The Symmetry of Generalizability Theory: Applications to Educational Measurement
Author(s): Jean Cardinet, Yvan Tourneur and Linda Allal
Reviewed work(s):
Source: Journal of Educational Measurement, Vol. 13, No. 2 (Summer, 1976), pp. 119-135
Published by: National Council on Measurement in Education
Stable URL: http://www.jstor.org/stable/1434233
Accessed: 15/03/2013 13:30

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.
Generalizability theory, as developed by Cronbach, Gleser, Nanda and Rajaratnam (1972), offers a more comprehensive and coherent framework than classical psychometric theory for the study of educational and psychological measurement. Nevertheless, these authors retain, at least in the examples they present, the traditional preoccupation of psychometrics, i.e., achieving the best possible differentiation of the persons tested.

We would like to show that such a limitation is unnecessary and that generalizability theory provides a powerful descriptive and analytic tool for other kinds of problems, where persons are not the central object of study. Curriculum evaluation, in particular, implies differentiation of educational objectives, of learning situations, of stages of progress, etc. In these cases, the between-subjects variability is more detrimental than helpful to the clarity of the results.

When research focuses on the conditions of measurement rather than on persons measured, it becomes necessary to transpose the dimensions of the measurement design so as to differentiate conditions while generalizing over persons. In order to clarify the way in which the dimensions of a design need to be treated, depending on the measurement problem under study, some new concepts must be introduced, in particular, the notions of face of differentiation and face of generalization as complementary aspects of a measurement design.

After introducing these concepts, an example will be presented, in order to show how the tools of generalizability theory may be extended to deal with a wide variety of measurement questions. Although implicit in the model of Cronbach, these extensions are not customary and can profitably be presented as suggestions for further research.

THE ORIGINS OF GENERALIZABILITY THEORY:
THE INTRACLASS COEFFICIENT

In a 1941 paper, Cyril Hoyt presented a procedure for estimating test reliability by analysis of variance, and illustrated the method using a specific example. We shall use the same example to establish a link between the ideas of Cronbach and those of his predecessors. Subsequently, we shall examine another example, with more complex data, to illustrate how generalizability theory can extend the same kind of analysis to multiple sources of variation.
Hoyt's Example

A test of 250 items was administered to 33 students in a class in botany in the College of Pharmacy of the University of Minnesota. The analysis of variance table is reproduced in Table 1:

Table 1
Analysis of variance for Hoyt's example

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among individuals</td>
<td>82.82933</td>
<td>32</td>
<td>2.58842</td>
</tr>
<tr>
<td>Among items</td>
<td>593.81927</td>
<td>249</td>
<td>2.38482</td>
</tr>
<tr>
<td>Individuals x items</td>
<td>1325.77673</td>
<td>7968</td>
<td>.16639</td>
</tr>
<tr>
<td>Total</td>
<td>2002.42533</td>
<td>8249</td>
<td>.24275</td>
</tr>
</tbody>
</table>

Calculation of the Variance Components

Computation of an F test makes use of the following well-known relations:
—The mean square interaction is an estimate of the pooled interaction and error variance: \( \sigma_{i,e}^2 \)
—The mean square among individuals is an estimate of \( \sigma_{i,e}^2 + n_e \sigma_i^2 \)
—The mean square among items is an estimate of \( \sigma_{i,e}^2 + n_i \sigma_i^2 \) where \( \sigma_i^2 \) is the true (i.e., systematic) variance between the \( n_r \) row means, \( \sigma_i^2 \) is the true variance between the \( n_c \) column means, \( n_r \) is the number of subjects, and \( n_c \) is the number of items.

The equations for estimating each variance component can be easily derived from these relations. For Hoyt's example, \( \hat{\sigma}_{i,e}^2 = .16639 \), \( \hat{\sigma}_i^2 = .00968812 \), and \( \hat{\sigma}_e^2 = .06722515 \).

Calculation of the Intraclass Coefficients

To obtain a coefficient of reliability (defined in classical test theory as the ratio of true variance to observed variance), an intraclass coefficient may be computed:

\[ r_g = \frac{\nu_g}{(\nu_g + \nu_e)} \]

where \( \nu_g \) is the true-variance component associated with factor G, and \( \nu_e \) the appropriate error-variance component. Such a coefficient can be interpreted as measuring the degree of differentiation between the levels of the factor being measured. Cronbach uses the term rho square (\( \rho^2 \)) to designate the intraclass coefficient within
the context of generalizability theory. We shall adopt the same symbol in the remainder of this text.

If the differentiation is made on the basis of a mean score rather than a single score, the error variance is reduced: it is divided by the number of observations on which the mean is based.

