^2(a) \tag{17}$$

and

$$E(MS_W) = \sigma^2(e) \tag{18}$$

where $K = (N^2 - \sum_j n_j^2)/N(J - 1)$ and $N = \sum_j n_j$. Unbiased estimates of $\sigma^2(e)$ and $\sigma^2(a)$ are given by

$$\hat{\sigma}^2(e) = MS_W \tag{19}$$

and

$$\hat{\sigma}^2(a) = (MS_B - MS_W)/K \ . \tag{20}$$

An estimate $\hat{\rho}^2(X,\zeta)$ of $\rho^2(X,\zeta)$ is given by

$$\hat{\rho}^2(X,\zeta) = \{(MS_B - MS_W)/K\}/\{(MS_B - MS_W)/K + MS_W\} \tag{21}$$

$$= (MS_B - MS_W)/\{MS_B + (K - 1)MS_W\} \ . \tag{22}$$

An estimate, $\hat{\rho}^2(\overline{X},\zeta)$, of the second-order partial replication reliability of the class mean $\rho^2(\overline{X},\zeta)$ is given by

$$\hat{\rho}^2(\overline{X},\zeta) = \{(MS_B - MS_W)/K\}/\{(MS_B - MS_W)/K + MS_W/\tilde{n}\} \tag{23}$$

$$= \tilde{n}(MS_B - MS_W)/\{\tilde{n}(MS_B + (K - \tilde{n})MS_W\}, \tag{24}$$

the same result given by the Spearman-Brown formula for the reliability of a test lengthened $\tilde{n}$ times.

The foregoing results can readily be generalized to the case of unequal variances for the errors by substitution of the average error variance for the common error variance in (6) through (18);

however, the form of the estimates is the same.

The linear model for the first-order partial replication method can be written, in true score terms, as

$$Y_{ij} = \zeta_j + \eta_{ij} \tag{25}$$

where Y_{ij} is the mean rating given instructor j by the i-th partial replicate of the class (i = 1, 2), ζ_j is the true score previously defined, and η_{ij} is the mean error for partial replicate i. The ANOVA model corresponding to (25) is

$$Y_{ij} = \mu + a_j + \xi_{ij} \tag{26}$$

where Y_{ij}, μ, and a_j are as previously defined, and ξ_{ij} is the mean ANOVA error for partial replicate i within instructor j's class. Again $\mu_j = \zeta_j$ and $\xi_{ij} = \eta_{ij}$, and the tautology is complete.

Assuming that the errors of the individual raters have equal variances, and assuming that the two partial replications in each class are of equal size, the variances for the mean errors (and the mean ratings) will be equal within each class, but unequal between classes of different sizes. Using the same approach as before, we can find the reliability for the item as the proportion of the average rating variability that is linearly predictable from the true scores, i.e.,

$$\rho^2(Y,\zeta) = \sigma^2(\zeta)/\overline{\sigma^2(Y)} = \sigma^2(a)/\{\sigma^2(a) + \overline{\sigma^2(\xi)}\} . \tag{27}$$

The mean error variance $\overline{\sigma^2(\xi)}$ is the average of the $\sigma^2(\xi_j) = \sigma^2(e)/(n_j/2) = 2\sigma^2(e)/n_j$. Thus

$$\rho^2(Y,\zeta) = \sigma^2(a)/\{\sigma^2(a) + 2\sigma^2(e)/\tilde{n}\} . \tag{28}$$

