PSYCHOMETRIKA—VOL. 30, NO. 4
DECEMBER, 1965

GENERALIZABILITY OF SCORES INFLUENCED BY MULTIPLE SOURCES OF VARIANCE

GOLDINE C. GLESER, LEE J. CRONBACH, AND NAGESWARI RAJARATNAM

UNIVERSITY OF ILLINOIS*

Generalizability theory concerns the adequacy with which a "universe" score can be inferred from a set of observations. In this paper the theory is applied to a universe in which observations are classifiable according to two independent variable aspects of the measuring procedure. Several types of universe scores are developed and the variance components ascertained for each type. The composition of expected observed-score variance and the adequacy of inference to a particular type of universe score is a function of the procedure used in gathering data. A generalizability study provides estimates of variance components which can be used in designing an efficient procedure for a particular decision purpose.

Any measuring procedure involves one or more variable aspects not specified in the operational definition of the procedure. Whatever choice the experimenter makes regarding these aspects—within broad limits—the observations remain acceptable as a basis for decision-making. One ordinarily interprets a child's Stanford-Binet score in the same manner whether Form L or M is used, whether Smith or Jones administers it, and whether it is obtained on Tuesday or Friday. The observation is accepted because it is considered representative of a universe of observations that might have been made on the child. Even the most momentary and local observation is generalized in interpretation; at the least, it is treated as representative of reports by other observers who might have made the observation at that time and place.

Every observation is to be regarded, therefore, as a sample from a universe of possible observations, any one of which might have satisfied the investigator. The subject's mean score over all these potential observations is his "universe" score; this, if it were available, would be the ideal basis for interpretations or decisions. Since universe scores can only be inferred, it is necessary to know how well such inferences can be made from observed scores. Generalizability theory is concerned with this question. Cronbach

*This study was conducted under Grant M-1839 from the National Institute of Mental Health while the authors were on the staff of the University of Illinois. Dr. Rajaratnam shared responsibility for the technical report of July, 1961 on which this paper is based. The present revision was made subsequent to her death in 1963. The present addresses of the other authors are: Goldine C. Gleser, Department of Psychiatry, Central Clinic, Cincinatti, 29, Ohio; Lee J. Cronbach, School of Education, Stanford University.

et al. [4, 5] have presented the theory for the one-facet universe. In this paper we extend the theory to multifacet universes.

In the multifacet universe, observations vary in more than one respect. For example, an investigator weighing a person on a given day may want to generalize to a universe consisting of all weights that could be recorded for that person at various times on that day on all weighing machines of a specified type by any appropriate observer. We shall refer to each of the identifying aspects of the observations—time, instrument, observer, etc.—as a *facet* [9]. One particular *condition* from each facet is employed in making each observation. In the example, the particular conditions of observation might be 11:00 A.M., the weighing machine in Dr. King's office, observer Nurse Thompson. The facets, together with the subjects, are sources of variance in observed data. Whatever the conclusion the investigator is interested in, one or more of these sources of variance contribute unwanted variance or error.

Several writers have recognized that unwanted variance comes from many sources, and that each definition of error changes the meaning of the "reliability coefficient." The most systematic of the early papers within a correlational framework is that of Gulliksen [8]. Theoretical treatments by Thorndike [15] and Cronbach [3], among others, argue that "true score" and "error" are to be defined differently depending upon the investigator's interest; for each definition a different experimental procedure must be used to estimate reliability. Subsequent writers have extended the familiar single-facet reliability study to treat multiple sources of variance. Lindquist [10] and Medley and Mitzel [12] give general formulas for the reliability of composite scores obtained when conditions of observation are sampled from a multifacet universe. (See also [1, 11, 13, 14].)

We extend these studies by giving more explicit attention to the universe of observations over which generalization is intended and to the experimental design for collecting measurements. We distinguish between the generalizability study and the decision study [cf. 4]. The *generalizability* (G) *study* collects data for the purpose of estimating the components of variance of a measuring procedure; the *decision* (D) *study* collects data for the purpose of making decisions or interpretations. The published estimates of reliability for a college aptitude test, for example, are based on a G study; but college personnel men employ these estimates to judge the accuracy of the scores they collect (D study) to make decisions about applicants. Sometimes, of course, the same data serve both purposes.

In this paper, we discuss observations classified with respect to only two facets, but the same principles extend to more elaborate designs. We restrict ourselves to observations obtained under conditions randomly sampled from the universe. This random-sampling assumption is much weaker than the assumption that variances arising under different conditions are equal and

that intercorrelations for the pairs of conditions are equal. These strong equivalence assumptions have appeared, in one guise or another, in such previous treatments as those of Lindquist and Burt.

Notation and Assumptions

Letting the subscript p denote persons, and i and j denote the conditions of observation with respect to two facets of a universe, we have a datum, X_{pij}. We assume that the number of conditions for any facet is infinite, although our argument can readily be adapted to a facet having a finite number of conditions [e.g., 2]. We make these further assumptions. (i) The universe of possible observations is completely crossed, i.e., for every person there is a possible observation corresponding to each and every pair of conditions formed by taking one condition from each facet. (ii) The conditions of i and j employed in the reliability study and in the decision data are randomly and independently sampled from the universe. (iii) There are no order effects.

We distinguish between matched and independent observations. Observations are *matched* with respect to a facet when one or more conditions are selected from that facet and observations are made on every subject under each selected condition. Example: all subjects respond to the same items, or are rated by the same judges. The term "matched" could be translated as "facet and persons crossed." Observations are *independent* with respect to a facet when conditions for that facet are selected independently for each observation without regard to subject or to the condition of other facets. Example: high school grades for job applicants, where only by chance would any two grades have been assigned by the same teacher. Observations sampled from a multifacet universe may be matched with respect to some facets and independent with respect to others. The various experimental designs that result will be discussed more fully below.

We shall be interested in certain means, M_p, M_{pi}, M_i, M_{ij}, etc.; each of these is the expectation of a subset of the possible scores. In our notation, the subscript applied to M indicates the parameters held constant in defining the subset. For example,

$$M_p = E(X_{pij} \mid p) \quad \text{and} \quad M_{ij} = E(X_{pij} \mid ij).$$

These means allow us to define components of the observed score. If an observation is made under condition i, for example, we expect the score to be increased over M by the component $M_i - M$, the "effect" associated with condition i. X_{pij} may be described as a sum of components:

$$\begin{aligned}(1) \qquad X_{pij} = {} & M + (M_p - M) + (M_i - M) + (M_j - M) \\ & + (M_{pi} - M_p - M_i + M) + (M_{pj} - M_p - M_j + M) \\ & + (M_{ij} - M_i - M_j + M) + e_{pij}.\end{aligned}$$

The term e_{pij} is a residual so that (1) is a tautology defining e_{pij}.

