DOCUMENT RESUME

ED 377 218 TM 022 386

AUTHOR Thompson, Bruce; Crowley, Susan
TITLE When Classical Measurement Theory Is Insufficient and Generalizability Theory Is Essential.
PUB DATE Apr 94
NOTE 19p.; Paper presented at the Annual Meeting of the Western Psychological Association (Kailua-Kona, HI, April 30, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Cutting Scores; Decision Making; Error of Measurement; *Estimation (Mathematics); *Generalizability Theory; Heuristics; *Measurement Techniques; Scores; Test Reliability; *Test Theory

ABSTRACT

Most training programs in education and psychology focus on classical test theory techniques for assessing score dependability. This paper discusses generalizability theory and explores its concepts using a small heuristic data set. Generalizability theory subsumes and extends classical test score theory. It is able to estimate the magnitude of multiple sources of error simultaneously, unlike classical theory, which enables only a single source of error to be considered at one time. Generalizability theory forces us to see that it is scores, not the tests themselves, that are reliable. Generalizability studies are the initial round of analyses that generate variance components for the sources of error in the study. Design studies use these variance components to answer questions about alternative measurement protocols. Generalizability analyses distinguish between decisions made in the context of cutoff scores (absolute decisions) and those that consider relative standing. An example involving 10 people who have taken a 4-item test each of 3 times illustrates application of generalizability theory. Six tables and two figures present details of the analysis. (Contains 11 references.) (SLD)

wpa.rev 5/7/94

WHEN CLASSICAL MEASUREMENT THEORY IS INSUFFICIENT AND GENERALIZABILITY THEORY IS ESSENTIAL

Bruce Thompson
Texas A&M University 77843-4225

Susan Crowley
Utah State University

Paper presented at the annual meeting of the Western Psychological Association, Kailua-Kona, Hawaii, April 30, 1994.

INTRODUCTION

Classical measurement theory has been used by psychologists since roughly the turn of the century. The results of analyses based on this theory continue to dominate the literature with coefficients of internal consistency reliability and test-retest reliability being commonly reported. The frequent reliance on classical test theory is somewhat unfortunate given the development of a broader and more powerful model for estimating the dependability of scores from behavioral measurements--**generalizability theory** (Cronbach, Gleser, Nanda, & Rajaratnum, 1972). As Jaeger (1991) notes, given the availability of this newer measurement theory,

> Thousands of social science researchers will no longer be forced to rely on outmoded [classical theory] reliability estimation procedures when investigating the consistency of their measurements. (Jaeger, 1991, p. x)

Although generalizability theory was developed several decades ago, the majority of training programs in education and psychology continue to focus on the more limited classical test theory techniques for assessing score dependability. Consequently, the purpose of the present paper is to discuss the elements of generalizability and to explore the concepts of this theory using a small heuristic data set.

REVIEW OF GENERALIZABILITY THEORY

The present discussion of the benefits of generalizability theory will be brief, given space limitations. For a more in-depth treatment of the topic, the reader is referred to Shavelson and Webb (1991), Eason (1991), and Brennan (1983).

Generalizability theory subsumes and extends classical test score theory. "G" theory is able to estimate the magnitude of the *multiple* sources of error simultaneously. Therefore, sources of error variance and interactions among these sources can be considered *simultaneously* in a single generalizability analysis. This is unlike classical test score analyses which allow for only a single source of error to be considered at one time. Not only does classical theory admit consideration of only one type of measurement error at a time, the theory does not consider the possible, completely independent or separate interaction effects of the sources of measurement error variance. For example, test-retest reliability estimates only consider variability (error) due to time. Similarly, internal consistency coefficients are based solely on variability (error) due to items. The test-retest analysis does <u>not</u> consider variability due to items, nor does the internal consistency analysis consider variability due to time. The classical test theory approach to score dependability is

graphically represented in Figure 1.