The first-order partial replication reliability of the mean rating for the entire class is then given by

$$\rho^2(\bar{Y},\zeta) = \sigma^2(a)/\{\sigma^2(a) + \overline{\sigma^2(\xi)}/2\} \tag{29}$$

$$= \sigma^2(a)/\{\sigma^2(a) + \sigma^2(e)/\tilde{n}\}, \tag{30}$$

the same parameter as in (16). The parameters $\rho^2(Y,\zeta)$ and $\rho^2(\bar{Y},\zeta)$ can be estimated in the same manner as before, using the linear model in (26). The relationships between the mean square between ($MS_{B'}$) and the mean square within ($MS_{W'}$) and the variances $\sigma^2(a)$ and $\overline{\sigma^2(\xi)}$ are given by the formulas

$$E(MS_{B'}) = \overline{\sigma^2(\xi)} + 2\sigma^2(a) \tag{31}$$

and

$$E(MS_{W'}) = \overline{\sigma^2(\xi)}. \tag{32}$$

Following the estimation techniques outlined earlier, we form estimates of $\rho^2(Y,\zeta)$ and $\rho^2(\bar{Y},\zeta)$, respectively, as

$$\hat{\rho}^2(Y,\zeta) = (MS_{B'} - MS_{W'})/(MS_{B'} + MS_{W'}) \tag{33}$$

and

$$\hat{\rho}^2(\bar{Y},\zeta) = (MS_{B'} - MS_{W'})/MS_{B'}\ . \tag{34}$$

Again, this procedure may be adapted to the case of unequal variances of the individual rater error scores by the substitution suggested previously for the second-order method.

Let us summarize the differences in the estimates $\hat{\rho}^2(\bar{X},\zeta)$ and $\hat{\rho}^2(\bar{Y},\zeta)$. In using $\hat{\rho}^2(\bar{X},\zeta)$ as an estimate of the reliability of the mean rating for the entire class, we are essentially finding the reliability for a single rater in the class and boosting that reliability

by the Spearman-Brown formula to represent the reliability of the mean rating for the entire class. If we assume equal error variances for all raters (parallel raters), then we still have to contend with the problem of unequal class frequencies through use of an unbalanced ANOVA design. If we use $\hat{\rho}^2(\overline{Y},\zeta)$ as an estimate of class mean reliability, we are using the mean rating for half the class as an approximation to the mean of the entire class. The Spearman-Brown formula is again used to reflect the reliability of the mean rating based on all ratings in each class. However, regardless of the equality or inequality of the error variances for the individual raters, we must use estimation procedures that take into account the unequal variances of the errors of the means based on half of each class.

Since both of the estimators $\hat{\rho}^2(\overline{X},\zeta)$ and $\hat{\rho}^2(\overline{Y},\zeta)$ are formed as ratios of unbiased estimators, neither method should produce an unbiased estimator of $\hat{\rho}^2(\overline{X},\zeta)$. Rather, both estimators should have means that are lower than the population parameter. Furthermore, it is not readily apparent which of the two methods should produce a better estimate of reliability in the sample. The following simulation was carried out to provide further insight into the problem.

In order to evaluate the adequacy of the two partial replication methods, two additional measures of reliability were examined which require two administrations of the evaluation item. The first of these is the intraclass correlation coefficient applied to truly replicated data. That is, the observations in the ANOVA model are the mean ratings for the entire class for each of two administrations of the item (see Lord and Novick, 1968, Ch. 7).

The second method involving truly replicated data is the sample product-moment correlation coefficient, computed across instructors between the class mean ratings on the two administrations of the item. Both of these methods yield estimates of the parameter given in (16), under the same set of assumptions, but neither estimate is guaranteed to be unbiased.

Method

Monte Carlo methods were used to simulate the sampling properties of the four estimates of reliability. A computer program was written to generate data representing ratings given twelve randomly selected instructors evaluated on an item with a given reliability. Reliabilities examined ranged from .90 to .30 in steps of .10. For each reliability, instructor true score variance was set at 1.0, and rater error variance was chosen in accordance with (16) for $\tilde{n} = 20$. Four sets of class frequencies were used to examine the differences in the various methods as a function of class size configuration. The class frequencies used in these simulations are given in Table 1.