Let us compute first the intraclass coefficient between subjects:

$$\rho_s^2 = \frac{\sigma_s^2}{\sigma_s^2 + (\sigma_{n,s}^2/n_s)} = \frac{0.0968812/0.0968812 + (0.16639/250)}{0.9357}$$

This value checks with Hoyt's result, obtained by another formula. Hoyt indicated that it would be possible to examine the between-items variance in order to determine the heterogeneity of the item difficulties. This is precisely the direction of analysis which we wish to develop. One may start by computing the intraclass coefficient between items.

$$\rho_i^2 = \frac{\sigma_i^2}{\sigma_i^2 + (\sigma_{n,i}^2/n_i)} = \frac{0.06722515/0.06722515 + (0.16639/33)}{0.9302}$$

The values of $\rho_s^2$ and $\rho_i^2$ indicate that the differentiation is completely satisfactory, both for subjects and for items, with the dimensions used in the design.

Directions of Development

Hoyt's formula, which did not make explicit the components of variance, was difficult to generalize. On the other hand, the rationale used by Cronbach and mentioned above for the computation of the generalizability coefficient lends itself quite easily to the introduction of other sources of variance. A second example will make this clear; it will show at the same time how Hoyt's proposal to examine the between-items variance can be realized.

To make the second example more meaningful, however, it is first necessary to summarize Cronbach's systematization, and to describe our own theoretical contribution to this model. We shall then be in a better position to discuss the various possibilities of applying generalizability theory to educational measures.

THE BASIC CONCEPTS

What Does Generalizability Mean?

A measurement instrument (test, rating scale, etc.) is useful only if the score it yields can give us information about something else. An observed score must at least inform us as to the expected value of other measures taken under equivalent conditions. This minimum requirement has traditionally been called "reliability."

The definition of an "equivalent" condition, however, is the heart of the problem. Using another set of questions, or repeating the same measure on another occasion, as was done traditionally, inevitably introduces sources of systematic variation. Classical test theory preferred to ignore them: each condition was supposed to estimate the same true score. Generalizability theory admits, on the contrary, that each observation belongs to a multitude of possible sets of observations. When conditions of a measurement situation can be maintained equivalent, the variability between one result and the next is likely to be limited. When measurement conditions are allowed to vary in one or several respects, the results are likely to be modified by the intervention of the corresponding sources of variation.
Cronbach concludes that a test is not reliable or unreliable; one can simply
generalize, to different degrees, from one observed score to the multiple means of the
different sets of possible observations. It follows that there are as many generalizability
coefficients as sets of observations.

What is a Facet in Cronbach's Sense?

Cronbach and his collaborators have introduced the concept of facet to designate an
aspect of the measurement situation which is allowed to vary from one observation to
the next. The set of moments (times) at which observations are made constitutes a
facet. The group of persons entitled to administer the instrument to the subjects may be
another. A facet is thus a set of conditions under which measurements can be carried
out. All conditions of a facet must be of the same kind, e.g., occasions, scorers, items,
etc..

In the presentation of generalizability theory by Cronbach and his collaborators, a
distinction is made between the dimension "persons" and the other dimensions of the
design (the characteristics of the situation that affect the measurements), to which the
term facet is applied. This follows from these authors’ implicit assumption that persons
are always the object of measurement.

As a consequence, in their book, the subjects are not counted as a facet: what they
call a two-facet design is the equivalent of a three-dimensional factorial design in
ANOVA (i.e., a design in which subjects are crossed with two other factors).

Still Cronbach and his collaborators do give one example (Cronbach et al., p. 203)
where schools, rather than the pupils themselves, are the object of study. This is a first
indication of the existence of other types of problems, where attention should not be
restricted to the differentiation of the people measured.

What is a Facet in our Terminology?

The introduction of asymmetry into a design in which no one dimension needs to be
given preferential treatment at the outset, seems to us to be deplorable. Our contribu-
tion is precisely to reestablish the symmetry of the design, by making the subjects a
facet like the others. By so doing, it becomes possible to extend generalizability theory
to a wider range of measurement problems than those explicitly dealt with in Cron-
bach's presentation. It becomes possible, in this perspective, to determine how precisely
a given measurement design differentiates its objects of study, whether persons, ques-
tions, methods, school systems, stages of learning, etc..

The practical importance of this distinction is easy to illustrate. If persons are the ob-
ject of study, maximum reliability will be obtained if the variance between persons is
large and if the items are all of medium difficulty. If, on the contrary, the object of study
is the degree to which different educational objectives have been attained within a given
system, maximum reliability will be obtained if the objectives vary widely in difficulty
(and, consequently, the between-items variance is large), while the variability between
subjects is as small as possible.