Simultaneous consideration both of multiple sources of error variance and of the interactions of these error sources is critical. An embedded assumption of many researchers using the classical score approach is that sources of error substantially overlap each other (e.g., the 15% of error from a test-retest analysis is the essentially the same error as the 10% or 15% measurement error detected in an internal consistency analysis) and that the sources of error do not interact to create additional error variance. As Thompson notes,

> in addition to being unique and cumulative, the sources [of error variance] may also interact to define disastrously large interaction sources of measurement error not considered in classical theory. The effects of these assumptions are all the more pernicious because of their unconscious character. (Thompson, 1991, p. 1072)

Since the goal of research is usually to generalize over items, occasions, test forms, administrations, etc., generalizability theory more closely honors the reality to which we wish to generalize.

Generalizability also forces us to see that it is particular scores, and **not** the tests themselves, that are reliable. Thus, the common telegraphic expression, indicating that "the test is reliable", is always literally untrue (Thompson, 1994). This habit of speech should be abandoned.

D-STUDIES

One source of frustration to those acquainting themselves with generalizability is the concept of D- and Design Studies and G- or Generalizability Studies. G-Studies are the initial round of analyses that generate variance components for the sources of error in the study. D-studies employ these variance components to answer questions about alternative measurement protocols. In addressing these "what if" questions (e.g., what if I administer only half of my items on two occasions? what if I use three raters instead of two?), the sources of error in the current assessment protocol can be pinpointed and the needed changes in the assessment regime specified to achieve a desired level of generalizability.

Additionally, these analyses can be used to make decisions regarding assessment with an acceptable cost-benefit ratio. If using 40 rather than 20 items yields the same improvement in score reliability as measuring three rather than two times with 20 items, and administering the longer measure twice is more cost effective, the researcher or practitioner is informed that the use of this more efficient measurement protocol is tenable.

RELATIVE VERSUS ABSOLUTE DECISIONS

Generalizability analyses distinguish between decisions made in the context of cutoff scores (absolute decisions), as against decisions only considering stability in which relative standing is a concern. Classical test theory does not admit a distinction between reliability involving *absolute* decisions made in the context of cutoff scores (e.g., intervention decisions invoking a cutoff score for an early intervention program) as against reliability involving decisions only considering stability in *relative* standing or rankings (e.g., always using the top 25% of samples for intervention even though score distributions over time may move up or down in an absolute sense).

This distinction can be important, particularly when decisions regarding intervention or subject selection are involved. Suppose that a child has the highest depression score in the school. That is relative position. If the score surpasses an identified criterion for moderate-severe depression (an absolute criterion), then an intervention will be conducted. If at the end of treatment the child still has the highest depression score in school, but it is now below the cut-off score, the treatment will likely be considered successful. In generalizability studies the coefficients that address reliability in the context of *relative* decisions, i.e., decisions only concerned with the stability of score rankings, are called *generalizability* coefficients. The coefficients that address reliability in the context of *absolute* decisions, i.e., decisions invoking cutoff score criteria, are called *phi* coefficients.

"G" THEORY ANALYSIS USING A HEURISTIC EXAMPLE

The reasons why generalizability theory is important have been summarized previously. Some of the required calculations will be summarized in the following example. Though these are automated in widely available software, this review may facilitate conceptual understanding of the analysis.

More thorough reviews of the importance of generalizability theory are available elsewhere (Eason, 1991; Shavelson, Webb, & Rowley, 1989; Thompson, 1991). More thorough treatments of the mechanics of the analysis are available from Shavelson and Webb (1991) or Webb, Rowley, and Shavelson (1989), or at the more advanced level from Brennan (1983, 1994).

Data

Table 1 presents a hypothetical data to illustrate these calculations. The example involves 10 people who have taken a four item test each of three times.

Classical Test Score Analyses

Table 2 presents all the classical test theory reliability estimates for the Table 1 data. Inspection of the table reveals that the internal consistency of the data varied considerably across occasions, from a high (and possibly acceptable level) of .77 to a distressingly low .19. Similarly, the test-retest coefficients ranged from a high of .68 to a low of -.08.