We use the following notation for expected mean squares (variances).

$$
\begin{aligned}
&V_{M_p} = E(M_p - M)^2; \text{ analogously for } V_{M_i} \text{ and } V_{M_j} . \\
&V_{M_{pi}} = E(M_{pi} - M_p - M_i + M)^2; \text{ analogously for } V_{M_{pj}} \text{ and } V_{M_{ij}} . \\
&V_{e_{pij}} = E(e_{pij}^2), \quad \text{since} \quad E(e_{pij}) = 0. \\
&V_{X_{pij}} = E(X_{pij} - M)^2.
\end{aligned} \tag{2}
$$

Other variance notations will be introduced as needed.

The variance of X_{pij} over population and universe is

$$
V_{X_{pij}} = V_{M_p} + V_{M_i} + V_{M_j} + V_{M_{pi}} + V_{M_{pj}} + V_{M_{ij}} + V_{e_{pij}}. \tag{3}
$$

That is, the variance of single observations over all p, i, and j is the sum of the variances contributed by each of the components of (1). Extension of (3) to universes with additional facets requires the addition of terms for the main effects of these other facets, and terms for their interactions with each other and with p, i, and j.

The score used in making decisions is often a composite of a number of observations. In this paper we assume that in the decision data each person's score is the average of scores obtained under k' conditions of i and m' conditions of j—e.g., ratings by m' judges on k' scales. We designate these sets of conditions by I and J, respectively. (Since k' and m' may be 1, there is no loss of generality.)

To this point we have been discussing the universe of all possible observations. The investigator may propose to generalize his D data to this universe or he may generalize over only certain selected facets, conditions of the other facets being fixed. His choice of a *universe of generalization* depends on his intended interpretation. Thus while measures of weight may vary with respect to occasion, instrument, and observer, an investigator who wants to study hour-to-hour fluctuations in weight would define the universe of generalization as including observations from all instruments and all observers, but on a fixed occasion. Another investigator who will use the same instrument throughout his entire study might propose to generalize over observers and over some range of occasions, but not over instruments. The universe score is defined differently for each universe of generalization. M_p is the universe score when the universe of generalization coincides with the universe of possible observations. M_{pI}· is the universe score when conditions I^* are fixed and generalization is over facet j. The *universe-score variance* is the variance of universe scores over the population of persons.

The problem as it presents itself ideally to an investigator is this. He has in mind a universe to which he proposes to generalize. This universe defines the "wanted" components of score variance. He hopes to plan a decision study that will produce scores from which accurate generalizations can be made.

Therefore, he conducts a G study, estimates the several components of variance, and from this information plans a D study such that unwanted components of score variance are minimized. Practical constraints, however, often restrict the investigator to a particular design for the D study, or to a limited class of not-necessarily-ideal designs. Whatever the design, he needs to estimate the accuracy of generalization of his decision data.

In the following discussion we first consider various types of universe of generalization and show the composition of universe-score variance for each type. Second, we determine the components of observed-score variance for various designs that may be used in D studies. Components included in observed-score variance but not in the variance of the appropriate universe score constitute error variance for that particular design and interpretation. A "coefficient of generalizability," which is the ratio of universe-score variance to expected observed-score variance, indicates how well one can generalize from one particular type of D data to one type of universe score. Next, we consider what information different types of G studies can give about the components of variance, and finally how such information is used to design a D study.

Universe Scores and Their Variance Components

Let us now consider several types of universe score to which generalization may be intended and identify the components of universe-score variance in each case. The material is summarized in Table 1.

TABLE 1

Composition of the Universe-Score Variance for Alternative Universes of Generalization

	Universe of Generalization	Composition of Variance*
A	All conditions of both i and j	V_{M_p}
B	All conditions of j crossed with k' fixed conditions of i	$V_{M_p} + (1/k')\ V_{M_{pi}}$
B'	All conditions of i crossed with m' fixed conditions of j	$V_{M_p} + (1/m')\ V_{M_{pj}}$
C	All conditions of j crossed with k' conditions of i fixed within person p	$V_{M_p} + (1/k')\ V_{M_{pi}} + (1/k')\ V_{M_i}$
C'	All conditions of i for m' conditions of j fixed within person p	$V_{M_p} + (1/m')\ V_{M_{pi}} + (1/m')\ V_{M_j}$

*The universe score is taken to be the expectation of the average score over observations. If the universe score is the expectation of the total for $k'm'$ items, all variance components are multiplied by $(k'm')^2$.

Universe A: Two Facets Variable

Sometimes the investigator is interested in generalizing over all identified facets of the universe. For example, if he is interested in determining the typical aggressiveness displayed by a subject regardless of the observer or of the persons with whom the subject is interacting, then the universe score is the expected value of all observations that could be made under any conjunction of observer $\times$ companion. For each observer i and companion j there is a potential score X_{pij} . The subject's universe score is M_p ; the universe-score variance is V_{M_p} . The universe score is the same whether the observed score is a single observation or the average of many observations. (Where the observed score is to be based on a total rather than an average, the variance given by our equations will need to be multiplied by k'^2, m'^2, or $(k'm')^2$, depending on whether summation is over conditions of i, j, or both.)

Universe B: One Facet Variable, One Fixed

Sometimes the investigator limits generalization to one facet. For example, in a study of job proficiency, he may wish to treat the scores as representative of ability, anywhere in a long time span; i.e., he will generalize over occasions. But he may use a fixed set of tasks for all subjects, with no generalization intended over a broader universe of tasks. The universe score is then the expected value $M_{pI\cdot}$ of observations on the set of tasks I^* on all the possible occasions j. The universe-score variance is $E(M_{pI\cdot} - M_{I\cdot})^2$. Fixing the conditions on one facet reduces the universe of generalization to a subset of the universe of possible observations; the universe score is the expected score for the person on the remaining facet.

It frequently happens that the fixed conditions for the decision study are unknown at the time the G study is conducted. For example, an investigator is interested in rating the morale of persons who will be observed on the same specific occasion, but the occasion might be any day on which the observation happens to be made. He wants to estimate the generalizability of the ratings made on that unspecified day, the universe of generalization consisting of ratings by all possible observers. If I^*, the subset of k' conditions of i, is unknown, the investigator can best use the *expected* universe-score variance for an I selected at random, i.e., $E(M_{pI} - M_I)^2$. Since

$$M_{pI} - M_I = (M_p - M) + \frac{1}{k'} \sum_{1}^{k'} (M_{pi} - M_p - M_i + M), \tag{4}$$

$$E(M_{pI} - M_I)^2 = V_{M_p} + \frac{1}{k'} V_{M_{pi}} .$$

Hence with one facet fixed but unspecified, and the other facet variable, the universe-score variance includes, along with V_{M_p} , the component $V_{M_{pi}}$

which represents the expected variance in response under condition i not attributable to variance in M_p .

There is a universe of generalization B' symmetric to B, which has conditions of facet i variable and J^* fixed. Universe-score variance for B' is given by (4) with i replaced by j and k' replaced by m'.