One example is enough to reject the usual assumption of psychometrics that it is al-
ways advisable to maximize the variance between subjects, while minimizing the
variance between conditions of observation.

At this point, it becomes necessary to distinguish between the two concepts that
remained confounded in the examples Cronbach chose: the population of subjects
measured and the objects of study needing maximum differentiation. A special terminology is needed to formulate this distinction.

**The Face of Generalization**

We shall call facet of generalization any source of variation which affects the measures taken of the objects under study. The facets of generalization are those aspects of the situation which create random errors of measurement, but also variations that are, in a sense, "systematic errors" of measurement.

Generalizability theory recognizes that any observation is subject to the influence of certain sources of variation that cannot be controlled. One can only establish a confidence interval to delimit the probable influence of all these factors combined. The facets of generalization are those which contribute to increase the confidence interval around the mean of all measures taken of the particular object under study. Common examples of facets of generalization would include moments (times) of test administration, scorers, topics of essay writing, etc..

The cartesian product of all the facets of generalization constitutes the set of all possible conditions of observation considered in the study. We call it the face of generalization. By convention, we shall always write the data matrix in such a way that its columns correspond to the face of generalization.

**The Face of Differentiation**

The fact that subjects constitute a facet of differentiation in traditional psychometrics is easy to understand. Since test scores represent little more than an ordinal scale, it is important to be able to differentiate the relative positions of subjects on the scale. A large range of variation between subjects is a condition favourable to the stability of their ranking, which is the basis for interpreting the scores derived from the test.

When subjects constitute a source of sampling error, as is the case with opinion polls, with surveys of educational attainment, with laboratory experiments, with curriculum evaluation studies and with many other fields of educational research, variability between persons is no longer desired. Other classes of objects must then be differentiated: the issues proposed, the systems compared, the treatments examined, etc..

Any study which is in some way comparative requires discrimination of the objects to be compared. A set of objects or characteristics which are to be compared constitutes a facet of differentiation in a study.

There may be more than one facet of differentiation in a given study if comparisons are to be made separately on several dimensions. In a laboratory experiment where the effects of several factors are tested by means of a factorial design, each hypothetical factor becomes a facet of differentiation. Replications (or higher order interactions considered as quasi-replications) represent the corresponding facets of generalization.

The cartesian product of all the facets of differentiation constitutes the set on which all possible comparisons between the objects of study are defined. We call it the face of differentiation. By convention, we shall always write the data matrix in such a way that its rows correspond to the face of differentiation.

**The Fixed and Random Conditions**

Cronbach’s identification of the conditions of observation of people with the face of generalization may cause other confusions to some superficial readers: we are referring to the distinction he makes between fixed and random conditions of a facet.
Conditions are random if the conditions used for a given facet are chosen randomly, or are considered, theoretically, to constitute a random sample. Testing occasions, scorers, test forms and items are often treated as random facets of a measurement design. Conditions are fixed for a facet whenever the conditions present in the design constitute the entire set in which we are interested. In this case, there is no need to take into account the sampling fluctuations of the conditions of the facet.

If the concepts of facets of generalization and differentiation are not introduced, one may be tempted to speak of generalization only for facets with random conditions. It is true that when conditions are fixed, there is no larger set of conditions to which we could generalize the observations we have obtained. Nevertheless, it is possible to generalize to the mean of fixed conditions as well as to the expected mean of randomly chosen conditions.

\[
\text{MOMENTS}
\]

\[
1 \quad 2
\]

\[
1 \quad . \quad S
\quad . \quad U
\quad . \quad B
\quad . \quad J
\quad . \quad E
\quad . \quad C
\quad . \quad T
\quad . \quad S
\]

\[
100
\]

\[
\text{Author} \quad \text{Author} \quad \text{Author}
\]

\[
A \quad B \quad C
\]

\[
\text{OBJECTS}
\]

\[
n_s = 100
\]

\[
n_o = 3
\]

\[
n_m = 2
\]

Figure 1. Design for the repeated application of a test of closure with texts by three authors.
With the distinction that we have introduced between the two types of facets, there is less danger of misinterpretation, for it should be clear that random and fixed conditions may exist on both faces. Fixing or not fixing the conditions of the facets of generalization has important consequences on the value of the generalizability coefficients. We have not found similar consequences for defining one way or the other the conditions of the facets of differentiation.

We may now apply the concepts presented above to the treatment of a hypothetical example, chosen to illustrate the diversity of the problems that may be attacked with the tools of generalizability theory.

