DOCUMENT RESUME

ED 068 534 TM 001 900

| | |
|---|---|
| AUTHOR | Baker, J. Philip |
| TITLE | Using Generalizability Theory and Multifacet Designs in the Validation of a Classroom Observation Instrument. |
| INSTITUTION | Stanford Univ., Calif. Stanford Center for Research and Development in Teaching. |
| SPONS AGENCY | Office of Education (DHEW), Washington, D.C. |
| REPORT NO | SCRDT-RDM-79 |
| BUREAU NO | BR-5-0252 |
| PUB DATE | Sep 71 |
| CONTRACT | OEC-6-10-078 |
| NOTE | 17p. |
| EDRS PRICE | MF-$0.65 HC-$3.29 |
| DESCRIPTORS | Analysis of Variance; Bias; *Classroom Observation Techniques; Data Analysis; *Generalization; Measurement Instruments; *Measurement Techniques; Reliability; *Test Reliability; Test Results; Theories |

ABSTRACT

The usefulness of generalizability theory in assessing the reliability of classroom observation instruments is illustrated, with a new index of reliability, called the coefficient of generalizability, given as an index of how well one can generalize from the instrument to the universe score according to the conditions of observation. Data from an instrument assessing verbal behavior are reanalyzed using generalizability theory. Three raters observed two classes of five student teachers, and three facets (class, occasion, and rater) were specified for study. Analysis of variance yielded the components of variance due to these facets, and ratio of the estimated universe. Score variance to the observed-score variance was computed as the coefficient of generalizability. The possible effect of various hypothetical conditions of observation on the coefficient of generalizability was considered. The results showed the coefficient of generalizability for one rater observing one class on one day was near zero for most of the summary variables for the instrument, but it increased when the number of observers and days of observation were increased. This was directly attributable to rater bias. Different observer training procedures might improve the usefulness of the instrument. (Author/LH)

Stanford Center for Research and Development in Teaching

SCHOOL OF EDUCATION   STANFORD UNIVERSITY

## STANFORD CENTER FOR RESEARCH AND DEVELOPMENT IN TEACHING

### Publication Resume

J. Philip Baker. Using Generalizability Theory and Multifacet Designs in the Validation of a Classroom Observation Instrument. Research and Development Memorandum No. 79. September 1971. 15 pp.

Purpose: To illustrate how generalizability theory is useful in assessing the reliability of classroom observation instruments. The conditions under which a given instrument is used vary from situation to situation, which changes the reliability of the instrument. A new index of reliability, called the coefficient of generalizability, is given as an index of how well one can generalize from the instrument to the universe score according to the conditions of observation.

Method: Data from an instrument assessing verbal behavior are reanalyzed using generalizability theory. Three raters observed two classes of five student teachers at the Stanford University School of Education on two occasions. Three facets (Class, Occasion, and Rater) were specified for study. Analysis of variance yielded the components of variance due to these facets, and ratio of the estimated universe. Score variance to the observed-score variance was computed as the coefficient of generalizability. The possible effect of various hypothetical conditions of observation on the coefficient of generalizability was considered.

Results: The coefficient of generalizability for one rater observing one class on one day was near zero for most of the summary variables for the instrument. The coefficient was increased when the number of observers and the number of days of observation were increased. This result was directly attributable to rater bias. Perhaps different observer training procedures could improve the usefulness of the instrument.

Target groups: Students of tests and measurements; educational researchers.

STANFORD CENTER
FOR RESEARCH AND DEVELOPMENT
IN TEACHING

Research and Development Memorandum No. 79

# USING GENERALIZABILITY THEORY AND MULTIFACET DESIGNS IN THE VALIDATION OF A CLASSROOM OBSERVATION INSTRUMENT

J. Philip Baker

School of Education
Stanford University
Stanford, California

September 1971

Published by the Stanford Center for Research and Developmen: in Teaching, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred. (Contract No. OEC-6-10-078, Project No. 5-0252-0501.)

## Introductory Statement

The Center is concerned with the shortcomings of teaching in American schools: the ineffectiveness of many American teachers in promoting achievement of higher cognitive objectives, in engaging their students in the tasks of school learning, and, especially, in serving the needs of students from low-income areas. Of equal concern is the inadequacy of American schools as environments fostering the teachers' own motivations, skills, and professionalism.