Insert Table 1 about here

A technique developed by Box and Muller (1958) and modified by Chen (1971) was used to select instructor true scores from the unit normal distribution. For the j-th instructor, n_j rater errors were selected from a normal distribution with mean zero and variance chosen as previously described. The error scores were added to the instructor's

true score to form simulated ratings. For the two methods based upon true replication of the item, a second set of rater errors was generated to provide the ratings for the second "administration" of the item. For the first-order partial replication method, a computer subroutine was written to randomize the observations in each class into two halves. Each of the four estimates of reliability was computed using the methods described in the previous section.

The procedure of generating instructor effects and rater errors, and computing the four estimates of reliability was repeated 1000 times for each combination of class frequencies and population reliability. A count was maintained of the number of estimates that exceeded the parameter value, the number that fell short of the parameter value, and the number of negative estimates. All negative estimates were set equal to zero after counting, and means and standard errors were computed for each method.

Results

Summary statistics for the partial replications methods are given in Table 2. The corresponding statistics for the two methods based on true replications are given in Table 3. Because of similarity in results and for ease of presentation, the results from the three conditions of unequal class frequencies have been averaged across conditions.

Insert Tables 2 and 3 about here

For all population reliabilities examined other than .3, the two methods based on true replications produced estimates with less bias on the average than did the two partial replications methods. Within the true replications methods, the intraclass correlation method tended to produce more bias on the average than did the sample correlation coefficient, although the difference between these two methods was never greater than .02. Within the two partial replications methods, there was no appreciable difference in average ratings for population reliabilities of .90 to .70. For reliabilities from .60 to .30, the first-order method produced a mean estimate that was closer to the parameter value than did the second-order method, with the two methods differing by as much as .05 in the most extreme cases.

For the population reliability equal to .30, the intraclass correlation computed on true replications produced the closest average estimate, the first-order partial replication method the next closest, the sample correlation tended to overestimate on the average, and the second-order partial replication method produced the most deviant average estimate, with a negative bias of approximately .04. Across the range of reliabilities and class frequencies, none of the four methods yielded an average estimate more than .1 deviant from the parameter value.

In terms of variability, the first-order partial replication method had the largest standard error, the second-order method had the smallest, and the two methods based on true replications produced intermediate values. Overall, standard errors for the first-order method were 30% to 40% larger than those of the second-order method, and the two true-replications methods tended to be 15% to 20% more variable than

the second-order partial replication method.

Both the first-order method and the intraclass correlation method on true replications appear to be unbiased in the median, since the parameter value is close to the medians of the sampling distributions of the 1000 obtained estimates. The second-order method tended to produce more underestimates than overestimates of the parameter values, with an average of 56% of the estimates being less than the population reliability. The sample correlation coefficient, on the other hand, exceeded the parameter an average of 55% of the time, across the range of reliabilities examined.

Finally the first-order partial replication method produced a greater proportion of negative estimates than did the other methods. For a population reliability of .30, for instance, nearly 30% of the first-order estimates were less than zero, almost twice as many as produced by the methods based on true replications, and 10% to 40% more than were produced by the second-order method.

Differences between equal and unequal class frequencies seemed to be negligible in their effect on the mean estimate or on the standard error, and produced no regular pattern of effects on the numbers of overestimates, underestimates, and negative estimates.

Discussion

In terms of bias, variability, and proportion of negative estimates, either of the methods based on true replication of the evaluation should be preferred over the first-order partial replication

method. This is not unreasonable, since both partial replication methods serve as approximations to true replication of the evaluation. To their credit, both partial replication methods do an acceptable job of estimation, considering that they deal with half as much information as do the true replications methods.

However, a decision between the two partial replications methods is not easy on the basis of the summary data. The first-order method leads to less bias but more variability; the second-order method produces fewer negative estimates, but more underestimates of the parameter value.

Perhaps the deciding point is the manner in which negative estimates are handled by the two partial replication methods. Each of the four methods examined will produce a negative estimate of reliability a certain proportion of the time. For any of the methods other than the first-order method, such an occurrence forces the researcher to decide between interpreting the findings as zero reliability, on the one hand, and running a new reliability study with a new sample of instructors, on the other.

