DOCUMENT RESUME

ED 142 577 TM 006 399

AUTHOR Naccarato, Richard W.; Gillmore, Gerald M.
TITLE The Application of Generalizability Theory to a College-Level French Placement Exam.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
REPORT NO EAC-76-24
PUB DATE Sep 76
NOTE 17p.

EDRS PRICE MF-$C.83 HC-$1.67 Plus Postage.
DESCRIPTORS *College Placement; College Students; *French; Higher Education; Measurement Techniques; *Statistical Analysis; *Test Reliability; *Tests
IDENTIFIERS *Generalizability Theory; Interrater Reliability

ABSTRACT

This paper involves an application of generalizability theory in assessing the dependability of a foreign language placement exam. The French Cloze test was administered to students within five levels of French classes and the results were scored by four different raters. Three specific generalizability coefficients are discussed along with implications of imposing three additional restrictions on the method by which future data are collected. The results show a very high item and student by item variance component and little variance due to the rater component. For this study adequate generalizability of students' scores was obtained for all generalizability coefficients using half of the total number of items on the exam and only one rater. Results indicate that in future decision studies involving tests of this nature each student should repond to the same set of items. Also, sufficient reliability may be obtained using only one rater per student but not all students need be rated by the same rater. (Author)

Educational Assessment Center

University of Washington

September 1976

76-24

The Application of Generalizability Theory to a College-Level French Placement Exam

Richard W. Naccarato

Gerald M. Gillmore

This paper involves an application of generalizability theory in assessing the dependability of a foreign language placement exam. The French cloze test was administered to students within five levels of French classes and the results were scored by four different raters. Three specific generalizability coefficients are discussed along with implications of imposing three additional restrictions on the method by which future data are collected.

The results show a very high item and student by item variance component and little variance due to the rater component. For this study adequate generalizability of students' scores was obtained for all generalizability coefficients using half of the total number of items on the exam and only one rater. Results indicate that in future decision studies involving tests of this nature each student should respond to the same set of items. Also, sufficient reliability may be obtained using only one rater per student but not all students need be rated by the same rater.

Educational Assessment Center Project: 563

The Application of Generalizability Theory to a College-Level French Placement Exam

A test, using the Cloze technique, was designed by Professor Victor Hanzeli of the University of Washington and pilot tested on UW students during Spring Quarter of 1975. This test consisted of five paragraphs in the French language, with 80 selected words deleted. The task of the students was to fill in the exact missing words. (Details of the development of the test may be obtained from Professor Hanzeli).

The test was administered to students in five classes of different levels during the final week of the quarter. The classes were numbered as follows: 103, 201, 202, 203, and 301 with 35, 16, 24, 24, and eight students respectively.

Each test was scored by the same four independent raters. Raters scored each item as follows: two points if the answer given was exactly that desired, one point if the answer given was a synonym, and zero points otherwise. Along with this scoring method, two others were possible: 1) An item could be considered correct only if it was scored two, and 2) An item could be considered correct if it was scored either one or two. The three methods yielded total scores which correlated very highly across all students ($r > .97$) and thus seemed unworthy of separate analyses. Arbitrarily, we chose the original method for all analyses to be reported here.

The purpose of this paper is to show how data collected in a design such as that described above can be analyzed through use of a generalizability theory (Cronbach, et al., 1972), so as to provide information on the dependability (reliability) of one's measurements within a variety of specific applications. Cronbach, et al., (1972) made a useful distinction between G-studies and D-studies.

The former are those done for the purpose of determining the magnitude of the various relevant sources of variance. The latter have as their purpose the providing of data for decision making. Data collected by a G-study can be also used for D-study purposes, however, the latter can be designed more efficiently if the former is done in advance.

The Design of the French Cloze Exam

In the present study the design of the G-study was a three-way completely crossed random effects analyses of variance design, with four raters rating 107 students on 80 items. For our purposes 30 items, which were equally dispersed among the five sections, were selected from the original 80 items on the exam. It was necessary to select only 30 items due to processing limitations of the computer program, however, as we shall see soon, it is possible to state reliabilities for any number of items.

Table 1 depicts the random effects model, where all effects are assumed to be sampled from infinite universes. Of course one may conceive of a fixed or mixed effects model in a situation such as this, however, for the sake of parsimony, the discussion will be limited to the random effects model. (See Kane and Brennan, in press, for a more detailed discussion.) In Table 1, students are designated S, raters R, and items I.

Insert Table 1 about here

The results of the analyses of variance of the data yielded by the French Cloze test are found in Table 2. Of particular interest are the estimated variance components. Relative to the others, the items and student by items components are very large. Raters and students by raters, on the other hand, are very small.