The Center employs the resources of the behavioral sciences--theoretical and methodological--in seeking and applying knowledge basic to achievement of its objectives. Analysis of the Center's problem area has resulted in three programs: Heuristic Teaching, Teaching Students from Low-Income Areas, and the Environment for Teaching. Drawing primarily upon psychology and sociology, and also upon economics, political science, and anthropology, the Center has formulated integrated programs of research, development, demonstration, and dissemination in these three areas. In the Heuristic Teaching area, the strategy is to develop a model teacher training system integrating components that dependably enhance teaching skill. In the program on Teaching Students from Low-Income Areas, the strategy is to develop materials and procedures for engaging and motivating such students and their teachers. In the program on Environment for Teaching, the strategy is to develop patterns of school organization and teacher evaluation that will help teachers function more professionally, at higher levels of morale and commitment.

The difficulty of constructing a reliable instrument for observing classroom behavior often arises from failure to understand that the instrument will be used under different conditions from those in which it was devised. The author suggests that Cronbach's coefficient of generalizability offers a solution to the problem. This paper was prepared as a part of the Training Studies project of the Heuristic Teaching program.

## Abstract

This paper illustrates how generalizability theory can be used in assessing the reliability of classroom observation instruments. A new index of reliability, called the coefficient of generalizability, is used to measure how well one can generalize from the observation instrument to the universe score in different conditions of observation. Data from an instrument assessing verbal behavior are reanalyzed. Three raters observed two classes of students on two days. The coefficient of generalizability for one rater observing one class on one day was near zero for most of the summary variables for the instrument; the coefficient was increased when the number of observers and the number of observations (days) were increased. This result could be directly attributed to rater bias. The instrument may be improved by changing rater training procedures.

# USING GENERALIZABILITY THEORY AND MULTIFACET DESIGNS IN THE VALIDATION OF A CLASSROOM OBSERVATION INSTRUMENT

J. Philip Baker

Reliability is a principal concern in the construction of instruments for observing classroom behavior. In the statistical literature various formulae exist for computing reliability coefficients. Unfortunately, these are sometimes used without full understanding of their meaning. One frequent cause of the misunderstanding and misinterpretation of reliability coefficients is that the conditions under which an instrument is constructed often are not the same as the conditions under which the instrument is used for evaluative or decision-making purposes. In such situations, previously published reliability information may no longer apply.

Cronbach et al. (in press) have outlined a solution to the problem: they recommend that instead of publishing only reliability data (which are valid only for the conditions defined in the validation process), the researcher should describe as many facets (general classes into which conditions of the study fall) of the observation process as might be important in the use of the instrument, and should include the variance component for each of these facets. The user is then able to construct coefficients appropriate to his own particular needs.

The Cronbach et al. report uses a new index of reliability, called the coefficient of generalizability, rather than the traditional reliability coefficient. For each facet there is a coefficient of generalizability (intraclass correlation coefficient) that serves as an index for generalizing from the instrument to the universe score defined by the facet. For example, although a student gets only a sample of mathematical questions on an ability test, a researcher would probably want to generalize the

The author, now at Stanford Research Institute, was formerly a Research Assistant at the Center.

the student's performance to all possible math questions of comparable concepts and difficulty (i.e., his math ability). The coefficient can be computed using data from the validation study in a manner prescribed by the conditions of decision study.

The following study will illustrate the use of a generalizability coefficient, using an actual observation instrument, and will demonstrate how the Cronbach procedure can shed new light on the validity of the instrument.

## Study of Verbal Behavior

Maria Flores-Hernandez (1970), in her doctoral dissertation at Stanford University, developed a classroom observation instrument to assess the verbal behavior of teacher and students during discussion periods. Each sentence of the transcripts of the discussions was rated in three general categories: Use of Previous Comments, Type of Conceptualization, and Level of Conceptualization. Each general category was broken down into smaller subcategories representing different types of verbal behavior, as follows:

I. Use of Previous Comments

A-A (Accepts-Advances): Speaker accepts previous idea or adds to idea.

R-R (Repeats-Rephrases): Speaker simply repeats or rephrases earlier comments.

EL (Elaborates): Speaker adds own ideas to earlier comments.

SU (Summarizes): Speaker summarizes what has been said earlier.

II. Type of Conceptualization

CM (Cognitive memory): Speaker is remembering earlier events.

CV (Convergent production): Speaker is synthesizing data to form a conclusion.