The first-order partial replication method, however, allows the researcher to reanalyze the data at hand, perhaps several times, by re-randomizing each class's data. A mean of several such reanalyses might be a reasonable non-zero estimate of the population reliability. To examine this possibility, data from the simulation using reliability equal to .30 was subjected to such reanalysis. This reliability was previously shown to lead to more negative estimates in each method than any other value examined. For each of ten simulated reliability studies, ten

estimates of the first-order partial replication reliability coefficient were computed. Each estimate was computed on a different randomization of the data, and any negative estimates were set equal to zero. For each simulated study, the ten estimates of reliability were averaged to yield a mean reliability estimate. The resulting ten mean estimates ranged from .036 to .638, with an overall mean of .323 for the ten studies.

This method of reanalysis bears further investigation, as do situations involving other class frequency configurations, different numbers of classes, and non-normal distributions of true and error scores. There is a definite need for simulation studies employing discrete ratings instead of the continuous ratings used in the present study, since most ratings are based on Likert-like scales with only a finite number of possible responses.

The present study dealt only with parallel raters, i.e., with equal variances for the error scores. In reality the variance of the errors may be related to the size of the class or type of class, in which case an assumption of tau equivalence (unequal variances) would be more easily defended.

The approach used in the present paper should be extended to higher-order designs. One of the advantages of the generic true score method is that it enables the user to look at multiple sources of variability, including multiple items, differing conditions of administration, occasions of evaluation, etc. Extension of the partial replication methods to higher designs would bring the theory one step closer to the reality of application.

In summary, for the fairly restrictive set of conditions examined in the present study, the first-order and second-order partial replication methods, both modifications of the generic true score approach, perform adequately when compared to methods requiring true replication of teacher evaluations. Of the two partial replication methods, the first-order offers less bias on the average, apparent unbiasedness in the median, and, most importantly, the opportunity to reanalyze data instead of re-running the study when a negative estimate is obtained. It stands as an acceptable alternative to the time-consuming and expensive methods based on test-retest estimation and enables the devotee of generalizability theory to utilize all the information collected in estimating reliability.

References

Box, G. E. P. & Muller, M. E. A note on the generation of random normal deviates. Annals of Mathematical Statistics, 1958, 29, 610-611.

Chen, E. H. A random normal number generator for 32-bit-word computers. Journal of the American Statistical Association, 1971, 66, 400-403.

Costin, F. A graduate course in the teaching of psychology: Description and evaluation. Journal of Teacher Education, 1968, 19, 425-432.

Costin, F. An empirical test of the "teacher-centered" versus "student-centered" dichotomy. Journal of Educational Psychology, 1971, 62, 410-412.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Doyle, K. O., jr. & Whitely, S. E. Student ratings as criteria for effective teaching. American Educational Research Journal, 1974, 11, 259-274.

Graybill, F. A. An Introduction to Linear Statistical Models, Vol. 1. New York: McGraw-Hill, 1961.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-183.

Lord, F. M. & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Scheffé, H. The Analysis of Variance. New York: Wiley, 1959.

Somers, L. G. & Southern, M. L. A rating scale for evaluation of teaching effectiveness for use with junior high school students. California Journal of Educational Research, 1974, 25, 128-133.

Winer, B. J. Statistical Principles in Experimental Design, 2d ed. New York: McGraw-Hill, 1971.