POSSIBLE APPLICATIONS

Presentation of an Example

Let us suppose that a cloze test has been constructed as a measure of reading comprehension. Three texts of the same length, written by different authors, have been chosen from a textbook in English Literature. Every seventh word has been omitted and the task is to fill in the exact word that is missing.

We may imagine that the test has been administered at two moments in time during a course of study dealing with the understanding of literature.

The data obtained can then be organized as a factorial design with three factors: (1) Subjects ($n_s = 100$), (2) Objects, i.e., authors ($n_o = 3$), and (3) Moments ($n_m = 2$), as shown in Figure 1.

Let us suppose that we arrive at the following breakdown of the total sum of squares (Table 2):

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects (S)</td>
<td>8201</td>
<td>99</td>
<td>83</td>
</tr>
<tr>
<td>Authors (O)</td>
<td>3216</td>
<td>2</td>
<td>1608</td>
</tr>
<tr>
<td>Moments (M)</td>
<td>1917</td>
<td>1</td>
<td>1917</td>
</tr>
<tr>
<td>S x O</td>
<td>1584</td>
<td>198</td>
<td>8</td>
</tr>
<tr>
<td>S x M</td>
<td>1683</td>
<td>99</td>
<td>17</td>
</tr>
<tr>
<td>O x M</td>
<td>804</td>
<td>2</td>
<td>402</td>
</tr>
<tr>
<td>S x O x M</td>
<td>396</td>
<td>198</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>17801</td>
<td>599</td>
<td></td>
</tr>
</tbody>
</table>

Table 2

Analysis of variance for the factorial design of Figure 1
Variance components can be computed for each source of variation. If we consider that subjects, authors and moments have been chosen at random, the factorial design satisfies the random model assumptions.

The mean squares and variance components of this design can be represented by a Venn diagram, as shown in Figure 2. Here, each circle or intersection of circles represents a mean square. Each mean square is the sum of certain variance components, multiplied by coefficients. The subscript of each variance component indicates the factor (or factors) to which it corresponds. The coefficient of each variance component corresponds to the factors which do not appear in the component's subscript. The equations for computing the variance components can be easily derived from this Venn diagram (see Table 3).

Measurement Questions To Be Treated

If we apply to these data the concepts presented above, a number of questions can now be examined.
Table 3

Equations for computing the variance components
and application to the example

\[
\hat{\sigma}^2_{som,e} = \frac{MS_{som}}{\text{som}} = 2
\]

\[
\hat{\sigma}^2_{om} = \frac{(MS_{om} - MS_{som})}{n_s} = 4
\]

\[
\hat{\sigma}^2_{sm} = \frac{(MS_{sm} - MS_{som})}{n_o} = 5
\]

\[
\hat{\sigma}^2_{so} = \frac{(MS_{so} - MS_{som})}{n_m} = 3
\]

\[
\hat{\sigma}^2_{m} = \frac{(MS_{m} - MS_{sm} - MS_{om} + MS_{som})}{n_s n_o} = 5
\]

\[
\hat{\sigma}^2_{o} = \frac{(MS_{o} - MS_{so} - MS_{om} + MS_{som})}{n_s n_m} = 6
\]

\[
\hat{\sigma}^2_{s} = \frac{(MS_{s} - MS_{so} - MS_{sm} + MS_{som})}{n_o n_m} = 10
\]

The face of differentiation could be composed of the subjects, or of the authors, or of the moments. It could also be formed by the combination of the facets Subjects and Authors, Subjects and Moments, or Authors and Moments.

The face of generalization is, of course, the complement of the face of differentiation. It will be composed of two facets when differentiation is desired on only one facet, and of only one facet in the other case. On this face, it is essential to distinguish facets having fixed and random conditions.

If the face of generalization were composed only of fixed facets, the coefficient of generalizability would then be equal to unity, as we would know everything we might wish to know. As this case is trivial, we shall consider only those cases where at least one facet has random conditions.

Altogether we obtain 12 possible cases of differentiation. They are presented in Table 4, with a brief explanation of the empirical meaning of each of the corresponding measurement designs. A set of data as simple as the one in our example is thus capable of providing information on a whole series of questions. This diversity might have passed unnoticed without the conceptual tools of generalizability theory.
Table 4
List of all possible differentiation problems
for a three-dimensional design