Partitioning of the Score Variance

As Brennan notes (1983), the variance components for the data to be analyzed are central because "they are the building blocks that provide a crucial foundation for all subsequent analyses" (p. 11). Variance components can be calculated using an ANOVA-type approach. There is the overall (grand) mean, and variance component for each "main" effect (i.e., Individuals, Occasion, Variables), and a variance component for each "interaction" effect (e.g., I x O, I x V, O x V, I x O x V). The variance components are represented graphically in Figure 2.

In generalizability theory, we partition the score variance into its various uncorrelated components. First, we partition the systematic variance (usually the variance associated with people, since we usually presume this is "true" variance when people truly vary as individuals) and error variance. Then we further partition the error variance into its main effect and interaction effect components.

ANOVA can be used to create orthogonal or uncorrelated partitions of variance. First, we compute the sum-of-squares (SOS) for the various variance sources, as illustrated in Table 3. These sum-of-squares are then converted into mean squares and then into variance components for scores, as illustrated in Table 4.

Next, the variance components for scores are converted in to variance components for means, as illustrated in Table 5. Table 6 illustrates the use of these variance components to estimate the proportion of score variance that is systematic for either making relative decisions in which only stability of score ordering is relevant, or for making absolute decisions in which stability of scores in relation to a score cutoff or external criterion (e.g., a passing score criterion) is relevant.

CONCLUSIONS

- Results of classical test theory and generalizability estimates of the dependability of the same scores can be very different.
- Classical test score estimates ranged from .77 to -.08, depending on the source of error considered. Generalizability

estimates considered **all** sources of error simultaneously and resulted in a "G" coefficient of .47 (phi coefficient = .42).

- Additional calculations using the variance components could identify specific ways that the dependability of the measurement could be increased (e.g., increase the number of items, increase the number of occasions).

- The result from the small heuristic data set demonstrate the **necessity** of moving away from a classical test theory approach whenever multiple sources of measurement error are presumed to exist simultaneously. These analyses further confirm other comparisons between generalizability theory and classical test theory presented in the literature (e.g., Crowley, Thompson, & Worchel, in press).

- Doctoral level training in measurement should focus on teaching generalizability theory as a more contemporary, broader, and more powerful paradigm for assessing score dependability.

References

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brennan, R. L. (1994). Variance components in generalizability theory. In C. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective. New York: Plenum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnum, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: John Wiley.

Crowley, S.L., Thompson, B., & Worchel, F.F. (in press). The Children's Depression Inventory: A comparison of generalizability and classical test score analyses. Educational and Psychological Measurement.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples (Vol. 1, pp. 83-98). In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. Greenwich, CT: JAI Press.

Jaeger, R. (1991). Foreword. In R.J. Shavelson & N.M. Webb, Generalizability theory: A primer (pp. ix-x). Newbury Park: SAGE Publications.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park: SAGE Publications.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

Thompson, B. (1994, January). It is incorrect to say "the test is reliable": Bad language habits can contribute to incorrect or meaningless research conclusions. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1989). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Table 1

Hypothetical Data for an Individuals x Occasion x Variables Measurement Protocol

	Rater #1				Rater #2				Rater #3			
	Variable				Variable				Variable			
Person	1	2	3	4	1	2	3	4	1	2	3	4
1	8	5	4	4	4	2	3	4	5	6	2	0
2	5	2	6	6	6	4	4	4	6	6	4	1
3	2	0	2	2	4	2	2	4	5	4	0	7
4	6	4	4	2	2	0	3	2	4	2	2	4
5	8	1	8	6	6	6	2	6	1	6	4	2
6	2	3	2	4	2	2	5	2	3	4	0	1
7	6	2	6	6	6	4	4	6	4	4	4	3
8	4	7	4	6	6	4	4	3	2	1	0	0
9	8	8	7	7	5	5	5	4	4	7	0	0
10	3	3	3	0	2	4	5	0	4	6	0	0

gen5a.wk1

Table 2

Classical alpha and Test-retest Correlation Coefficients For the Table 1 Data

		Administration		
Administration	α	First	Second	Third
First	.7737	1.0000		
Second	.6311	.6787 (46.06%)	1.0000	
Third	.1875	-.0850 (00.72%)	.1331 (01.77%)	1.0000

Note. Squared test-retest correlation coefficients are presented in parentheses as percentages.