Table 1

Class Frequency Patterns and Graybill's K Factors Used in the Simulation

Pattern Number	Frequencies	K
1	20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20	20.000
2	10, 10, 10, 30, 30, 30, 30, 30, 30, 30, 30, 30	24.727
3	12, 12, 12, 12, 12, 12, 60, 60, 60, 60, 60, 60	34.545
4	10, 10, 10, 24, 24, 24, 30, 30, 30, 40, 40, 40	25.587

Table 2
Summary Statistics for the First-Order Method, $\rho^2(\bar{Y},\zeta)$, and the Second-Order Method, $\rho^2(\bar{X},\zeta)$, for 1000 Replications of 12 Classes

Reliability	Method *	Mean†	Std. Error†	% High	% Low	% Neg
.9	$\rho^2(\bar{Y},\zeta)$ (=)	.874	.0922	49.2	50.8	.1
	$\rho^2(\bar{Y},\zeta)$ (≠)	.877	.0905	50.1	49.9	.03
	$\rho^2(\bar{X},\zeta)$ (=)	.877	.0678	46.5	53.5	0.0
	$\rho^2(\bar{X},\zeta)$ (≠)	.874	.0048	42.2	57.8	0.0
.8	$\rho^2(\bar{Y},\zeta)$ (=)	.752	.1694	50.0	50.0	.4
	$\rho^2(\bar{Y},\zeta)$ (≠)	.757	.1672	50.8	49.2	.8
	$\rho^2(\bar{X},\zeta)$ (=)	.757	.1308	46.2	53.8	.4
	$\rho^2(\bar{X},\zeta)$ (≠)	.755	.1233	42.2	57.8	.1
.7	$\rho^2(\bar{Y},\zeta)$ (=)	.636	.2241	49.4	50.6	3.6
	$\rho^2(\bar{Y},\zeta)$ (≠)	.643	.2255	50.5	49.5	3.1
	$\rho^2(\bar{X},\zeta)$ (=)	.641	.1764	45.2	54.8	1.5
	$\rho^2(\bar{X},\zeta)$ (≠)	.641	.1652	42.2	57.8	.7
.6	$\rho^2(\bar{Y},\zeta)$ (=)	.532	.2561	49.1	50.9	7.8
	$\rho^2(\bar{Y},\zeta)$ (≠)	.540	.2671	50.7	49.3	8.1
	$\rho^2(\bar{X},\zeta)$ (=)	.529	.2076	45.0	55.0	4.8
	$\rho^2(\bar{X},\zeta)$ (≠)	.533	.1933	43.3	56.7	2.7
.5	$\rho^2(\bar{Y},\zeta)$ (=)	.441	.2709	48.4	51.6	15.1
	$\rho^2(\bar{Y},\zeta)$ (≠)	.451	.2776	50.9	49.1	14.2
	$\rho^2(\bar{X},\zeta)$ (=)	.427	.2229	44.6	55.4	9.6
	$\rho^2(\bar{X},\zeta)$ (≠)	.433	.2070	44.6	55.4	6.4
.4	$\rho^2(\bar{Y},\zeta)$ (=)	.364	.2722	48.8	51.2	22.1
	$\rho^2(\bar{Y},\zeta)$ (≠)	.376	.2809	50.9	49.1	22.2
	$\rho^2(\bar{X},\zeta)$ (=)	.336	.2263	44.7	55.3	16.3
	$\rho^2(\bar{X},\zeta)$ (≠)	.342	.2089	44.5	55.5	11.9
.3	$\rho^2(\bar{Y},\zeta)$ (=)	.298	.2668	48.4	51.6	29.4
	$\rho^2(\bar{Y},\zeta)$ (≠)	.314	.2760	50.1	49.9	29.4
	$\rho^2(\bar{X},\zeta)$ (=)	.260	.2182	44.1	55.9	26.5
	$\rho^2(\bar{X},\zeta)$ (≠)	.261	.1995	44.5	55.5	20.1

* "=" refers to the results for the equal-frequencies conditions; "≠" refers to the average of the results for the unequal-frequencies conditions

† Negative reliability estimates were set equal to zero in computing means and standard errors.