DV (Divergent production): Speaker is taking a new approach to the problem, is being creative, or is changing the subject.

EV (Evaluation): Speaker is judging the worth, beauty, or correctness of something.

III. Level of Conceptualization

Da (Date): Speaker is working with facts.

Co (Concepts): Speaker is working with concepts (one at a time).

Ge (Generalization): Speaker is working with a system of concepts.

A catch-all category called Administration was used by the raters for sentences that did not fit in any of the three general categories of the system. Sentences assigned to the catch-all category (9.5 percent of the sentences in all) were not used in the analysis. In addition to the ratings along the three categorical dimensions, actuarial counts of the speaker (Teacher or Pupil) and the sentence type (Statement or Question) were taken.

Observations were conducted in the classrooms of five Stanford interns in the Secondary Teacher Education Program (STEP). Each teacher had two classes. Each class was observed on two consecutive days. A stenographer was present in each classroom, and tape recordings were also made of the discussions. From these sources, typescripts were made, and these typescripts were then rated by three raters.

The Use of Previous Comments category originally had an additional subcategory--Ignores (the previous comment)--but only 14 out of some 16,000 sentences were assigned to this category. Thus it was excluded and those observations were dropped.

Although each of the three major categories could receive a numerical "score" for each sentence, it was difficult to decide which of the subcategories represented the low end of the scale, and which the high end. It was decided to assign each sentence to a category that represented the intersection of the ratings in each of the three major categories; this meant that there were now 48 variables (4 x 4 x 3). For example, a sentence was placed in the A-A,CV,Co box if it received ratings of A-A in the Use of Comments category, CV in the Type of Conceptualization category, and Co in the Level of Conceptualization category.

Initially, the investigators were interested in looking for differences in the distribution of the tallies for each of the variables between Teacher-Statements, Teacher-Questions, Pupil-Statements, and Pupil-Questions. A computer program was written which tallied the ratings (a) for each class discussion in the four above-mentioned sentence types, (b) for each class regardless of sentence type, (c) for each sentence type across classes, and (d) for the total number of sentences across classes and sentence type. Nearly 15,000 entries were made in the summary table.

For each table, both raw scores and proportions were computed. Since the proportions were small--usually less than 0.15--they were transformed using the arc sine transformation to normalize their distributions.

The initial analyses were made using three facets: Class (2 levels), Day (2 levels), and Rater (3 levels). A fourth variable, that of Intern, entered into the reanalysis of the data; it was this variable over which I wished to generalize. Accordingly, a four-way analysis of variance (ANOVA) was performed, each of the 48 variables receiving separate analysis. The mean squares obtained from each analysis were used in continuing the generalizability study.

At this point it is appropriate to explain the generalizability methodology used, before proceeding to the results of the ANOVA.

## Generalizability Techniques

The first step in this procedure is to identify components of variance. Components of variance can be extracted from mean squares by algebraic methods. For each mean square in the ANOVA table, an expression representing its composition can be written. For example, in a fully crossed design, each main effect term contains terms of all other interaction terms plus the main effect term; these terms are multiplied by the product of the sample sizes of the variables that do not appear in each term.

The notation used in this discussion is as follows: $p$ , $i$ , and $j$ represent the main effects in the ANOVA table. $p$ stands for the population that is the object of generalization (usually persons), and $i$ and $j$ are the sets of conditions (facets) that can also be objects of generalization. $pi$ , $pj$ , $ij$ , and $pij$ are the interaction terms in the ANOVA table. $\sigma^2(p)$ represents the variance due to $p$ . The variance due to the other terms in the table is similarly notated. $n_p$ , $n_i$ , and $n_j$ refer to the sample sizes of $p$ , $i$ , and $j$ respectively.