Table 3

Partitioning the Variability in the Table 1 Scores

$SOS_I = (\Sigma\ (\bar{X}_i - \bar{X})^2)\ (o=3\ (v=4))$

$= ((3.92 -3.683)^2 +(4.50 -3.683)^2 +(2.83 -3.683)^2 +(2.92 -3.683)^2 +$
$(4.67 -3.683)^2 +(2.50 -3.683)^2 +(4.58 -3.683)^2 +(3.42 -3.683)^2 +$
$(5.00 -3.683)^2 +(2.50 -3.683)^2)\ (o=3\ (v=4))$

$= (0.0544 +0.6669 +0.7225 +0.5878 +$
$0.9669 +1.4003 +0.8100 +0.0711 +$
$1.7336 +1.4003)\ (o=3\ (v=4))$

$= 8.4139\ (o=3\ (v=4)) = 8.4139\ (12) =$ **100.9668**

$SOS_O = (\Sigma\ (\bar{X}_o - \bar{X})^2)\ (i=10\ (v=4))$

$= ((4.40 -3.683)^2 +(3.70 -3.683)^2 +(2.95 -3.683)^2)\ (i=10\ (v=4))$

$= (0.5136 +0.0003 +0.5378)\ (i=10\ (v=4))$

$= 1.0517\ (i=10\ (v=4)) = 1.0517\ (40) =$ **42.0666**

$SOS_V = (\Sigma\ (\bar{X}_v - \bar{X})^2)\ (i=10\ (o=3))$

$= (4.43 -3.683)^2 +(3.80 -3.683)^2 +(3.30 -3.683)^2 +(3.20 -3.683)^2)$
$(i=10\ (o=3))$

$= (0.5625 +0.0136 +0.1469 +0.23361)$
$(i=10\ (o=3))$

$= 0.9567\ (i=10\ (o=3)) = 0.9567\ (30) =$ **28.7000**

$SOS_{IO} = (\Sigma\ (\bar{X}_{io} - \bar{X})^2)\ (v=4) - SOS_I - SOS_O$

$= ((5.25 -3.683)^2 +(3.25 -3.683)^2 +(3.25 -3.683)^2 +$
$(4.75 -3.683)^2 +(4.50 -3.683)^2 +(4.25 -3.683)^2 +$
$(1.50 -3.683)^2 +(3.00 -3.683)^2 +(4.00 -3.683)^2 +$
$(4.00 -3.683)^2 +(1.75 -3.683)^2 +(3.00 -3.683)^2 +$
$(5.75 -3.683)^2 +(5.00 -3.683)^2 +(3.25 -3.683)^2 +$
$(2.75 -3.683)^2 +(2.75 -3.683)^2 +(2.00 -3.683)^2 +$
$(5.00 -3.683)^2 +(5.00 -3.683)^2 +(3.75 -3.683)^2 +$
$(5.25 -3.683)^2 +(4.25 -3.683)^2 +(0.75 -3.683)^2 +$
$(7.50 -3.683)^2 +(4.75 -3.683)^2 +(2.75 -3.683)^2 +$
$(2.25 -3.683)^2 +(2.75 -3.683)^2 +(2.50 -3.683)^2)$
$(v=4) - SOS_I - SOS_O$

$= (2.4544 +0.1878 +0.1878 +$
$1.1378 +0.6669 +0.3211 +$
$4.7669 +0.4669 +0.1004 +$
$0.1003 +3.7378 +0.4669 +$
$4.2711 +1.7336 +0.1878 +$
$0.8711 +0.8711 +2.8336 +$
$1.7336 +1.7336 +0.0044 +$
$2.4544 +0.3211 +8.6044 +$
$14.5669 +1.1378 +0.8711 +$
$2.0544 +0.8711 +1.4003)\ (v=4) - SOS_I - SOS_O$