Insert Table 2 about here

Generalizability of the French Cloze Exam

There are three primary coefficients of generalizability of interest regarding the French Cloze test. The first of these, $\varepsilon\rho^2(R,I)$, is the case where we desire to generalize results over both items and raters, considering conditions for both of these facets to be samples from some larger universe of conditions. The coefficient $\varepsilon\rho^2(R)$ will stand for the case where raters are sampled from an infinite universe of raters, and the conditions of the item facet are assumed to exhaust all possible conditions of that facet, i.e., items are assumed to be finite. Case III will denote the situation where we are considering the items to be sampled from an infinite set and the raters to be a finite sample. The generalizability coefficient for this situation will be denoted $\varepsilon\rho^2(I)$.

Which of these three generalizability coefficients will be appropriate for a D-study depends upon its purpose. If one wants a score to be an estimate of what one would obtain responding to any infinite set of items all measuring ability in the French language and having the test scored by any of a large number of qualified raters, the $\varepsilon\rho^2(R,I)$ is most appropriate. If, on the other hand, generalization beyond the set of items or raters used is not desired, then $\varepsilon\rho^2(R)$ or $\varepsilon\rho^2(I)$ is appropriate. (The fourth logical coefficient, with raters and items both finite, is not estimable from the data of this design - see Kane and Brennan in press.)

Beyond the three possible situations or cases which we may consider for our decision study there are also three additional restrictions we may wish to impose on the decision model. The first of these restrictions is to assume

that all students will respond to the same set of items. This is different from the assumption that the items came from a finite set of items of a similar nature. Choosing to use $\varepsilon\rho^2(R)$ rather than $\varepsilon\rho^2(R,I)$ relates to the set of items to which we want to generalize. Choosing to employ restriction I relates to how we plan to administer the test. If all students get the same items, then the variance component for items does not enter into the generalizability calculation, because scores are based on a sum over the same items. Not employing restriction I implies a nested design where each student potentially receives a different set of items.

Restriction II on the full model implies that we have the condition that all students are rated by the same raters. In this case rater variance may be ignored between students since it is assumed that any effect due to different raters has an equal effect over all students.

Finally, the two previously mentioned restrictions can be combined into the situation where all students have the same raters and respond to the same exact set of items (restriction III). Again, it depends upon the purposes of the D-study as to whether or not these restrictions are appropriate.

Formulation and Discussion of the Generalizability Coefficients

This particular section will deal with the statistical formulation of the generalizability coefficients $\varepsilon\rho^2(R,I)$, $\varepsilon\rho^2(R)$, and $\varepsilon\rho^2(I)$ under the full model and with the three restrictions previously mentioned. The formulations are in accordance with recommended procedures for forming generalizability coefficients according to Cronbach, et al., (1972, chapter 3). The various formulas are presented for reference in Table 3.

Insert Table 3 about here

$\varepsilon\rho^2(R,I)$ - Generalizability Case I - Items and Raters Infinite

If we wish to generalize the results of this study over both items and raters, considering each of these facets to be samples randomly drawn from infinite universes of "similarly defined" items and raters, the appropriate coefficient of generalizability is $\varepsilon\rho^2(R,I)$. The "universe score" for a student is defined to be the expected value of his/her average score on the exam, taken over all possible samples of items and raters. The expected observed score variance for student mean scores (σ^2_{obs}), under the full model, is composed of universe score variance σ^2_s and error variance, and is represented by the formula:

$$(1) \qquad \sigma^2_s + \frac{1}{n_r}\sigma^2_r + \frac{1}{n_i}\sigma^2_i + \frac{1}{n_r}\sigma^2_{rs} + \frac{1}{n_i}\sigma^2_{is} + \frac{1}{n_r n_i}\sigma^2_{ri} + \frac{1}{n_r n_i}\sigma^2_e \quad ,$$

where n_r is the number of raters and n_i is the number of items involved in the decision study. Since we plan to generalize our results over both items and raters in Case I, the universe score variance (the numerator of $\varepsilon\rho^2(R,I)$) is found by taking the limits of σ^2_{obs} as n_r and n_i approach infinity. This leaves only σ^2_s as the estimated universe score variance. The resulting generalizability coefficient then is:

$$(2) \qquad \varepsilon\rho^2(R,I) = \frac{\sigma^2_s}{\sigma^2_{obs}} \quad ,$$

the ratio of universe and observed score variance.

The primary purpose of the generalizability study is to obtain estimates of the variance components (as shown in Table 2) so that we may formulate the

appropriate, or desired, coefficient for future decision studies. If we wish to estimate the dependability of student scores for a future decision study, it is only necessary to substitute into Formula 2 the appropriate number of raters and items for that study after the variance components have been estimated by the G-study.