Table 3
Summary Statistics for the Intraclass Correlation, ρI, and the Sample Correlation, r, for 1000 Replications of 12 Classes

Reliability	Method*	Mean†	Std. Error†	% High	% Low	% Neg
.9	ρI (=)	.874	.0728	48.0	52.0	0.0
	ρI (≠)	.886	.0775	51.1	48.9	0.0
	r (=)	.886	.0714	54.0	46.0	0.0
	r (≠)	.890	.0759	57.1	42.9	0.0
.8	ρI (=)	.780	.1158	50.8	49.2	0.0
	ρI (≠)	.772	.1340	51.2	48.8	.03
	r (=)	.793	.1166	57.2	42.8	0.0
	r (≠)	.787	.1341	56.4	43.6	.03
.7	ρI (=)	.666	.1738	50.1	49.9	.5
	ρI (≠)	.669	.1785	51.3	48.7	.4
	r (=)	.682	.1778	55.1	44.9	.7
	r (≠)	.685	.1823	55.8	44.2	.5
.6	ρI (=)	.569	.1957	50.9	49.1	1.0
	ρI (≠)	.569	.2054	51.2	48.8	1.8
	r (=)	.587	.2015	54.0	46.0	1.3
	r (≠)	.587	.2111	54.5	45.5	2.0
.5	ρI (=)	.459	.2145	48.9	51.1	3.5
	ρI (≠)	.475	.2282	51.5	48.5	4.5
	r (=)	.477	.2234	52.1	47.9	4.2
	r (≠)	.493	.2361	54.9	45.1	5.0
.4	ρI (=)	.390	.2220	50.8	49.2	7.6
	ρI (≠)	.386	.2387	50.7	49.3	10.4
	r (=)	.407	.2315	53.6	46.4	7.7
	r (≠)	.405	.2488	53.2	46.8	10.5
.3	ρI (=)	.299	.2191	49.4	50.6	15.4
	ρI (≠)	.309	.2344	50.7	49.3	17.8
	r (=)	.313	.2285	51.6	48.4	15.2
	r (≠)	.325	.2466	52.6	47.4	18.7

* "=" refers to the results for the equal-frequencies condition; "≠" refers to the average of the results for the unequal-frequencies conditions

† Negative reliability estimates were set equal to zero in computing means and standard errors.

APPENDIX

COMPLETE SIMULATION RESULTS

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH

RELIABILITY EQUAL TO 0.90

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 2.25

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.874	0.0922	49.2	50.8	0.1
1	SECOND ORDER	0.877	0.0678	46.5	53.5	0.0
1	INTRACLASS	0.877	0.0728	48.0	52.0	0.0
1	CORRELATION	0.896	0.0714	54.0	46.0	0.0
2	FIRST ORDER	0.876	0.0949	49.5	50.5	0.1
2	SECOND ORDER	0.874	0.0671	42.9	57.1	0.0
2	INTRACLASS	0.879	0.0775	49.1	50.9	0.0
2	CORRELATION	0.887	0.0762	54.6	45.4	0.0
3	FIRST ORDER	0.880	0.0843	51.6	48.4	0.0
3	SECOND ORDER	0.873	0.0707	41.5	58.5	0.0
3	INTRACLASS	0.880	0.0762	50.0	50.0	0.0
3	CORRELATION	0.890	0.0735	56.3	43.7	0.0
4	FIRST ORDER	0.874	0.0922	49.3	50.7	0.0
4	SECOND ORDER	0.874	0.0686	42.1	57.9	0.0
4	INTRACLASS	0.884	0.0767	54.2	45.8	0.0
4	CORRELATION	0.892	0.0781	60.3	39.7	0.0