Universe C: One Facet Variable, One Facet Fixed within Person p

In universe B the same fixed conditions were used with every subject. It is conceivable that the investigator will instead allow the conditions of facet i to vary from subject to subject, but will hold the conditions of i fixed for any one subject. For example, he might ask each subject to rate his spouse on items related to adjustment. While the spouse is fixed for the subject, each subject has a different spouse. Generalization is intended over some other facet, such as items in this example.

The universe of generalization for each person is defined by the k' fixed conditions of i (call this set I_p) and the variable conditions of j. The universe score is M_{pI_p} . The expected value of the universe score in this case is M, since I_p ranges over all values of i as p varies over the population. The universe-score variance is

$$(5) \qquad E(M_{pI_p} - M)^2 = E\left[\frac{1}{k'}\sum_{i_p=1}^{k'}(M_{pi_p} - M)^2\right]$$

$$= \frac{1}{k'}V_{M_{pi}} + \frac{1}{k'}V_{M_i} + V_{M_p} .$$

Again there is a universe of generalization C', with fixed conditions J_p and variable conditions i. The universe-score variance is given by (5) with appropriate modification.

Components of Expected Observed-Score Variance

We now consider the observed-score variance in the decision data. This variance is a function of the variances of the means of the particular judges, items, etc. used in the study, rather than universe variance components which derive from all conditions. Usually, however, the data of the G study are not collected under the precise combination of conditions that will appear in the decision study. Hence the *expected* observed variance for a randomly sampled set of functions is the best available estimate of the observed variance for the D data. This variance is a function of universe and population components.

The composition of observed-score variance in the D study depends upon its experimental design. Decision data may be gathered according to designs ranging from complete matching (every person tested under the same set of crossed conditions) to procedures in which conditions are selected independ-

ently for each observation. Once a design is specified we can determine which components contribute to the expected variance of observed scores in decision data collected according to that design. Estimates of these components then yield estimates of the unknown observed-score variance.

We assume that decisions are to be based on scores obtained by averaging $k'm'$ observations for each person. (In some examples to follow, k' or m' is one.) For each of the designs applicable to the two-facet universe, we shall examine the algebraic composition of the observed variance—rather, of its expectation—over all sets of $k'm'$ observations per person that might be sampled from the universe.

Table 2 summarizes the statements about composition of expected observed variance to be developed below. Designs 1, 2, 4, and 5 permit generalization to universe A or B; while it is technically posible to generalize to C, such use is unlikely. Design 3 is primarily useful in generalizing to A. Designs 6 and 7 permit generalization to A or C, but not to B because I is not fixed. Design 8 is useful only in generalizing to A.

Design 1: *Matched Data, i and j Crossed*

In design 1, decision data are matched over persons and conditions of facets i and j are crossed. The investigator selects, by random sampling, k' conditions of i and m' conditions of j, and each person is observed under all $k'm'$ paired conditions. The resulting X_{pij} form a complete three-way cross-classification. An example would be a study in which a checklist of symptoms of tension is filled out by a set of raters, each of whom examines all subjects.

The observed score, $X_{pI^*J^*}$, for person p is the average of $k'm'$ observations for him. It can be expressed as follows

$$
\begin{aligned}
(6)\quad X_{pI^*J^*} &= \frac{1}{k'm'} \sum_{i=1}^{k'} \sum_{j=1}^{m'} X_{pij} \\
&= M + (M_p - M) + \frac{1}{k'} \sum_{i=1}^{k'} (M_i - M) + \frac{1}{m'} \sum_{j=1}^{m'} (M_j - M) \\
&\quad + \frac{1}{k'} \sum_{i=1}^{k'} (M_{pi} - M_p - M_i + M) + \frac{1}{m'} \sum_{j=1}^{m'} (M_{pj} - M_p - M_j + M) \\
&\quad + \frac{1}{k'm'} \sum_{i=1}^{k'} \sum_{j=1}^{m'} (M_{ij} - M_i - M_j + M) + \frac{1}{k'm'} \sum_{i=1}^{k'} \sum_{j=1}^{m'} e_{pij} ,
\end{aligned}
$$

where the i and j are conditions within sets I^* and J^*.

The expectation of $X_{pI^*J^*}$ over the population of persons is

$$
\begin{aligned}
(7)\quad M_{I^*J^*} &= E(X_{pI^*J^*}) = M + \frac{1}{k'} \sum_{i=1}^{k'} (M_i - M) \\
&\quad + \frac{1}{m'} \sum_{j=1}^{m'} (M_j - M) + \frac{1}{k'm'} \sum_{i=1}^{k'} \sum_{j=1}^{m'} (M_{ij} - M_i - M_j + M).
\end{aligned}
$$

TABLE 2

Composition of Expected Observed Variance for Decision Data Obtained by Various Designs

Experimental Design		Components and Their Weights*						
Facet i	Facet j	V_{Mp}[a]	$V_{M_{pi}}$[b]	V_{M_i}[c]	$V_{M_{pj}}$[b']	V_{M_j}[c']	$V_{M_{ij}}$	$V_{e_{pij}}$
1 Matched**	Matched; crossed with i	1	$1/k'$		$1/m'$			$1/k'm'$
2 Matched	Matched; nested in i	1	$1/k'$		$1/k'm'$			$1/k'm'$
2' Matched; nested in j	Matched	1	$1/k'm'$		$1/m'$			$1/k'm'$
3 Matched; independent of j	Matched; independent of i	1	$1/k'm'$		$1/k'm'$			$1/k'm'$
4 Matched	Nested in p; crossed with i	1	$1/k'$		$1/m'$	$1/m'$	$1/k'm'$	$1/k'm'$
4' Nested in p; crossed with j	Matched	1	$1/k'$	$1/k'$	$1/m'$		$1/k'm'$	$1/k'm'$
5 Matched	Independent	1	$1/k'$		$1/k'm'$	$1/k'm'$	$1/k'm'$	$1/k'm'$
5' Independent	Matched	1	$1/k'm'$	$1/k'm'$	$1/m'$		$1/k'm'$	$1/k'm'$
6 Nested in p	Nested in p; crossed with i	1	$1/k'$	$1/k'$	$1/m'$	$1/m'$	$1/k'm'$	$1/k'm'$
7 Nested in p	Independent	1	$1/k'$	$1/k'$	$1/k'm'$	$1/k'm'$	$1/k'm'$	$1/k'm'$
7' Independent	Nested in p	1	$1/k'm'$	$1/k'm'$	$1/m'$	$1/m'$	$1/k'm'$	$1/k'm'$
8 Independent	Independent	1	$1/k'm'$	$1/k'm'$	$1/k'm'$	$1/k'm'$	$1/k'm'$	$1/k'm'$

*Each score is the average of observations under k' conditions of i and m' conditions of j.

**In each case the term "matched" could be translated as "crossed with p."

[a]Always included in universe-score variance.

[b]Always included in universe-score variance if I is fixed in the universe of generalization.

[c]Included in universe-score variance only if I is fixed within persons.

[b']Same as (b) for symmetric cases where J is fixed.