The formulation of $\varepsilon\rho^2(R,I)$ changes if we wish to impose one of the restrictions, previously mentioned, on the design of the decision study. Table 3 describes the effects of each of the three restrictions upon the appropriate generalizability coefficient. Recall that choosing to employ one of these restrictions relates to how we plan to carry out the future decision study, and is part of the decision model. Which variance components are to be included in the calculation of the generalizability coefficient depends strictly on the purpose and design of the D-study.

$\varepsilon\rho^2(R)$ - Generalizability Case II - Items Finite and Raters Infinite

$\varepsilon\rho^2(R)$ is the coefficient we would employ if the desire is to generalize the results of this exam over raters, but not over items. The universe score for this coefficient is the expected value of a student's average score on the exam, as given by a random sample of raters from the domain of raters, using a finite set of these items in the D-study. With this coefficient we do not consider the items to be a sample from any larger set of items, and wish only to generalize our results for these items or some subset of them. Universe score variance is now equal to the observed score variance of Formula 1, as the number of raters (n_r) approaches infinity. The generalizability coefficient for this case is then given by:

$$(3) \qquad \varepsilon\rho^2(R) = \frac{\sigma^2_s + \frac{1}{n_i}\sigma^2_i + \frac{1}{n_i}\sigma^2_{is}}{\sigma^2_{obs}}$$

In Case I, where the universe score was defined as an expected value over an infinite set of items, the variance component $\frac{1}{n_i}\sigma^2_{si}$ was considered to be error variance. Now, since the items in the exam are assumed to exhaust the universe of items, we cannot consider the sampling of student by item interactions as being error. The differential response of students to the items is now legitimately a part of the universe score (as is item variance).

$\varepsilon\rho^2(I)$ - Generalizability Case III - Items Infinite and Raters Finite

The third generalizability coefficient, $\varepsilon\rho^2(I)$, is obtained if we desire to generalize the results of the exam over items, but not beyond the finite number of raters in the G-study. This coefficient is basically a measure of internal consistency of the items on the French Cloze exam. $\varepsilon\rho^2(I)$ is approximately equal to the expected correlation of any two measures of student performance on the items, based on an independent sample of items from the domain of items, and a common sample of raters from this finite set of raters. The universe score for $\varepsilon\rho^2(I)$ is defined as the expected value of the average student score on the exam, as given by a finite number of raters, using a random sample of items. Universe score variance is now composed of student variance, rater variance, and student by rater variance. The remaining four terms of the observed score variance (σ^2_{obs}) are considered differentiated error variance. The formulation of this third coefficient is as follows:

$$(4) \qquad \varepsilon\rho^2(I) = \frac{\sigma^2_s + \frac{1}{n_r}\sigma^2_r + \frac{1}{n_r}\sigma^2_{rs}}{\sigma^2_{obs}} \quad .$$

Results and Discussion

Values for the three generalizability coefficients from the French Cloze test under the full model and the three restrictions are found in Table 4.

For illustrative purposes, we have chosen four combinations: one item, one rater; 40 items, one rater; 80 items, one rater; and 80 items, four raters.

Insert Table 4 about here

Perusal of Table 4 reveals several important relationships, especially in the context of efficiently designing future D-studies. First notice that $\varepsilon\rho^2(R,I)$ is very close in magnitude to comparable values of $\varepsilon\rho^2(I)$. This is a direct result of the relatively small variance of the rater by student interaction. One implication of this is that it makes little difference whether one wants to treat raters as finite or infinite. Another implication is that reliabilities based on interitem consistency will not seriously overestimate $\varepsilon\rho^2(R,I)$.

Values of $\varepsilon\rho^2(R)$ tend to be much larger than either $\varepsilon\rho^2(R,I)$ or $\varepsilon\rho^2(I)$. Thus, treating items as finite has profound consequences on resulting generalizability coefficients. Furthermore, in most educational settings it would be a mistake to do so, since we are typically measuring general knowledge of a content rather than knowledge specific to the questions asked. Measures of inter-rater reliability will tend to grossly overestimate the values of $\varepsilon\rho^2(R,I)$.

Comparable values of the generalizability coefficients in the full model and restriction II are nearly equal, as are comparable values of the generalizability coefficients for restriction I and III. These similarities are a direct result of the relatively small variance component of raters. However, the item variance component is larger and causes restriction I and III to yield coefficients which are higher than those of the full model and restriction II. The decision-making implications of this are twofold. One does not

need to have every rater rate every student. Students can be nested within raters. However, unless the number of items is great (at least 40), one should have every student respond to the same set of items.