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH

RELIABILITY EQUAL TO 0.60

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 5.00

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.752	0.1694	50.0	50.0	0.4
1	SECOND ORDER	0.758	0.1308	46.2	53.8	0.4
1	INTRACLASS	0.780	0.1158	50.8	49.2	0.0
1	CORRELATION	0.793	0.1166	57.2	42.8	0.0
2	FIRST ORDER	0.757	0.1700	50.4	49.6	0.8
2	SECOND ORDER	0.755	0.1229	43.1	56.9	0.1
2	INTRACLASS	0.773	0.1156	51.5	48.5	0.1
2	CORRELATION	0.789	0.1334	58.0	42.0	0.1
3	FIRST ORDER	0.762	0.1622	51.9	48.1	0.7
3	SECOND ORDER	0.756	0.1212	42.0	58.0	0.0
3	INTRACLASS	0.767	0.1382	49.6	50.4	0.0
3	CORRELATION	0.781	0.1411	54.3	45.7	0.0
4	FIRST ORDER	0.753	0.1694	50.1	49.9	0.8
4	SECOND ORDER	0.754	0.1257	42.1	57.9	0.2
4	INTRACLASS	0.776	0.1281	52.4	47.6	0.0
4	CORRELATION	0.790	0.1277	56.9	43.1	0.0

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH RELIABILITY EQUAL TO 0.70

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 0.57

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.636	0.2241	49.4	50.6	3.6
1	SECOND ORDER	0.641	0.1764	45.2	54.8	1.5
1	INTRACLASS	0.666	0.1738	50.1	49.9	0.5
1	CORRELATION	0.682	0.1778	55.1	44.9	0.7
2	FIRST ORDER	0.642	0.2291	50.5	49.5	3.4
2	SECOND ORDER	0.638	0.1694	43.0	57.0	1.0
2	INTRACLASS	0.666	0.1766	50.5	49.5	0.5
2	CORRELATION	0.680	0.1800	54.7	45.3	0.6
3	FIRST ORDER	0.646	0.2245	50.5	49.5	2.6
3	SECOND ORDER	0.645	0.1581	42.1	57.9	0.2
3	INTRACLASS	0.672	0.1789	51.2	48.8	0.3
3	CORRELATION	0.698	0.1819	56.3	43.7	0.3
4	FIRST ORDER	0.639	0.2229	50.5	49.5	3.4
4	SECOND ORDER	0.639	0.1682	41.5	58.5	0.9
4	INTRACLASS	0.668	0.1800	52.1	47.9	0.4
4	CORRELATION	0.684	0.1849	56.4	43.6	0.5

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES. USING 12 CLASSES AND AN ITEM WITH RELIABILITY EQUAL TO 0.60

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 13.33

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.532	0.2561	49.1	50.9	7.8
1	SECOND ORDER	0.529	0.2076	45.0	55.0	4.8
1	INTRACLASS	0.569	0.1957	50.9	49.1	1.0
1	CORRELATION	0.587	0.2015	54.0	46.0	1.3
2	FIRST ORDER	0.541	0.2651	51.2	48.8	8.5
2	SECOND ORDER	0.529	0.1997	43.7	56.3	4.9
2	INTRACLASS	0.573	0.2020	54.6	45.4	1.9
2	CORRELATION	0.594	0.2069	57.1	42.9	1.9
3	FIRST ORDER	0.543	0.2627	51.0	49.0	8.5
3	SECOND ORDER	0.540	0.1836	44.1	55.9	1.4
3	INTRACLASS	0.569	0.2083	50.5	49.5	1.8
3	CORRELATION	0.586	0.2149	54.5	45.5	2.3
4	FIRST ORDER	0.535	0.2555	49.8	50.2	7.4
4	SECOND ORDER	0.531	0.1965	42.2	57.8	2.8
4	INTRACLASS	0.565	0.2059	48.4	51.6	1.6
4	CORRELATION	0.582	0.2114	51.9	48.1	1.9