[c']Same as (c) for symmetric cases where J is fixed within persons.

This equation includes those terms of (6) that are constant over persons, all others having an expectation of zero as a consequence of their definitions. The population variance, $E(X_{pI^*J^*} - M_{I^*J^*})^2$, involves both the variances and covariances of the specific means, but the *expected* variance of $X_{pI^*J^*}$ for an unspecified set of k' matched conditions of IJ is

$$V_{X_{pIJ}} = V_{M_p} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{m'} V_{M_{pj}} + \frac{1}{k'm'} V_{e_{pij}} . \tag{8}$$

The control introduced by matching data has eliminated variance due to facet i, facet j, and their interaction from expected observed-score variance, because these effects are the same for all subjects.

Design 2: Data Matched, j Nested in i

In design 2, the decision data are matched over persons, but the conditions of j are nested in i. That is, k' conditions of i are selected and then m' conditions of j are independently selected for each such i. Example: each of several specified raters (I^*) observes all subjects on two occasions, but the raters do not make their observations simultaneously; occasions vary from rater to rater. In this design i and j are confounded. The observed score for person p, which we denote as $X_{pI^*J_i^*}$, is the average of $k'm'$ observations. Then the expected observed-score variance is given by

$$V_{X_{pIJ_i}} = V_{M_p} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{k'm'} V_{M_{pj}} + \frac{1}{k'm'} V_{e_{pij}} . \tag{9}$$

The expected observed-score variance for design 2 differs from that of design 1 only in that $V_{M_{pj}}$ is divided by $k'm'$ instead of m'. Thus the contribution from this source of variance is reduced here.

Design 3: Data Matched, Conditions of i and j Independent

In design 3 decision data are matched over persons, but conditions of i and j are selected randomly, independently of each other. Thus some number of paired conditions $i \times j$ are selected and each person is observed under every combination. Facets i and j are confounded so that this design resembles the unifacet model, $i \times j$ forming a single complex facet. An example would be the administration of four alternate forms of a test, one on each of four different occasions, to a group of subjects. The number of paired conditions will be represented by $k'm'$ to enable comparison of formulas resulting from various designs. However, in this design there are $k'm'$ conditions of i and $k'm'$ conditions of j involved. The expected observed-score variance is

$$V_{X_{pIJ}} = V_{M_p} + \frac{1}{k'm'} (V_{M_{pi}} + V_{M_{pj}} + V_{e_{pij}}). \tag{10}$$

Since this design minimizes the contributions of both $V_{M_{pi}}$ and $V_{M_{pj}}$ to expected observed variance, it is the most efficient design for generalization to universe A whenever both $V_{M_{pi}}$ and $V_{M_{pj}}$ are relatively large.

Design 4: Matched on i, j Nested within Persons, i and j Crossed

In the fourth design, certain values of i are used as conditions for observing all persons. Values of j are selected independently for each person, producing sets J_p . Each j in J_p is paired with each i to obtain one observation. For each person there is a k' by m' matrix of scores. In this design, p and j are completely confounded, j being nested within p. An example is the application of the rating scale of symptoms of tension (i) to subjects, each of whom is judged by a different rater or raters (j). Another example: certain judges (i) working simultaneously observe each subject in turn, but the occasions (j) differ from subject to subject.

The observed score in this design may be denoted $X_{pI^*J_p}$. Equation (6) indicates the composition of observed score if j is replaced by j_p . The expectation of $X_{pI^*J_p}$ over persons is simply

$$M + \frac{1}{k'} \sum_{i=1}^{k'} (M_i - M),$$

all other terms vanishing.

Over all sets I that might be used in decision-making, the expected observed-score variance is

$$(11) \qquad V_{X_{pIJ_p}} = V_{M_p} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{m'} V_{M_{pj}} + \frac{1}{m'} V_{M_j} + \frac{1}{k'm'} V_{M_{ij}} + \frac{1}{k'm'} V_{e_{pij}} .$$

Two components of variance associated with j are added when the design is changed from completely matched to matched-within-persons-on-j (cf. Table 2).

Design 5: Matched on i, Independent on j

In the fifth design all persons are observed under the same conditions I^*. But for each $p \times i$ combination, different values of j are selected, independent of the j for any other p or i. Example: the same group of raters (i) observe all subjects, but observations occur on different occasions (j) for each subject-rater pair. This allows a two-way classification (subjects $\times$ raters) with m' entries (occasions) per cell. The conditions j are different for each p,i pair and hence we shall designate them as j_{pi} .

The observed score obtained by averaging all $k'm'$ scores for p is denoted X_{pI^*J} . The expectation (over all sets of I) of the observed-score variance is

$$(12) \qquad V_{X_{pIJ}} = V_{M_p} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{k'm'} V_{M_{pj}} + \frac{1}{k'm'} V_{M_j} + \frac{1}{k'm'} V_{M_{ij}} + \frac{1}{k'm'} V_{e_{pij}} .$$

The expected observed-score variance for this design contains the same components as design 4 (cf. Table 1), but the effect of V_{M_j} and $V_{M_{pj}}$ is reduced since each of these is divided by the total number of observations $k'm'$.

Design 6: i and j Crossed, Both Nested within Persons

The sixth case is that in which k' conditions of i and m' conditions of j are selected independently for each person and all pairs are formed. Observations of persons p are made under each of the $k'm'$ pairs selected for him. For example, acceptance of peers is studied by having the subject rate a limited number of peers (i), all on the same occasions (j), neither peers nor occasions being the same for different subjects.

The observations form a k' by m' matrix for each subject, but the conditions of i and j are both dependent on p. The average of these observations, $X_{pI_pJ_p}$, may be expressed as in (6), replacing i and j by i_p and j_p, respectively. Assuming that the conditions of i and j are randomly sampled, the expectation of these scores in the population is M, since all other components vary with p and hence have expectations of zero. The variance of observed scores in the population does not differ from the expected observed variance in this case. (This is also true of designs 7 and 8, below.)

$$(13) \qquad V_{X_{pI_pJ_p}} = \underset{p}{E}(X_{pI_pJ_p} - M)^2$$

$$= V_{M_p} + \frac{1}{k'} V_{M_i} + \frac{1}{m'} V_{M_j} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{m'} V_{M_{pj}} + \frac{1}{k'm'} V_{M_{ij}} + \frac{1}{k'm'} V_{e_{pij}} .$$

Design 6 (cf. Table 2) allows all components of X_{pij} to enter the observed variance. It therefore yields the largest observed variance of all designs, and hence the lowest generalizability, other things being equal.