$= 61.1166\ (v=4) -100.9668 -42.0666$

$= 244.4664 -100.9668 -42.0666 =$ **101.4333**

$SOS_{IV} = (\Sigma (\underline{X}_{iv} - \overline{X})^2)\ (o=3) - SOS_I - SOS_V$

$= ((5.67 -3.683)^2 +(4.33 -3.683)^2 +(3.00 -3.683)^2 +(2.67 -3.683)^2 +$
$(5.67 -3.683)^2 +(4.00 -3.683)^2 +(4.67 -3.683)^2 +(3.67 -3.683)^2 +$
$(3.67 -3.683)^2 +(2.00 -3.683)^2 +(1.33 -3.683)^2 +(4.33 -3.683)^2 +$
$(4.00 -3.683)^2 +(2.00 -3.683)^2 +(3.00 -3.683)^2 +(2.67 -3.683)^2 +$
$(5.00 -3.683)^2 +(4.33 -3.683)^2 +(4.67 -3.683)^2 +(4.67 -3.683)^2 +$
$(2.33 -3.683)^2 +(3.00 -3.683)^2 +(2.33 -3.683)^2 +(2.33 -3.683)^2 +$
$(5.33 -3.683)^2 +(3.33 -3.683)^2 +(4.67 -3.683)^2 +(5.00 -3.683)^2 +$
$(4.00 -3.683)^2 +(4.00 -3.683)^2 +(2.67 -3.683)^2 +(3.00 -3.683)^2 +$
$(5.67 -3.683)^2 +(6.67 -3.683)^2 +(4.00 -3.683)^2 +(3.67 -3.683)^2 +$
$(3.00 -3.683)^2 +(4.33 -3.683)^2 +(2.67 -3.683)^2 +(0.00 -3.683)^2)$
$(o=3) - SOS_I - SOS_V$

$= (3.9336 +0.4225 +0.4669 +1.0336 +$
$3.9336 +0.1003 +0.9669 +0.0003 +$
$0.0003 +2.8336 +5.5225 +0.4225 +$
$0.1003 +2.8336 +0.4669 +1.0336 +$
$1.7336 +0.4225 +0.9669 +0.9669 +$
$1.8220 +0.4669 +1.8225 +1.8225 +$
$2.7220 +0.1225 +0.9669 +1.7336 +$
$0.1003 +0.1003 +1.0336 +0.4669 +$
$3.9336 +8.9003 +0 1003 +0.0003 +$
$0.4669 +0.4225 +1.0336 +13.5669)$
$(o=3) - SOS_I - SOS_V$

$= 69.7666\ (o=3) -100.9668 -28.7000$

$= 209.3 -100.9668 -28.7000 = \mathbf{79.6333}$

$SOS_{OV} = (\Sigma (\overline{X}_{ov} - \overline{X})^2)\ (i=10) - SOS_O - SOS_V$

$= (5.20 -3.683)^2 +(3.50 -3.683)^2 +(4.60 -3.683)^2 +$
$(4.30 -3.683)^2 +(4.30 -3.683)^2 +(3.30 -3.683)^2 +$
$(3.70 -3.683)^2 +(3.50 -3.683)^2 +(3.80 -3.683)^2 +$
$(4.60 -3.683)^2 +(1.60 -3.683)^2 +(1.80 -3.683)^2)$
$(i=10) - SOS_O - SOS_V$

$= (2.3003 +0.0336 +0.8403 +$
$0.3803 +0.3803 +0.1469 +$
$0.0003 +0.0336 +0.0136 +$
$0.8403 +4.3403 +3.5469)\ (i=10) - SOS_O - SOS_V$

$= 12.8566\ (i=10) -42.0666 -28.7000$

$= 128.5666 -42.0666 -28.7000 = \mathbf{57.8000}$

$SOS_{IOV} = (\Sigma (\overline{X}_{iov} - \overline{X})^2) - SOS_I - SOS_O - SOS_V - SOS_{IO} - SOS_{IV} - SOS_{OV}$