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH RELIABILITY EQUAL TO 0.50

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 20.00

CLASS SIZE	PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
	1	FIRST ORDER	0.441	0.2709	48.4	51.6	15.1
	1	SECOND ORDER	0.427	0.2229	44.6	55.4	9.6
	1	INTRACLASS	0.459	0.2145	48.9	51.1	3.5
	1	CORRELATION	0.476	0.2234	52.1	47.9	4.2
	2	FIRST ORDER	0.454	0.2814	52.2	47.8	14.0
	2	SECOND ORDER	0.429	0.2131	44.6	55.4	7.6
	2	INTRACLASS	0.468	0.2267	49.2	50.8	4.4
	2	CORRELATION	0.485	0.2364	52.8	47.2	4.5
	3	FIRST ORDER	0.455	0.2780	51.3	48.7	14.5
	3	SECOND ORDER	0.440	0.1972	45.8	54.2	4.1
	3	INTRACLASS	0.472	0.2249	50.5	49.5	4.2
	3	CORRELATION	0.493	0.2324	54.6	45.4	4.9
	4	FIRST ORDER	0.443	0.2733	49.3	50.7	14.1
	4	SECOND ORDER	0.430	0.2107	43.5	56.5	7.4
	4	INTRACLASS	0.463	0.2309	54.8	45.2	5.0
	4	CORRELATION	0.503	0.2396	57.4	42.6	5.5

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH

RELIABILITY EQUAL TO 0.40

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 30.00

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.364	0.2722	48.8	51.2	22.1
1	SECOND ORDER	0.336	0.2263	44.7	55.3	16.3
1	INTRACLASS	0.390	0.2220	50.8	49.2	7.6
1	CORRELATION	0.407	0.2315	53.6	46.4	7.7
2	FIRST ORDER	0.382	0.2852	52.2	47.8	22.1
2	SECOND ORDER	0.337	0.2159	44.6	55.4	14.0
2	INTRACLASS	0.390	0.2392	52.3	47.7	10.5
2	CORRELATION	0.410	0.2468	54.9	45.1	9.8
3	FIRST ORDER	0.381	0.2821	50.5	49.5	22.0
3	SECOND ORDER	0.348	0.1982	45.4	54.6	9.4
3	INTRACLASS	0.376	0.2474	47.5	52.5	11.7
3	CORRELATION	0.394	0.2565	50.7	49.3	12.1
4	FIRST ORDER	0.366	0.2744	50.0	50.0	22.4
4	SECOND ORDER	0.339	0.2126	43.6	56.4	12.4
4	INTRACLASS	0.393	0.2296	52.3	47.7	8.9
4	CORRELATION	0.411	0.2410	54.0	46.0	9.5

SUMMARY STATISTICS FOR THE FOUR RELIABILITY ESTIMATES, USING 12 CLASSES AND AN ITEM WITH RELIABILITY EQUAL TO 0.30

INSTRUCTOR VARIANCE 1.00

RATER ERROR VARIANCE 46.67

CLASS SIZE PATTERN	METHOD	MEAN	STD. ERROR	% HIGH	% LOW	% NEGATIVE
1	FIRST ORDER	0.298	0.2668	48.4	51.6	29.4
1	SECOND ORDER	0.260	0.2182	44.1	55.9	26.5
1	INTRACLASS	0.299	0.2191	49.4	50.6	15.4
1	CORRELATION	0.313	0.2285	51.6	48.4	15.2
2	FIRST ORDER	0.323	0.2805	51.9	48.1	29.4
2	SECOND ORDER	0.259	0.2062	44.9	55.1	23.2
2	INTRACLASS	0.307	0.2309	50.7	49.3	17.0
2	CORRELATION	0.322	0.2433	52.2	47.8	17.8
3	FIRST ORDER	0.317	0.2796	49.7	50.3	28.8
3	SECOND ORDER	0.265	0.1876	45.0	55.0	15.6
3	INTRACLASS	0.312	0.2390	50.3	49.7	17.7
3	CORRELATION	0.329	0.2502	53.0	47.0	18.4
4	FIRST ORDER	0.301	0.2678	48.8	51.2	30.1
4	SECOND ORDER	0.258	0.2047	43.7	56.3	21.4
4	INTRACLASS	0.309	0.2332	51.1	48.9	18.7
4	CORRELATION	0.325	0.2464	52.7	47.3	19.8