Design 7: i Nested within Persons, j Independent

For each person, k' conditions of i are selected independently, and for each consequent $p \times i$ combination, m' conditions of j are selected independently. Example: each subject is judged by two raters (i) on each of several occasions (j); the raters differ from subject to subject, and the occasions regarding which each rater reports differ from rater to rater. The data can be arranged in n rows of k' cells each, with m' observations per cell. The variance of observed scores, $X_{pI_pJ_{pi}}$, is

$$(14) \qquad V_{X_{pI_pJ_{pi}}} = V_{M_p} + \frac{1}{k'} V_{M_i} + \frac{1}{k'} V_{M_{pi}} + \frac{1}{k'm'} V_{M_j} + \frac{1}{k'm'} V_{M_{pj}} + \frac{1}{k'm'} V_{M_{ij}} + \frac{1}{k'm'} V_{e_{pij}} .$$

This variance includes all components, but compared to that of design 6

where both facets are nested within persons the contributions of V_{M_i} and $V_{M_{pi}}$ are reduced by a factor of k'. It differs from the variance in the matched-independent case (design 5) only by the added component V_{M_i}. (See Table 1.)

Design 8: Both i and j Independent

The last design is that in which the conditions of i and j are selected randomly and independently for each observation, $k'm'$ observations being made on each person. In effect, $k'm'$ conditions of ij are randomly and independently selected for each person. An example would be speech samples of a person as he talked to various companions, no companion appearing on more than one occasion. Another example is high school grades of job applicants considered as reports from various teachers, each having observed the applicant on different occasions. In this design, p, i, and j are completely confounded.

The variance of observed scores $X_{p[IJ]_p}$ (where IJ is bracketed to emphasize that it is the ij combination that is selected) is given by

$$(15)\quad V_{X_{p[IJ]_p}} = V_{M_p} + \frac{1}{k'm'}\left[V_{M_i} + V_{M_j} + V_{M_{pi}} + V_{M_{pj}} + V_{M_{ij}} + V_{e_{pij}}\right].$$

Since, for this design, increasing the number of conditions of either facet implies utilizing more conditions of both facets, every variance term except V_{M_p} is divided by the total number of observations per person, $k'm'$.

Coefficient of Generalizability Expressed in Terms of Variance Components

We define the accuracy of generalization as the ratio of expected universe-score variance to expected observed-score variance. This variance ratio, which we shall call VR, can be shown to be a lower bound to the expected squared correlation between observed score and universe score. For any particular universe of generalization and decision-study design, VR can be expressed in terms of a ratio of the appropriate variance components. When these variance components can be estimated from a G study, an estimate of VR can be obtained.

Without for the moment considering how variance components are estimated, we may express VR for any particular universe of generalization and decision-study design by an appropriate selection of numerator and denominator from Tables 1 and 2, respectively. Not all possible ratios are meaningful since a certain decision-study design may be incompatible with one or more of the universes of generalization. For example, if one desired to generalize to universe B or B' with I or J fixed, he would not employ a D study in which the conditions of i and j vary from person to person (designs 6, 7, 7', and 8). Similarly, designs 1, 2, 2', 5, and 8 are not compatible with generalization to universe C or C'.

Several of the ratios formed in this way have appeared in the literature before; for example, Lindquist's [10] formula 180 applies to design 1 and universe A, whereas his formulas 177 and 178 apply to design 1 with universe B and B', respectively. Nested designs have been ignored in the previous literature on reliability with the one exception of Lindquist's formula 181.

Estimation of Components from Matched G Data

Whenever possible, data for a generalizability study should be collected by a matched design with conditions of the facets crossed and with two or more conditions used for each facet. Then no facets are confounded and all components can be estimated. This not only gives a more complete understanding of the sources of error in inferring universe scores from observed score, but permits an investigator to design his decision study for greatest precision with any fixed cost of data collection [10]. Regardless of the design the investigator ultimately uses for a D study, the completely crossed G study provides all the information he could need for evaluating VR.

It is necessary to assume that the conditions of i to be employed in the D study, and the conditions of i employed in the G study, are random samples from the same universe; likewise for j and for p. G data are collected for n persons, on k conditions of i and m conditions of j, where k and m are greater than one. A three-way analysis of variance gives the following mean squares.

Source	df	Sum of Squares	Mean Square
Persons	$n-1$	SS_p	$MS_p = SS_p/(n-1)$
Facet i	$k-1$	SS_i	$MS_i = SS_i/(k-1)$
Facet j	$m-1$	SS_j	$MS_j = SS_j/(m-1)$
Persons $\times$ Facet i	$(n-1)(k-1)$	SS_{pi}	$MS_{pi} = SS_{pi}/(n-1)(k-1)$
Persons $\times$ Facet j	$(n-1)(m-1)$	SS_{pj}	$MS_{pj} = SS_{pj}/(n-1)(m-1)$
Facet $i \times$ Facet j	$(k-1)(m-1)$	SS_{ij}	$MS_{ij} = SS_{ij}/(k-1)(m-1)$
Residual	$(n-1)(k-1)(m-1)$	SS_r	$MS_r = SS_r/(n-1)(k-1)(m-1)$

The variance components may be estimated by the following formulas.

$$
\begin{aligned}
\hat{V}_{epij} &= MS_r\,, \\
\hat{V}_{Mpi} &= (MS_{pi} - MS_r)/m, \\
\hat{V}_{Mpj} &= (MS_{pj} - MS_r)/k, \\
\hat{V}_{Mij} &= (MS_{ij} - MS_r)/n, \\
\hat{V}_{Mj} &= (MS_j - MS_{pj} - MS_{ij} + MS_r)/nk, \\
\hat{V}_{Mi} &= (MS_i - MS_{pi} - MS_{ij} + MS_r)/mn, \\
\hat{V}_{Mp} &= (MS_p - MS_{pi} - MS_{pj} + MS_r)/km.
\end{aligned} \tag{17}
$$

TABLE 3

Illustrative Data for a Completely Matched G Study

Persons	I		II		III		IV		V		VI		T_p
	J_1	J_2	J_1	J_2	J_1	J_2	J_1	J_2	J_1	J_2	J_1	J_2	
a	2	1	2	2	2	0	4	1	2	1	2	1	20
b	2	3	2	0	0	0	3	3	3	3	1	2	22
c	1	3	3	6	1	2	2	1	2	3	2	1	27
d	4	1	2	0	2	0	3	1	3	3	2	1	22
e	4	5	3	2	3	4	5	5	3	3	2	2	41
f	2	1	2	0	0	0	1	2	1	1	2	1	13
g	2	1	3	3	0	0	1	2	2	1	1	1	17
h	1	2	2	0	0	0	2	1	2	1	2	1	14
i	3	4	4	3	0	0	3	3	2	3	0	1	26
j	3	4	3	1	0	1	2	3	2	2	3	1	25
k	2	2	3	3	0	1	3	1	3	5	2	4	29
l	3	3	4	4	1	1	6	6	4	6	2	4	44
Total	29	30	33	24	9	9	35	29	29	32	21	20	300

Variance Analysis

Source	df	Sum of Squares	Mean Square
Persons	11	84.1667	7.6515
Judges	1	1.0000	1.0000
Items	5	64.6667	12.9333
P × J	11	19.5000	1.7727
P × I	55	81.1667	1.4758
I × J	5	4.3333	.8667
Residual	55	34.1666	.6212
Total	143	289.0000	

An example of the analysis of completely matched reliability data is given in Table 3. Two psychiatrists each rated patients on six items at the time of interview. (The number of patients, 12, is much smaller than one would desire. The small k and m also lead to a large sampling error in the variance estimates.) The items, scaled 0 to 6, referred to the intensity of symptoms of anxiety (e.g., agitation, pupillary dilation). The population variance estimates obtained from this sample are as follows.