$n_i\sigma^2(p) + \sigma^2(pi,e)$ is interpreted as the sum of the product of the sample size of i times the variance due to p plus the variance due to the p x i interaction (this last term being confounded with the error, or residual, variance in this case). The following examples illustrate mean square expressions for 2- and 3-way ANOVA tables.

| | Source | Expected Mean Square<br>Component Expression |
|---|---|---|
| 2-way: | p | $n_i\sigma^2(p) + \sigma^2(pi,e)$ |
| | i | $n_p\sigma^2(i) + \sigma^2(pi,e)$ |
| | pi,e | $\sigma^2(pi,e)$ |
| 3-way: | p | $n_in_j\sigma^2(p) + n_i\sigma^2(pj) + n_j\sigma^2(pi) + \sigma^2(pij,e)$ |
| | i | $n_pn_j\sigma^2(i) + n_p\sigma^2(ij) + n_j\sigma^2(pi) + \sigma^2(pij,e)$ |
| | j | $n_pn_i\sigma^2(j) + n_i\sigma^2(pj) + n_p\sigma^2(ij) + \sigma^2(pij,e)$ |
| | pi | $n_j\sigma^2(pi) + \sigma^2(pij,e)$ |
| | pj | $n_i\sigma^2(pj) + \sigma^2(pij,e)$ |
| | ij | $n_p\sigma^2(ij) + \sigma^2(pij,e)$ |
| | pij,e | $\sigma^2(pij,e)$ |

In looking at the 2-way table, it can be seen that the value of the pi,e term (also called the residual) is simply the estimate of the pi and e (error) components provided by ANOVA computer programs: the residual is computed by subtracting the grand mean, row, and column effects from each observation in turn, squaring this result, summing over all observations, and dividing by the appropriate number of degrees of freedom. The pi,e component, then, in analyses where there is only one observation in each cell, represents whatever variance is not accounted for by the main effects (rows and columns). If there is more than one observation per cell, then the pi,e term becomes an interaction term (pi), and an additional term (e) is available as the residual, or error, term. This discussion is confined to the one observation per cell situation. In looking at the expression for the p term, it can be seen that by subtracting the pi,e term, $n_i$ times the variance of p remains $(n_i\sigma^2(p))$. This is because the variance of p has been sampled $n_i$ times, and the average has not yet been computed in the mean square term. Dividing $n_i\sigma^2(p)$ by $n_i$ leaves us with $\sigma^2(p)$, which is the variance attributable to facet p. The same reasoning applies to extracting the component for i.

The 3-way table presents a more complex situation, but the same principles apply. The value of the residual component $(\sigma^2(pij,e))$ is directly available as the mean square for the residual. To obtain the components for each source in the table, these steps may be followed.

1. Subtract the residual value from each of the interaction mean square values. Divide each of these results by the appropriate n value. (E.g., the pi component is obtained by subtracting $MS_{pij,e}$ from $MS_{pi}$ and dividing the result by $n_j$, since j is the variable not present in the pi expression.) We now have the components of variance for each interaction term.
2. The formula for computing the component for j is
$$\sigma^2(j) = MS_j - MS_{ij} - MS_{pj} + MS_{pij,e} .$$
The residual term is added, rather than subtracted, at the end of the expression because both $MS_{pj}$ and $MS_{ij}$ contain the residual (so it is subtracted twice), but the expression for $\sigma^2(j)$ contains the residual only once.
3. The p and i components may be extracted using similar equations:
$$\sigma^2(i) = MS_i - MS_{pi} - MS_{ij} + MS_{pij,e} ,$$
$$\sigma^2(p) = MS_p - MS_{pi} - MS_{pj} + MS_{pij,e} .$$

Having derived the components of variance for each source, the next step is to find the discrepancy between the observed scores for each variable and the universe mean; this is done in a manner that represents an intention to generalize over all or selected facets. To generalize over all facets, $X_{pIJ...} - M_p$ is needed. The $M_p$ indicates that generalization is over all facets, and the capital IJ denotes all conditions i and all j. The symbol for the discrepancy score is $\Delta_{pIJ}$. Then the within-class standard deviation is $\sigma(\Delta)$. $\sigma^2(\Delta)$ can be represented in terms of components of $\Delta$. Using the Cronbach et al. notation, $M_p\sim$ will stand for all terms of the component $M_p - M$. Similarly $M_{pI}\sim$ will stand for $M_{pI} - M_p - M_I + M$.