<table>
<thead>
<tr>
<th>No</th>
<th>Different.</th>
<th>Generaliz.</th>
<th>Rand.</th>
<th>Fix.</th>
<th>The problem is to obtain measures differentiating:</th>
<th>whatever may be the:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>O</td>
<td>M</td>
<td></td>
<td>the students' comprehension of texts at two moments during the course</td>
<td>authors</td>
</tr>
<tr>
<td>2</td>
<td>S</td>
<td>M</td>
<td>O</td>
<td></td>
<td>the students' comprehension of texts written only by the three authors</td>
<td>moment</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>O &amp; M</td>
<td></td>
<td></td>
<td>the student's comprehension of literary texts</td>
<td>authors and the moment</td>
</tr>
<tr>
<td>4</td>
<td>O</td>
<td>S</td>
<td>M</td>
<td></td>
<td>the difficulty of the three authors, at two moments of the course</td>
<td>students</td>
</tr>
<tr>
<td>5</td>
<td>O</td>
<td>M</td>
<td>S</td>
<td></td>
<td>the difficulty of the three authors, for a given group of students</td>
<td>moment</td>
</tr>
<tr>
<td>6</td>
<td>O</td>
<td>S &amp; M</td>
<td></td>
<td></td>
<td>the difficulty of the three authors</td>
<td>students and moment</td>
</tr>
<tr>
<td>7</td>
<td>M</td>
<td>S</td>
<td>O</td>
<td></td>
<td>the effect of a course dealing with three given authors</td>
<td>students</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>O</td>
<td>S</td>
<td></td>
<td>the effect of a Literature course on a given group of students</td>
<td>authors</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>S &amp; O</td>
<td></td>
<td></td>
<td>the effect of a course dealing with the understanding of literature</td>
<td>students and the authors</td>
</tr>
<tr>
<td>10</td>
<td>S &amp; O</td>
<td>M</td>
<td></td>
<td></td>
<td>the comprehension of each author, by a given group of students</td>
<td>moment</td>
</tr>
<tr>
<td>11</td>
<td>S &amp; M</td>
<td>O</td>
<td></td>
<td></td>
<td>the students' comprehension of texts, at each of two moments during the course</td>
<td>authors</td>
</tr>
<tr>
<td>12</td>
<td>O &amp; M</td>
<td>S</td>
<td></td>
<td></td>
<td>the difficulty of three given authors</td>
<td>students</td>
</tr>
</tbody>
</table>

128
As shown in Table 4, three main fields of enquiry may be considered, depending on the type of differentiation needed: the ability of the students, the difficulty of the texts and the effect of the course. Moreover, the interpretation of the scores obtained will, in each case, depend on the kind of generalization required. We may fix one facet of generalization and describe the results in the particular context of the levels observed for the fixed facet. We may want, on the contrary, to consider the measures obtained as representative of all possible conditions of the facet of generalization, even those that were not observed. Logically, the precision of the measure should be higher in the first case than in the second. We can now verify this mathematically by computing generalizability coefficients for each of these situations.

**Coefficients of Generalizability**

A generalizability coefficient is an intraclass coefficient which extends the classical definition of reliability (defined as the ratio of true variance to observed variance).

The numerator, estimating the true variance, contains the components of variance corresponding to the face of differentiation. (If conditions are fixed for some facets of generalization, the numerator contains also the interaction of these facets with the face of differentiation.)

The denominator contains the expected variance of the observed scores that pertain to the face of differentiation for the chosen design. This variance includes, besides the true variance, the interaction of the faces of differentiation and generalization. (If a facet of generalization is nested in a facet of differentiation, it also adds its own component to the variance of the observed scores.)

We speak of an “expected” variance because the procedure is divided into two stages. In a G-study (G for Generalizability), the different variance components which affect the measures are estimated on the basis of a complete factorial design. In a D-study (D for Decision) which follows, a modified version of the G-study design is chosen: one estimates the variance of the scores for this projected design (without really observing them), by adding up the variance components implied. The precision of the new design can then be tested by the computation of a generalizability coefficient.

The coefficient $p^2$ has two main functions. According to Cronbach’s presentation, it is especially useful to determine, before conducting a D-study, the adaptations of the measurement design that are needed to attain a desired level of generalizability within the imposed limitations of cost, time, effort, etc. For example, one may come to the conclusion that the number of conditions making up a certain facet will have to be increased.

The second use that we suggest is to determine the adequacy of a given design for various measurement purposes. As an illustration of this process, Table 5 gives the formulas appropriate to the different combinations of facets presented in Table 4, and applies them to our hypothetical example.