$=((8-3.683)^2 +(5-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 +$
$(3-3.683)^2 +(4-3.683)^2 +(5-3.683)^2 +(6-3.683)^2 +(2-3.683)^2 +(0-3.683)^2 +$
$(5-3.683)^2 +(2-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 +$
$(4-3.683)^2 +(4-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 +(1-3.683)^2 +$
$(2-3.683)^2 +(0-3.683)^2 +(2-3.683)^2 +(2-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 +$
$(2-3.683)^2 +(4-3.683)^2 +(5-3.683)^2 +(4-3.683)^2 +(0-3.683)^2 +(7-3.683)^2 +$
$(6-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 +(2-3.683)^2 +(0-3.683)^2 +$
$(3-3.683)^2 +(2-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 +(2-3.683)^2 +(4-3.683)^2 +$
$(8-3.683)^2 +(1-3.683)^2 +(8-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +$

$$
\begin{aligned}
&(2-3.683)^2 +(6-3.683)^2 +(1-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 + \\
&(2-3.683)^2 +(3-3.683)^2 +(2-3.683)^2 +(4-3.683)^2 +(2-3.683)^2 +(2-3.683)^2 + \\
&(5-3.683)^2 +(2-3.683)^2 +(3-3.683)^2 +(4-3.683)^2 +(0-3.683)^2 +(1-3.683)^2 + \\
&(6-3.683)^2 +(2-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 + \\
&(4-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(3-3.683)^2 + \\
&(4-3.683)^2 +(7-3.683)^2 +(4-3.683)^2 +(6-3.683)^2 +(6-3.683)^2 +(4-3.683)^2 + \\
&(4-3.683)^2 +(3-3.683)^2 +(2-3.683)^2 +(1-3.683)^2 +(0-3.683)^2 +(0-3.683)^2 + \\
&(8-3.683)^2 +(8-3.683)^2 +(7-3.683)^2 +(7-3.683)^2 +(5-3.683)^2 +(5-3.683)^2 + \\
&(5-3.683)^2 +(4-3.683)^2 +(4-3.683)^2 +(7-3.683)^2 +(0-3.683)^2 +(0-3.683)^2 + \\
&(3-3.683)^2 +(3-3.683)^2 +(3-3.683)^2 +(0-3.683)^2 +(2-3.683)^2 +(4-3.683)^2 + \\
&(5-3.683)^2 +(0-3.683)^2 +(4-3.683)^2 +(6-3.683)^2 +(0-3.683)^2 +(0-3.683)^2) \\
&- SOS_I - SOS_O - SOS_V - SOS_{IO} - SOS_{IV} - SOS_{OV}
\end{aligned}
$$