Variance due to differences in the anxiety of patients over all raters and symptoms $\hat{V}_{M_p} = 0.419$

Variance due to differences in symptom means over all raters	$\hat{V}_{M_i} = 0.471$
Variance due to differences in average judgment of raters	$\hat{V}_{M_j} = -.014$
Variance due to differences in patient's standing from one symptom to another	$\hat{V}_{M_{pi}} = 0.427$
Variance due to differences among raters in their over-all judgments of specific patients	$\hat{V}_{M_{pj}} = 0.192$
Variance due to differences in symptom means from rater to rater	$\hat{V}_{M_{ij}} = 0.021$
Residual variance	$\hat{V}_{e_{pij}} = 0.621$

The large variance components are those arising from M_p, M_i, M_{pi}, and the residual. The variance of M_p reflects individual differences, probably the information of greatest interest to the investigator. The large $\hat{V}_{M_i}$ indicates that differences in symptom means are substantial. Whether a group of individuals receives high or low absolute scores depends on the symptoms used to assess anxiety. $\hat{V}_{M_{pi}}$ is also large, indicating that persons may score relatively high on some symptoms of anxiety and low on others. The person's standing on any one symptom therefore has limited value as an indication of his standing on the universe of symptoms. The small variance for M_j indicates that raters differ very little in their use of the scale. ($\hat{V}_{M_j}$ is in fact negative; such a value could arise only from a zero or small positive variance in the universe.) $\hat{V}_{M_{pj}}$ shows that raters vary in their over-all rating of the same patients, but this variation is relatively small. The residual variance is large, i.e., there is considerable variance among observations not accounted for by main effects associated with persons, judges, or symptoms, nor by their pairwise interactions. This residual of course includes all the "random errors" of observation and judgment and the triple interaction $p \times i \times j$.

The sizes of these components imply that it is especially important, if generalization to universe A, B', or C' is intended, to increase k', the number of symptoms rated in the decision data. This reduces the contributions from M_i, M_{pi}, and e_{pij}, all of which are irrelevant variance for these purposes. Another possibility to consider, as can be seen from Table 3, is to redefine the universe by dropping symptom III, which is evidently not closely related to the other symptoms.

Whatever the design to be used in collecting decision data, we may substitute appropriate estimates for the variances required according to Tables 1 and 2. Suppose, for example, matched decision data will be obtained using six symptoms and two judges (design 1, $k' = 6$, $m' = 2$), and generalization is intended to the universe of judges for a fixed set of six symptoms (universe B). We then estimate the ratio of "wanted" (universe-score) variance to observed variance to be

$$\frac{12^2[.419 + (.427/6)]}{12^2[.419 + (.427/6) + (.192/2) + (.621/12)]} = \frac{.490}{.638} = .768.$$

For generalization over both symptoms and judges (universe A), however, VR is only .657. That is to say, the broader generalization (and probably the one of greater scientific significance) involves greater error than the more limited generalization.

If it is not feasible for this investigator to match judges in collecting decision data, so that different pairs of judges rate each person (symptoms still being matched), then he would use row 4 of Table 2 to obtain the denominator for VR. It appears that control on judges can be relaxed with little loss in accuracy of generalization, since $\hat{V}_{M_j}$ is negative and $\hat{V}_{M_{ij}}$ is very small.

Variance Estimates Obtained from Other Experimental Designs

While there are obvious advantages in studying generalizability by means of a completely matched design, such a study is not always practicable. We shall therefore discuss other designs, indicating what variance components

TABLE 4

Components (or Combinations of Components) That Can Be Estimated from Various G-Study Designs

Design for G Study		Components Estimated*						
Facet i	Facet j	V_{M_p}	$V_{M_{pi}}$	V_{M_i}	$V_{M_{pj}}$	V_{M_j}	$V_{M_{ij}}$	$V_{e_{pij}}$
1. Matched	Matched; crossed with i	×	×	×	×	×	×	×
2. Matched	Matched; nested in i	×	×	×	3	2	2	3
2′. Matched; nested in j	Matched	×	3	1	×	×	1	3
3. Matched	Matched; independent of i	×	3	2	3	2	2	3
4. Matched	Nested in p; crossed with i	×	×	×	2	2	3	3
4′. Nested in p; crossed with j	Matched	×	1	1	×	×	3	3
5. Matched	Independent	×	×	×	3	3	3	3
5′. Independent	Matched	×	3	3	×	×	3	3
6. Nested in p	Nested in p; crossed with i	×	1	1	2	2	3	3
7. Nested in p	Independent	×	1	1	3	3	3	3
7′. Independent	Nested in p	×	3	3	2	2	3	3
8. Independent	Independent	×	3	3	3	3	3	3

*In any row, components designated by the same numeral are tied to each other and cannot be estimated separately.

TABLE 5

Applicability of Various G-Study Designs to Estimate Expected Observed Variance in Decision Data

Design for G Study	Design for D Study											
	1	2	2′	3	4	4′	5	5′	6	7	7′	8
1	×	×	×	×	×	×	×	×	×	×	×	×
2		×					×			×		×
2′			×					×			×	×
3				×								×
4					×		×		×	×	×	×
4′						×		×	×	×	×	×
5							×			×		×
5′								×			×	×
6									×	×	×	×
7										×		×
7′											×	×
8												×

An × indicates that the G design provides the information required to estimate generalizability for the D study.

they can estimate and how they can be used to analyze generalizability for various types of decision data. In designs other than design 1, confounding reduces us to estimating certain variance components in combination. The components that can be estimated separately from any given design are indicated by an × in Table 4; a common numeral indicates components that can be estimated only in combination. In design 4, for example, p and j are confounded; as a result the estimation procedure produces an estimate of $V_{M_i} + V_{M_{pi}}$ and another for $V_{M_{ij}} + V_{e_{pij}}$.

The use of this information is indicated in Table 5. This is based on a comparison of Table 4—what can be estimated—with Table 1—what one desires to estimate. Note that for any design in the D study, the same design used in the G study (with or without $k = k'$, $m = m'$) estimates the pertinent variances and VR.

We shall describe the estimation procedure more fully for only one design: design 4 in which data are matched on i and nested within persons on j, with i and j crossed. Assume that k conditions of facet i are used throughout, m conditions of facet j are selected randomly and independently for each subject, i and j are crossed, and that n randomly selected subjects are observed. Then the analysis of variance given on page 413 can be performed.