If we examine the values found for $p^2$, it would appear that the design is insufficient to measure adequately a student’s comprehension of texts, independently of the specifically chosen authors and of the various amounts of study (Formula 3: $p^2 = .720$). On the other hand, if we limit our conclusions to these three authors only, the generalizability of the total of the three subscores, whatever the moments, is near the usually admitted level of .80 (Formula 2). If we interchange O and M, fixing the moments by measuring the students always at the same two moments in their training, but admitting other kinds of texts, we obtain a coefficient of .904 (Formula 1). This shows that the design...
Table 5
Formulas for computing the generalizability coefficient
for the three-dimensional designs in Table 4,
and results for the case of the example

<table>
<thead>
<tr>
<th>$N^0$ of design</th>
<th>Diffe-</th>
<th>Generaliz.</th>
<th>Formula of $\rho^2$</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rent.</td>
<td>Rand. Fix.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>S O M</td>
<td></td>
<td>$\sigma_s^2 + (1/n_m) \sigma_{sm}^2$</td>
<td>0.904</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_s^2 + (1/n_m) \sigma_{sm}^2 + (1/n_o) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>S M O</td>
<td></td>
<td>$\sigma_s^2 + (1/n_o) \sigma_{so}^2$</td>
<td>0.792</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_s^2 + (1/n_m) \sigma_{sm}^2 + (1/n_o) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>S O &amp; M</td>
<td></td>
<td>$\sigma_s^2$</td>
<td>0.720</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_s^2 + (1/n_m) \sigma_{sm}^2 + (1/n_o) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>O S M</td>
<td></td>
<td>$\sigma_o^2 + (1/n_m) \sigma_{om}^2$</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_o^2 + (1/n_m) \sigma_{om}^2 + (1/n_s) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>O M S</td>
<td></td>
<td>$\sigma_o^2 + (1/n_s) \sigma_{so}^2$</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_o^2 + (1/n_m) \sigma_{om}^2 + (1/n_s) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>O S &amp; M</td>
<td></td>
<td>$\sigma_o^2$</td>
<td>0.746</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\sigma_o^2 + (1/n_m) \sigma_{om}^2 + (1/n_s) \sigma_{so}^2 + (1/n_o m) \sigma_{som,e}^2$</td>
<td></td>
</tr>
</tbody>
</table>

would be quite adequate to test verbal comprehension if the moments of observation were fixed.

The comparison of the Formulas 4, 5 and 6, and of the values of the corresponding coefficients, shows that it is not possible to reliably differentiate the levels of difficulty of the authors without fixing the moments of observation. (Another possibility would be, of course, to increase the sampling of moments: $\rho^2$ would reach .81 if there were three occasions of observation during the same period, and .85 if there were four.)
Table 5 (continued)
Formulas for computing the generalizability coefficient
for the three-dimensional designs in Table 4,
and results for the case of the example

<table>
<thead>
<tr>
<th>N° of design</th>
<th>Different.</th>
<th>Generaliz.</th>
<th>Formula of $\rho^2$</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>M</td>
<td>S</td>
<td>$\frac{\sigma^2_m + (1/n_o) \sigma^2_{om}}{\sigma^2_m + (1/n_o) \sigma^2_{om} + (1/n_s) \sigma^2_{sm} + (1/n_n) \sigma^2_{som,e}}$</td>
<td>.992</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>O</td>
<td>$\frac{\sigma^2_m}{\sigma^2_m + (1/n_o) \sigma^2_{om} + (1/n_s) \sigma^2_{sm} + (1/n_n) \sigma^2_{som,e}}$</td>
<td>.791</td>
</tr>
<tr>
<td>9</td>
<td>M</td>
<td>S &amp; O</td>
<td>$\frac{\sigma^2_m}{\sigma^2_m + (1/n_o) \sigma^2_{om} + (1/n_s) \sigma^2_{sm} + (1/n_n) \sigma^2_{som,e}}$</td>
<td>.783</td>
</tr>
<tr>
<td>10</td>
<td>S &amp; O</td>
<td>M</td>
<td>$\frac{\sigma^2_s + \sigma^2_o + \sigma^2_{so}}{\sigma^2_s + \sigma^2_o + \sigma^2_{so} + 1/n_m (\sigma^2_{sm} + \sigma^2_{om} + \sigma^2_{som,e})}$</td>
<td>.776</td>
</tr>
<tr>
<td>11</td>
<td>S &amp; O</td>
<td>O</td>
<td>$\frac{\sigma^2_s + \sigma^2_m + \sigma^2_{sm}}{\sigma^2_s + \sigma^2_m + \sigma^2_{sm} + 1/n_o (\sigma^2_{so} + \sigma^2_{mo} + \sigma^2_{som,e})}$</td>
<td>.870</td>
</tr>
<tr>
<td>12</td>
<td>O &amp; M</td>
<td>S</td>
<td>$\frac{\sigma^2_o + \sigma^2_m + \sigma^2_{om}}{\sigma^2_o + \sigma^2_m + \sigma^2_{om} + 1/n_s (\sigma^2_{so} + \sigma^2_{sm} + \sigma^2_{som,e})}$</td>
<td>.993</td>
</tr>
</tbody>
</table>

Examining the following coefficient (Formula 7: $\rho^2 = .992$), we discover that it is perfectly possible to measure reliably the effect of the course. A much smaller number of students would certainly suffice. However, the conclusions have to be limited to the three authors chosen. Generalization to other authors (Formulas 8 and 9) leads to insufficient differentiation ($\rho^2 = .791$ and .783).