$$
\begin{aligned}
=(&18.6336 + 1.7336 + 0.1003 + 0.1003 + 0.1003 + 2.8336 + \\
&0.4669 + 0.1003 + 1.7336 + 5.3670 + 2.8336 + 13.5669 + \\
&1.7336 + 2.8336 + 5.3670 + 5.3670 + 5.3670 + 0.1003 + \\
&0.1003 + 0.1003 + 5.3670 + 5.3670 + 0.1003 + 7.2002 + \\
&2.8336 + 13.5669 + 2.8336 + 2.8336 + 0.1003 + 2.8336 + \\
&2.8336 + 0.1003 + 1.7336 + 0.1003 + 13.5669 + 11.0003 + \\
&5.3670 + 0.1003 + 0.1003 + 2.8336 + 2.8336 + 13.5669 + \\
&0.4669 + 2.8336 + 0.1003 + 2.8336 + 2.8336 + 0.1003 + \\
&18.6336 + 7.2002 + 18.6336 + 5.3670 + 5.3670 + 5.3670 + \\
&2.8336 + 5.3670 + 7.2002 + 5.3670 + 0.1003 + 2.8336 + \\
&2.8336 + 0.4669 + 2.8336 + 0.1003 + 2.8336 + 2.8336 + \\
&1.7336 + 2.8336 + 0.4669 + 0.1003 + 13.5669 + 7.2002 + \\
&5.3670 + 2.8336 + 5.3670 + 5.3670 + 5.3670 + 0.1003 + \\
&0.1003 + 5.3670 + 0.1003 + 0.1003 + 0.1003 + 0.4669 + \\
&0.1003 + 11.0002 + 0.1003 + 5.3670 + 5.3670 + 0.1003 + \\
&0.1003 + 0.4669 + 2.8336 + 7.2002 + 13.5669 + 13.5669 + \\
&18.6336 + 18.6336 + 11.0003 + 11.0003 + 1.7336 + 1.7336 + \\
&1.7336 + 0.1003 + 0.1003 + 11.0003 + 13.5669 + 13.5669 + \\
&0.4669 + 0.4669 + 0.4669 + 13.5669 + 2.8336 + 0.1003 + \\
&1.7336 + 13.56691 + 0.1003 + 5.3670 + 13.5669 + 13.5669) \\
&- SOS_I - SOS_O - SOS_V - SOS_{IO} - SOS_{IV} - SOS_{OV} \\
&= 555.9666 - 100.9668 - 42.0666 - 28.7000 \\
&\qquad - 101.4333 - 79.6333 - 57.8000 = \mathbf{145.3666}
\end{aligned}
$$

gen5a.wk1

Table 4

The Conversion of Sums-of-Squares to Mean Squares
and Then to Score Variance Components

Source	(SOS	/df	=MS)	+	(MS	MS	MS)	=	Sum
i	(100.9668	/ 9	=11.2185)	+	(-5.6352	-2.9494	+2.6920)	=	5.3259
o	(42.0666	/ 2	=21.0333)	+	(-5.6352	-9.6333	+2.6920)	=	8.4568
v	(28.7000	/ 3	= 9.5667)	+	(-2.9494	-9.6333	+2.6920)	=	-0.3240
io	(101.4333	/18	= 5.6352)	+	(-2.6920			=	2.9432
iv	(79.6333	/27	= 2.9494)	+	(-2.6920			=	0.2574
ov	(57.8000	/ 6	= 9.6333)	+	(-2.6920			=	6.9414
iov	(145.3666	/54	= 2.6920)					=	2.6920

Source	Sum	/	(k (k))	=	Sum	/	Product	Variance Components
i	5.3259	/	(o=3 (v=4))		5.3259	/	12	0.4438
o	8.4568	/	(i=10 (v=4))		8.4568	/	40	0.2114
v	-0.3240	/	(i=10 (o=3))		-0.3240	/	30	-0.0108
io	2.9432	/	(v=4)		2.9432	/	4	0.7358
iv	0.2574	/	(o=3)		0.2574	/	3	0.0858
ov	6.9414	/	(i=10)		6.9414	/	10	0.6941
iov	2.6920				2.6920			2.6920

Table 5

Conversion of Score Variance Components
to Variance Components for Means

Source	Variance Component$_S$	Frequency	Variance Component$_M$
i	0.44382	1	0.44382
o	0.21141	3	0.07047
v	0[a]	4	0
io	0.73580	3	0.24526
iv	0.08580	4	0.02145
ov	0.69413	12	0.05784
iov	2.69197	12	0.22433

Note. Variance Component$_S$ is the variance component for *single observations* originating in a given variance source. Variance Component$_M$ is the variance component for *mean scores* originating in a given variance source. The "object of measurement" (here i individual people, who are presumed to truly vary) has its frequency set to 1.

[a]Negative variance components are set to 0 prior to further calculations.