In this formulation, a dot indicates summation over the subscript it

Analysis of Variance When k Conditions of Facet i Are Used Throughout, m Conditions of Facet j Are Selected Randomly and Independently for Each Subject, i and j Are Crossed, and n Randomly Selected Subjects Are Observed

Source	df	Sum of Squares	Mean Square
Persons	$n - 1$	$\mathrm{SS}_p = \frac{\sum_p X_{p..}^2}{mk} - \frac{X_{...}^2}{mnk}$	$\mathrm{MS}_p = \frac{\mathrm{SS}_p}{n - 1}$
Facet i	$k - 1$	$\mathrm{SS}_i = \frac{\sum_i X_{.i.}^2}{mn} - \frac{X_{...}^2}{mnk}$	$\mathrm{MS}_i = \frac{\mathrm{SS}_i}{k - 1}$
Facet j within persons	$n(m - 1)$	$\mathrm{SS}_{jwp} = \sum_p \left[\frac{\sum_j X_{p.j}^2}{k} - \frac{X_{p..}^2}{mk} \right]$	$\mathrm{MS}_{jwp} = \frac{\mathrm{SS}_{jwp}}{n(m - 1)}$
Persons $\times$ Facet i	$(n - 1)(k - 1)$	$\mathrm{SS}_{pi} = \frac{\sum_p \sum_i X_{pi.}^2}{m} - \mathrm{SS}_p - \mathrm{SS}_i - \frac{X_{...}^2}{mnk}$	$\mathrm{MS}_{pi} = \frac{\mathrm{SS}_{pi}}{(n - 1)(k - 1)}$
Residual	$n(n - 1)(k - 1)$	$\mathrm{SS}_r = \sum_p \sum_i \sum_j X_{pij}^2 - \frac{\sum_p \sum_i X_{pi.}^2}{m} - \mathrm{SS}_{jwp}$	$\mathrm{MS}_r = \frac{\mathrm{SS}_r}{n(m - 1)(k - 1)}$
Total	$nmk - 1$		

replaces. Then

$$
\begin{aligned}
\text{est}\,(V_{M_iI} + V_{epiI}) &= \mathrm{MS}_r\,, \\
\hat{V}_{M_{pi}} &= (\mathrm{MS}_{pi} - \mathrm{MS}_r)/m, \\
\text{est}\,(V_{M_I} + V_{M_{pI}}) &= (\mathrm{MS}_{iwp} - \mathrm{MS}_r)/k, \\
\hat{V}_{M_i} &= (\mathrm{MS}_i - \mathrm{MS}_{pi})/mn, \\
\hat{V}_{M_p} &= (\mathrm{MS}_p - \mathrm{MS}_{iwp} - \mathrm{MS}_{pi} + \mathrm{MS}_r)/mk.
\end{aligned} \tag{18}
$$

While separate estimates of V_{Mpi} , V_{Mi} , and V_{Mp} are obtained, the residual term also contains the ij interaction, and the variance due to differences in mean for conditions of facet j cannot be separated from that due to the pj interaction. Thus, unless both V_{M_j} and $V_{M_{ij}}$ are zero, this design gives an unsatisfactory estimate of expected observed variance for decision data matched on both i and j (designs 1 — 3). However, it can be used to obtain the expected observed variance of data gathered by designs 4, 5, 6, 7, 7′, and 8 (cf. Table 5).

Discussion

The concept of generalization underlying our model is one that serves to unify and clarify much of the disparate material on reliability. It brings to focus the fact that a test, a rating scale, or other measuring instrument has many "reliabilities," one for each combination of intended interpretation (universe of generalization) and data-collection design. Each VR indicates the accuracy with which one can make a particular generalization on the basis of decision data obtained under a specified procedure.

One fundamental feature of our model is the assumption that conditions of observation are randomly and independently sampled from a multifacet universe of possible conditions. This assumption does not imply equivalence of scores in the classical sense of equal variability of scores under all conditions and equal intercorrelations, whereas the assumption made by Lindquist [10] that the components of observed score are independent random variables does imply classical equivalence. Removal of the equivalence assumption improves the model as a description of data from ratings and test items where variances are rarely uniform. However, it also makes possible a different coefficient of generalizability to M_p for every set of $k'm'$ conditions in designs 1, 2, 3. Ratings made by one rater, for example, may agree more with the universe score than those made by another rater. We estimate the *expected* coefficient of generalizability for matched data under any set of $k'm'$ conditions that might be drawn from the universe, rather than the specific coefficient for a particular set of conditions. With matched data the expected coefficient is identical to the specific coefficient *only* when conditions are equivalent in the classical sense. The specific coefficient would be preferred, in principle, where one knows what conditions will be used in the matched D study, but

it can almost never be estimated (see [4], pp. 148–149). With data where conditions of observation vary randomly with persons, the variance of observed scores in the population is the same for every set of observations, and no distinction is required between the specific and the expected coefficient of generalizability.

While we have limited this paper to conditions randomly sampled from an infinitely large two-facet universe in which all combinations of pairs of conditions theoretically exist, it is possible to extend our model. The extension to a universe of more than two facets follows directly from the present development. The random-sampling model can be altered to consider stratified sampling on one or more facets. Results for generalization to a finite universe of conditions may be obtained with the aid of the pigeonhole model and expected mean squares presented by Cornfield and Tukey [2]. Using this model, Medley and Mitzel [12] have presented rules for obtaining variance components from a completely crossed G study in which conditions of one facet are assumed fixed. It should be noted that the resulting VRs are algebraically equivalent to generalization to our universe B (or B'), but apply logically only to the conditions of the fixed-facet used in the G study since these conditions exhaust the universe for that facet under their assumptions.

The most important function of a G study is to estimate variance components. This information not only can be used to design the D study but also may answer important substantive questions. Endler, Hunt, and Rosenberg [6], for example, were interested in the extent to which "anxiousness" as a trait is dependent on the situation in which the subject is observed and the modes of response considered in judging anxiety. They constructed an S–R Inventory of Anxiousness in which the subject indicates the intensity of his feeling on 14 five-point response scales such as: "heart beats faster," "get an uneasy feeling," etc., considering eleven situations in turn. Data from this inventory are matched, with situations and response scales crossed (design 1). The investigators originally attempted to answer their questions by a comparison of mean squares, which is incorrect. In a further study (in preparation) they interpret variance components. The estimates for one sample (169 male and female students at Pennsylvania State University) are as follows.

Subject (S)	.103
Situation (Sit)	.094
Mode of response (M-R)	.438
S × Sit	.179
S × M-R	.200
Sit × M-R	.118
Residual	.660

Residual variance, mode of response, and the S × Sit and S × M–R interactions make the largest contributions to variance in the universe of

observations. The large M–R component indicates that any assessment of anxiety depends to a considerable extent on which modes of behavior are considered. Any D study should use the same set of M–R for all persons—a remark that is less obvious than it may seem. Endler *et al.* [7] have pointed out that in ratings and observations of anxiety the observer is very often left to decide for himself what symptoms or cues he will attend to, a procedure which evidently attenuates the generalizability of the ratings. While the smaller components for Sit and Sit × M–R suggest that matching on situations is less vital, this too is desirable in the D study. Some situations produce more anxiety, on the average, than others, and different situations tend to elicit different patterns of anxiety responses.