Row 10 of Table 5 shows that we cannot interpret the three subscores of each student irrespective of his degree of coursework ($\rho^2 = .776$). It is possible, however, to measure
reliably his ability at given moments during his training, irrespective of the texts introduced in the cloze test ($\rho^2 = .870$). The last row of the table shows that the difficulty of the authors at specific moments is perfectly measurable with 100 students; much fewer would certainly suffice.

**Confidence Intervals**

The term "true score," in a classical test theory, was ambiguous since the diverse observation procedures developed by psychometricians yielded non-convergent estimates of a person's true score. In Cronbach's theory, this expression is replaced by "universe score," much more precise, since it designates the expected value of a subject's score, i.e., it is the mean of all possible observations that could be made in a given universe.

This use of sampling theory makes it possible to apply the classical technique of confidence intervals. One can determine a margin of variation within which the universe parameter will be found a given proportion $(1 - \alpha)$ of the time.

The variation of a person's observed score with respect to its expected value can be estimated. In our three-dimensional example, the error variance for a subject's score is given by the formula:

$$
\hat{\sigma}_{\text{OM}}^2 = \frac{1}{n_o} \hat{\sigma}_o^2 + \frac{1}{n_m} \hat{\sigma}_m^2 + \frac{1}{n_o n_m} \hat{\sigma}_{om}^2 + \frac{1}{n_o} \hat{\sigma}_{os}^2 + \frac{1}{n_m} \hat{\sigma}_{sm}^2 + \frac{1}{n_o n_m} \hat{\sigma}_{som,e}^2
$$

Assuming that the deviations from the mean are distributed normally, it is possible to conclude that there is a probability of $(1 - \alpha)$ that the universe score lies within the interval

$$
\bar{X}_{sOM} \pm z_{\alpha/2} \hat{\sigma}_{\text{OM}}
$$

where $\bar{X}_{sOM}$ is the mean of the subject $s$ across randomly chosen levels of $O$ and $M$ forming the matrix $O \times M$.

One can also (by symmetry) determine the confidence interval around the mean score of a cloze test composed of $n_o$ subtests applied to $n_s$ subjects on a given occasion (moment), the levels of the facets $O$ and $S$ being chosen randomly.

$$
\hat{\sigma}_{\text{MSO}}^2 = \frac{1}{n_s} \hat{\sigma}_s^2 + \frac{1}{n_o} \hat{\sigma}_o^2 + \frac{1}{n_s n_o} \hat{\sigma}_{so}^2 + \frac{1}{n_s} \hat{\sigma}_{sm}^2 + \frac{1}{n_o} \hat{\sigma}_{om}^2 + \frac{1}{n_s n_o n_m} \hat{\sigma}_{som,e}^2
$$

A problem of practical importance in preparing criterion-referenced tests is to determine the confidence interval for the mean level of success attained on a series of subtests (or items) belonging to the universe defined by a given objective. Sub-tests having levels of success which fall outside these limits probably do not belong to the universe. In terms of our example, the problem would be to determine whether the difficulty of one author would fall outside the normal range of variation. The error variance (around the expected value of the scores for all randomly chosen subjects and moments) would be:

$$
\hat{\sigma}_{\text{SMO}}^2 = \frac{1}{n_s} \hat{\sigma}_s^2 + \frac{1}{n_m} \hat{\sigma}_m^2 + \frac{1}{n_s n_m} \hat{\sigma}_{sm}^2 + \frac{1}{n_s} \hat{\sigma}_{so}^2 + \frac{1}{n_m} \hat{\sigma}_{om}^2 + \frac{1}{n_s n_m} \hat{\sigma}_{som,e}^2
$$

This formula could be used to estimate the probable range of random fluctuations affecting the true difficulty of one author.
This formula does not, however, take test items into account. In the design of our example, they constitute a hidden facet. As this may be misleading, we are led to introduce a fourth facet, Q, i.e., items or questions. This facet is nested within authors. Its conditions are random, since the procedure used to create the cloze test actually draws the items at random.