In the reanalysis of the Flores-Hernandez study, $M_p$ represents the mean of the universe of classes conducted by the interns. Since there are three obvious facets which could be examined, the notation for observed scores could be $X_{pIJK}$. Breaking down the terms of $X_{pIJK}$ yields:

$$M + M_p\sim + M_I\sim + M_J\sim + M_K\sim + M_{pI}\sim + M_{pJ}\sim + M_{pK}\sim + M_{IJ}\sim + M_{IK}\sim + M_{JK}\sim +$$

$$M_{pIJ}\sim + M_{pIK}\sim + M_{pJK}\sim + M_{IJK}\sim + M_{pIJK,e}\sim .$$

The terms of $M_p$ are $M + M_p\sim$. $\Delta_{pIJK}$ is found by subtracting the expression for $M_p$ from the expression for $X_{pIJK}$. This leaves $M_I\sim + M_J\sim + M_K\sim + M_{pI}\sim + \ldots + M_{IJK}\sim + M_{pIJK,e}\sim$ . The variance components of each of these terms is divided by their frequency of occurrence within each class (within p), and the sum of the resulting values is the value of $\sigma^2(\Delta)$. This term represents the variance of the within-class discrepancy.

Of final interest in this analysis will be the coefficient of generalizability, which has already been introduced as a reliability coefficient for a particular universe of generalization. It is estimated by the ratio of the estimated universe-score variance to the expected observed-score variance.

## Results of Generalizability Reanalysis

In this study, all 15 terms of the ANOVA table, which would normally appear in a 4-way fully crossed design, did not appear, since classes were nested within interns. Also, results of the first analysis indicated that the component for interns was relatively small. It was decided to eliminate the Intern facet from the analysis, and consider only classes. This left 10 classes, 3 raters, and 2 days in a fully crossed design.

The data using the 48-variable analysis were discouraging. It appeared that rater bias provided most of the variance in the scores, and that reliability was virtually nonexistent. It was then decided to look at marginal totals of the 11 main classifications (A-A, R-R, EL, SU, CM, CV, DV, EV, Da, Co, Ge).

Table 1 shows the ANOVA summary for the A-A category across the Teacher-Pupil and Statement-Question dimensions.

TABLE 1

Three-Way ANOVA Table for Category A-A

| Source | Mean Square | df | Variance Component | Proportion of Variance Accounted for |
|---|---|---|---|---|
| Class (C) | 0.00618 | 9 | -0.0011 | 0.000 |
| Rater (R) | 1.0475 | 2 | 0.0517 | 0.876 |
| Day (D) | 0.00798 | 1 | 0.0002 | 0.003 |
| CR | 0.01238 | 18 | 0.0056 | 0.095 |
| CD | 0.00188 | 9 | 0.0002 | 0.004 |
| RD | 0.00227 | 2 | 0.0001 | 0.002 |
| Residual | 0.00121 | 18 | 0.0012 | 0.020 |

The most arresting feature of this table is that 87 percent of the variance of the scores in the A-A category comes from rater disagreement. In addition, nearly 10 percent of the variance is due to rater bias involving the class being rated. Finally, only 2 percent of the variance is unaccounted for, using this design.

Table 2 shows how the discrepancy (unaccounted) variance--$\sigma^2(\Delta)$--was computed. The variance of the discrepancy scores is 0.02. It might appear that an error of about 0.04 (95 percent level) is acceptable, but when we look at the generalizability coefficient, a different picture emerges.

The coefficient of generalizability is estimated by the ratio of the estimated universe-score variance to the expected observed-score variance. The expected observed-score variance is made up of all components where c is found: c, cR, cD, cRD. The c component is negative and is treated as zero. The other components total to 0.0022. The universe score is estimated by the c component, which is zero. The generalizability coefficient, then, is 0.0000/0.0022 = 0.0000. Four of the 11 classification variables in this study had generalizability coefficients of zero, a discouraging result to say the least.

TABLE 2

Components of Scores for a Two-Facet Analysis of Variable A-A

| $X_{cRD}$ | $M_c$ | cRD | Component of Variance | Frequency Within c | Contribution to $\sigma^2(\Delta)$ |
|---|---|---|---|---|---|
| $M$ | $M$ | | | | |
| $M_c^{\sim}$ | $M_c^{\sim}$ | | | | |
| $M_R^{\sim}$ | | $M_R^{\sim}$ | 0.0517 | 3 | 0.0172 |
| $M_D^{\sim}$ | | $M_D^{\sim}$ | 0.002 | 2 | 0.0001 |
| $M_{cR}^{\sim}$ | | $M_{cR}^{\sim}$ | 0.0056 | 3 | 0.0019 |
| $M_{cD}^{\sim}$ | | $M_{cD}^{\sim}$ | 0.0002 | 2 | 0.0001 |
| $M_{RD}^{\sim}$ | | $M_{RD}^{\sim}$ | 0.0001 | 6 | 0.0000 |
| $M_{cRD,e}^{\sim}$ | | $M_{cRD,e}^{\sim}$ | 0.0012 | 6 | 0.0002 |
| | | | | Total | 0.0194 = 0.02 |

Note: Capital letters as subscripts indicate that these components are considered fixed in this study, i.e., the universe of raters is the 3 raters used in the study, and the universe of days is the 2 days on which the classes were observed. The value of $\sigma^2(\Delta)$, then, applies to this study only; other values can be computed for differing conditions of R and D.