The variance estimate for S, compared to those for S × Sit and S × M–R, indicates that the general trait of "anxiousness" makes only a limited contribution to score variance; Ss react differentially to situations and utilize different patterns of response. A considerable proportion of the variance arises from the residual, which includes both the triple interaction (S × Sit × M–R) and a component describing variance over observations within cells. Retest data would be required to isolate the interaction and thus determine to what extent each person has a unique response pattern for each situation.

The above example allows us to demonstrate another important conclusion. The Spearman-Brown concept of reliability as a function of "length of test" is not applicable to a multifacet instrument. While an investigator improves generalizability by increasing k' and m', no simple formula suffices to relate VR to the over-all number of observations. Accuracy of generalization is a function of the sample size on each facet. Consider the efficiency of various crossed designs for an S–R Inventory, all of which elicit 144 responses per subject, but for which the number of situations (k') and the number of response modes (m') are varied. The estimated observed variance (assuming that the score on the instrument will be the average response) and the corresponding VR for several possible values of k' and m' are shown in Table 6. Each change in the k', m' balance alters VR even though the "length" of the instrument remains constant. For this example the maximum VR is obtained for k' and m' nearly equal, but in general the optimum balance depends on the relative size of the interaction components.

The analysis thus far, illuminating as it is, has not led us to the correct answer for all purposes. If the investigator wants to generalize only to M_p he would do better to use design 3 for his inventory by pairing each situation with a *different* mode of response. For a 144-item test this would reduce both the S × Sit and S × M–R contributions by a factor of 144, yielding an estimated observed variance of .107 and a VR of .94. With design 3, increasing the number of observations increases both k' and m' equally; hence the VR increases in accord with the Spearman-Brown formula.

The investigator may, however, want to preserve information about S's reaction to particular situations: i.e., he may want to form one subtest

for each situation i and use the observed $\bar{X}_{pi}$ as a basis for generalizing over the universe of responses to the corresponding universe score M_{pi} . Each subtest will have length m', and the number of subtests will be k'. The expected generalizability of each subtest is indicated at the bottom of Table 6. The components for both S and S × Sit now contribute to universe-score variance. As the number of situations is increased, with fixed total test length, the accuracy of generalization to each of the M_{pi} decreases, so that the investigator has to compromise between "bandwidth" and "fidelity." The $k'm'$ pair that produced optimum generalizability over both situations and responses (VR near .73) permits generalization to twelve or so different M_{pi} for each person with a VR of about .80. If the investigator is willing to learn about only four specific situations, the coefficient of generalizability for the specific inferences rises to .92, but that for generalization to M_p drops to .64. Hence an investigator with multiple purposes must make some delicate compromises in designing his D study.

It would be highly desirable for developers of tests to design G studies so that they could report in the test manual the magnitude of each component of variance that one might hypothesize to be important. There has been discussion in the APA *Technical Recommendations* and elsewhere of a distinction between the coefficient of stability and the coefficient of equivalence for a test; reference is also made to a coefficient of internal consistency. From the viewpoint of the present paper, the first describes generalizability over occasions, while the second and third describe generalizability over samples of

TABLE 6

Composition of Observed Variance for an S-R Inventory of 144 Responses Using Differing Numbers of Situations (k') and Responses (m')

Component	Variance Estimate	Multiplier	k':	144	36	12	4	1
			m':	1	4	12	36	144
Subject (p)	.10	1		.100	.100	.100	.100	.100
S × Sit (pi)	.18	$1/k'$		.001	.005	.015	.045	.180
S × M-R (pj)	.20	$1/m'$		.200	.050	.017	.006	.001
Residual	.66	$1/k'm'$		.005	.005	.005	.005	.005
Estimated observed variance of average score on 144 responses				.306	.160	.137	.156	.286
VR: percentage of variance attributable to M_p				.33	.62	.73	.64	.35
Estimated variance of average subtest score ($k' = 1$)				1.140	.495	.352	.304	.286
VR: percentage of variance attributable to M_{pi}				.25	.57	.80	.92	.98

items. No distinction between "equivalence" and "internal consistency" is required when items are experimentally independent and have been randomly sampled from a universe. A suitable G study for these purposes is one in which each person is measured on at least two occasions with several items or at least two test forms. (For many tests, it would be desirable to augment the design by including scorers or examiners as an additional possible source of variance.) This type of analysis not only yields information about the stability of scores over occasions and the equivalence of items but also gives an estimate of the interaction component. Furthermore, the information is in a form which can be used by investigators asking different questions and using different designs to collect D data as well as by those desiring only the two customary coefficients. For these reasons we recommend reporting estimated components of variance in test manuals. If, in addition, some coefficient is reported, the manual should make very clear to what universe of generalization and D-study design the coefficient pertains.

REFERENCES

[1] Burt, C. Test reliability estimated by analysis of variance. *Brit. J. statist. Psychol.*, 1955, **8**, 103–118.

[2] Cornfield, J. and Tukey, J. W. Average values of mean squares in factorials. *Ann. math. Statist.*, 1956, **27**, 907–949.

[3] Cronbach, L. J. Test reliability: its meaning and determination. *Psychometrika*, 1947, **12**, 1–16.

[4] Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. Theory of generalizability: a liberalization of reliability theory. *Brit. J. statist. Psychol.*, 1963, **16**, 137–163.

[5] Cronbach, L. J., Gleser, G. C., and Rajaratnam, N. Generalizability of stratified-parallel tests. *Psychometrika*, 1965, **30**, 39–56.

[6] Endler, N. S., Hunt, J. M., and Rosenstein, A. J. An S-R inventory of anxiousness. *Psychol. Monogr.*, 1962, No. 536.

[7] Endler, N. S. and Hunt, J. M. Sources of behavioral variance as measured by the S-R inventory of anxiousness. Mimeographed (December, 1964).

[8] Gulliksen, H. The content reliability of a test. *Psychometrika*, 1936, **1**, 189–194.

[9] Guttman, L. The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. Glencoe, Ill.: Free Press, 1954.

[10] Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

[11] Mahmoud, A. F. Test reliability in terms of factor theory. *Brit. J. statist. Psychol.*, 1955, **8**, 119–135.

[12] Medley, D. M. and Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963. Pp. 247–328.

[13] Moonan, W. J. Computational illustrations of the internal and external consistency analysis of examination responses. *J. exp. Educ.* 1957, **25**, 181–190.

[14] Stanley, J. C. Analysis of three-way orthogonal classifications, with applications to rater bias and trait independence. Unpublished manuscript, 1959.

[15] Thorndike, R. L. Research problems and techniques. *Report No. 3*, AAF Aviation Psychol. Program Research Reports. U. S. Government Printing Office, 1947.

Manuscript received 2/22/65