The difficulty of an author fluctuates due to random sampling of the levels of the generalization facets, S, M, and Q. Random variations affect the main effects of these factors, their interactions between themselves, and their interactions with O, the differentiation facet. The preceding formula becomes:

$$\hat{\sigma}_{soQM}^2 = \frac{1}{n_s} \sigma^2_s + \frac{1}{n_q} \sigma^2_q + \frac{1}{n_m} \sigma^2_m + \frac{1}{n_q n_m} \sigma^2_{qm} + \frac{1}{n_s n_q} \sigma^2_{sq} + \frac{1}{n_s n_m} \sigma^2_{sm}
+ \frac{1}{n_s} \sigma^2_{so} + \frac{1}{n_m} \sigma^2_{om} + \frac{1}{n_q n_m} \sigma^2_{om} + \frac{1}{n_s n_q n_m} \sigma^2_{s Q M,e}.$$  

The same principle can be applied to the case of an individual subject. For instance, in an individualized type of instruction, a student may be required to attain a minimum score on the cloze test concerning one author, before being allowed to study the others. A confidence interval can be determined around this minimum passing score. Setting $n_s = 1$, $\sigma^2_s$, and $\sigma^2_q = 0$ in the above formula, we get the error variance for establishing this confidence interval: $\hat{\sigma}_{soQM}^2$, for any given level of S and O.

It is also possible to define confidence intervals for the difference between the results of two subjects, or for the difference between the levels of difficulty of two authors. In fact, any comparative decision, such as the determination of the objectives best attained by an educational program, implies such confidence intervals.

The Different Theories of Educational Tests

The examples given above suggest that generalizability theory can be applied to a great variety of measurement problems. We would like now to describe briefly three frequent problems of educational measurement, and show that, in each case, a particular theoretical framework can be derived from this theory. It would seem that other problems, for which a precise methodology does not yet exist, could be attacked by making use of the symmetry of the model.

Theory of norm-referenced tests: differentiation of subjects. This first case corresponds to the classical theory of aptitude tests, in which the facet of differentiation is the subjects, and the facet of generalization, the items or the moments (i.e., reliability estimated by parallel-form or by test-retest procedures). Educational norm-referenced tests apply this methodology directly.

Theory of educational survey instruments: differentiation of objectives. Educational research is frequently interested in measuring the results of a teaching program, i.e., determining the objectives which can or cannot be achieved by the pupils. Therefore, it is the degree to which the objectives are attained which must be measured with precision. The differences between pupils, far from being necessary, are, at best, unimportant.

The inversion of the usual direction of differentiation is easily acceptable if it is realized that the object of the study is the school system. In this case, the test should in-
clude a wide enough range of difficulty to permit the evaluation of all objectives, those clearly attained and those not reached at all. It should differentiate as much as possible the levels of performance attained on tasks pertaining to different objectives.

Theory of progress tests: differentiation of occasions. Other educational tests aim at describing the progress of students through their learning program, distinguishing several stages in the acquisition of the final objective.

By following the principles exposed above, it appears clearly that, if the subjects and the objectives constitute facets of generalization, the only admissible source of differentiation is the third facet: the occasions.

Yet, most tests constructed with this intent up to now have aimed at differentiating the subjects to the maximum degree, thus mixing tests measuring progress and tests measuring an aptitude. These two kinds of measurement could be quite different.

INTERPRETATION IN TERMS OF SYSTEMS ANALYSIS

In summary, it appears that generalizability theory, as presented by Cronbach and his collaborators, can be extended further than the examples they present might lead one to believe.

Generalizability theory offers a comprehensive frame of reference, consistent with the global outlook introduced in education by the ideas of curriculum and system. For instance, considering the educational system as a set of elements in interaction calls our attention to the relationship that exists between the results of the pupils and the methods of teaching, the conditions of examination, etc.. Similarly, generalizability theory may offer the concepts and methods needed for an appropriate assessment of the various dimensions of the educational system: institutional and personal factors, objectives, moments, etc..

Of all the measurement rationales presented in the social sciences, Cronbach’s proposal seems to be the most adequate for embracing the large number of factors that intervene in a social system. Generalizability theory makes it possible to analyze and compare the efficiency of different possible designs for gathering information. It can also usefully guide further efforts to construct new measuring instruments. Furthermore, it suggests new ways of adapting measurement procedures to the particular combination of facets of differentiation and generalization characterizing each situation.

REFERENCES


AUTHORS


TOURNEUR, YVAN Address: Faculté des Sciences Psychopédagogiques, 21 Place du Parc, 7000 Mons, Belgium. Title: Assistant Professor. Degrees: Lic. Pédag., Lic. Psychol. Bruxelles,