Finally, it is clear that increases in both raters and items will increase generalizability. However, the impact of each successive increase becomes increasingly less. In the present case, adequate generalizability is obtained with only one rater and forty items for both $\varepsilon\rho^2(R,I)$ and $\varepsilon\rho^2(I)$. This is especially true if all students respond to the same set of items. If the number of items is increased to 80, generalizability exceeds .90. Increases in raters produce very little increase in generalizability and the number of raters can probably be reduced to one.

Table 1

Random Effects ANOVA for 3-Way Completely Crossed Design

Source of variance	df	E(ms)
S	s-1	$\sigma^2(e) + r\sigma^2(si) + i\sigma^2(sr) + ri\sigma^2(p)$
R	r-1	$\sigma^2(e) + s\sigma^2(ri) + i\sigma^2(sr) + si\sigma^2(r)$
I	i-1	$\sigma^2(e) + s\sigma^2(ri) + r\sigma^2(si) + rs\sigma^2(i)$
SR	(s-1)(r-1)	$\sigma^2(e) + i\sigma^2(sr)$
SI	(s-1)(i-1)	$\sigma^2(e) + r\sigma^2(si)$
RI	(r-1)(i-1)	$\sigma^2(e) + s\sigma^2(ri)$
SRI(e)	(r-1)(s-1)(i-1)	$\sigma^2(e)$

r = number of raters

i = number of items

s = number of students

Table 2

The Analysis of Variance Summary Table

Source	ss	df	ms	$\hat{\sigma}^2$
S	1154.81	106	10.89	.074
R	13.89	3	4.63	.001
I	2937.49	29	101.29	.231
SR	38.34	318	.12	.001
SI	6090.42	3074	1.98	.475
RI	61.98	87	.71	.006
SRI(e)	738.55	9222	.08	.080

Table 3

Generalizability Formulas

n_r = number of raters n_i = number of items

Case I: Items and Raters Infinite

$$\varepsilon\rho^2(R,I) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{1}{n_r}\sigma_r^2 + \frac{1}{n_i}\sigma_i^2 + \frac{1}{n_r}\sigma_{rs}^2 + \frac{1}{n_i}\sigma_{is}^2 + \frac{1}{n_r n_i}\sigma_{ri}^2 + \frac{1}{n_r n_i}\sigma_e^2}$$

Let σ_{obs}^2 = the denominator of $\varepsilon\rho^2(R,I)$.

Case II: Items Finite and Raters Infinite

$$\varepsilon\rho^2(R) = \frac{\sigma_s^2 + \frac{1}{n_i}\sigma_i^2 + \frac{1}{n_i}\sigma_{is}^2}{\sigma_{obs}^2}$$

Case III: Items Infinite and Raters Finite

$$\varepsilon\rho^2(I) = \frac{\sigma_s^2 + \frac{1}{n_r}\sigma_r^2 + \frac{1}{n_r}\sigma_{rs}^2}{\sigma_{obs}^2}$$

Restrictions of the Full Model

Restriction 1: All students respond to the same set of items.

Eliminate $\frac{1}{n_i}\sigma_i^2$ from universe and observed score variance.

Restriction 2: All students are rated by the same raters.

Eliminate $\frac{1}{n_r}\sigma_r^2$ from universe and observed score variance.

Restriction 3: All students have the same items and raters.

Eliminate $\frac{1}{n_i}\sigma_i^2$, $\frac{1}{n_r}\sigma_r^2$, and $\frac{1}{n_r n_i}\sigma_{ri}^2$ from universe and observed score variance.

Table 4

Generalizability Coefficients

	I=No. of Items R=No. of Raters	Full Model	Res. I	Res. II	Res. III
CASE I					
$\varepsilon\rho^2(R,I)$ Items Infinite Raters Infinite	I=1, R=1	.083	.113	.083	.114
	I=40, R=1	.771	.822	.779	.831
	I=80, R=1	.860	.892	.871	.902
	I=80, R=4	.892	.925	.892	.925
CASE II					
$\varepsilon\rho^2(R)$ Items Finite Raters Infinite	I=1, R=1	.618	.836	.619	.845
	I=40, R=1	.896	.956	.905	.966
	I=80, R=1	.930	.964	.941	.976
	I=80, R=4	.964	.999	.964	.999
CASE III					
$\varepsilon\rho^2(I)$ Items Infinite Raters Finite	I=1, R=1	.084	.114	.085	.115
	I=40, R=1	.781	.833	.789	.843
	I=80, R=1	.872	.904	.882	.915
	I=80, R=4	.895	.929	.895	.929

References

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley & Sons, 1972.

Kane, M. T. and Brennan, R. L. The generalizability of class means. Review of Educational Research (in press).