## Possible Conclusions from Generalizability Theory

Turning to some of the variables that had positive coefficients, one can see how the generalizability study using the multifacet approach can suggest how best to use an observation instrument in decision studies. Since the variance components, which are summed to obtain the values for computing the ratio, are divided by the number of times they are sampled in the design, it is possible to determine how the coefficient varies from design to design. The CV variable will be used for illustration.

Table 3 shows how the coefficient can assume different values when the number of raters and the number of observations (days) are varied. It is apparent that increasing the number of raters reduces the variance component due to raters, since that component is sampled more often. Increasing the number of days increases the generalizability coefficient.

TABLE 3

Values of $\sigma^2(\Delta)$ and the Generalizability Coefficient as a Function of Design in a Decision Study (Variable CV)

| Com-ponent | Variance Estimate | Contribution to Universe Score Variance | No. | Contribution to $\sigma^2(\Delta)$ for: (Contribution to expected observed variance for:) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Raters = | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 |
| | | | Days = | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Class | 0.0065 | 0.0065 | | - | - | - | - | - | - | - | - | - |
| | | | | (0.0065 | .0065 | .0065 | .0065 | .0065 | .0065 | .0065 | .0065 | .0065) |
| Rater | 0.0121 | -- | | 0.0121 | .0121 | .0121 | .0060 | .0060 | .0060 | .0040 | .0040 | .0040 |
| | | | | ( - | - | - | - | - | - | - | - | - ) |
| Day | 0.0000 | -- | | 0.0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 | .0000 |
| | | | | ( - | - | - | - | - | - | - | - | - ) |
| CR | 0.0010 | -- | | 0.0010 | .0010 | .0010 | .0005 | .0005 | .0005 | .0003 | .0003 | .0003 |
| | | | | (0.0010 | .0010 | .0010 | .0005 | .0005 | .0005 | .0003 | .0003 | .0003) |
| CD | -0.0009 | -- | | | | | variance treated as zero | | | | | |
| | | | | ( | | | variance treated as zero | | | | | ) |
| RD | -0.0005 | -- | | | | | variance treated as zero | | | | | |
| | | | | ( - | - | - | - | - | - | - | - | - ) |
| CRD,e | 0.0101 | -- | | 0.0101 | .0050 | .0030 | .0050 | .0030 | .0020 | .0030 | .0020 | .0010 |
| | | | | (0.0101 | .0050 | .0030 | .0050 | .0030 | .0020 | .0030 | .0020 | .0010) |
| | | 0.0065 | $\sigma^2(\Delta)$ = | 0.0232 | .0181 | .0165 | .0116 | .0091 | .0082 | .0077 | .0060 | .0055 |
| | | | Gen. Coeff. = | 0.37 | .52 | .60 | .54 | .68 | .75 | .64 | .76 | .82 |

Note: Generalizability Coefficient = $\frac{\text{Universe-score variance}}{\text{Expected observed-score variance}}$

The figure 0.82 indicates a respectable level of reliability, but it requires 3 raters looking at a class on 3 different occasions to obtain that value. Simply increasing the number of observations and using one rater does not provide sufficient reliability for firm scores.

The CV variable was the only one for which a coefficient greater than 0.8 was obtained (Co had a coefficient of 0.75, using 3 raters and 3 days); all the other variables were below 0.7.

The difficulty with this instrument appears to be that rater variability accounts for most of the variance in the observed scores. This finding might suggest that training procedures for the raters should be improved, or that the rating procedures are too involved and cumbersome for raters to use accurately and reliably. The instrument probably cannot be used without revision along one or both of these lines.

## References

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for single and multiple observations. New York: John Wiley & Sons (in press).

Flores-Hernandez, M. Teachers' class discussion variables related to student participation and opinion. Unpublished doctoral dissertation, Stanford University, 1970.