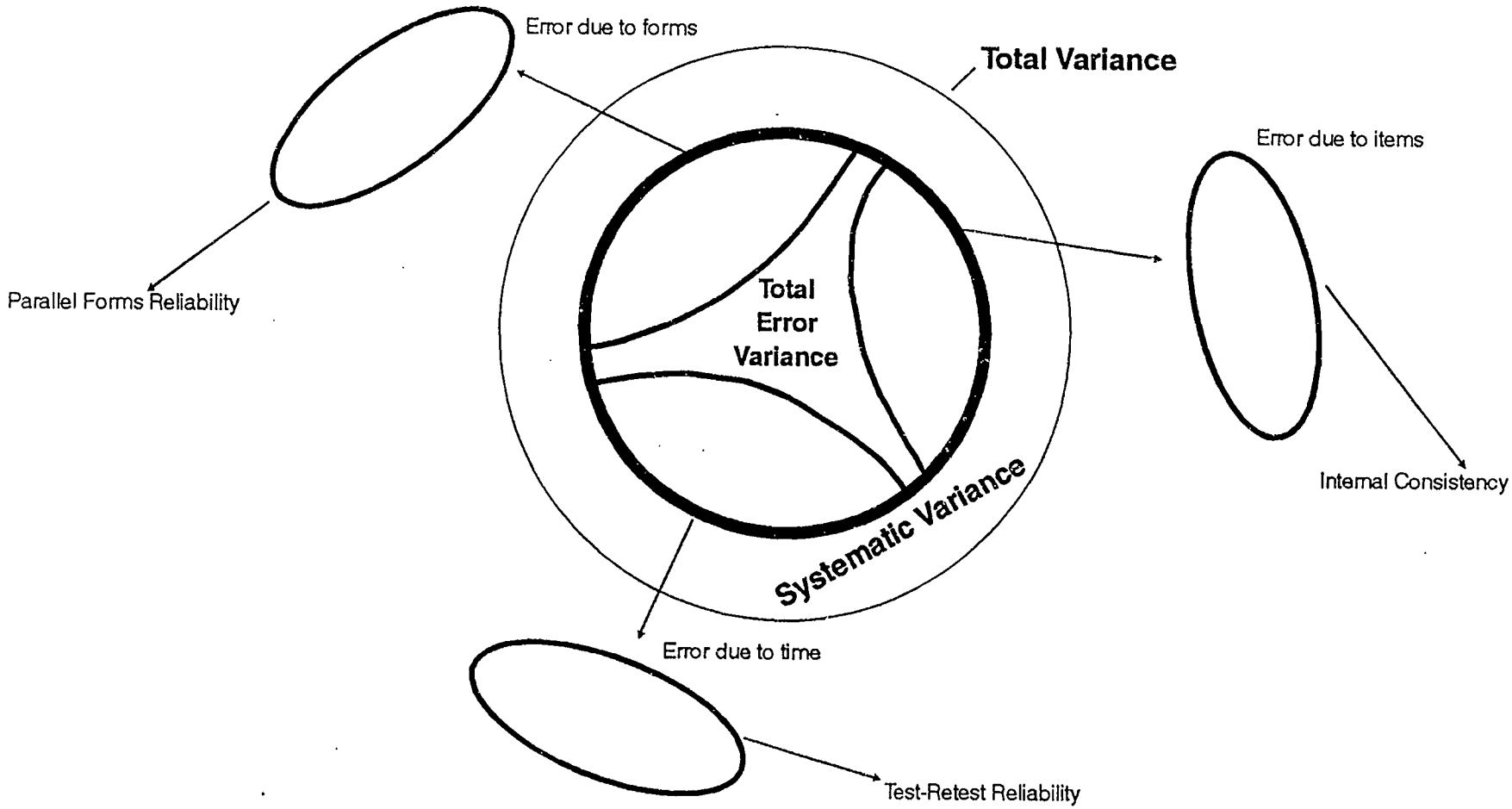
Table 6

The Conversion of Variance Components Into
Proportions of Total Variance That are Systematic or Reliable

Source/ Statistic	Variance Component$_M$	%age	Systematic Variance	Error Variance (Relative)	Error Variance (Absolute)
i	0.4438	41.74	0.4438		
o	0.0705	6.28			0.0705
v	0	0			0
io	0.2453	23.07		0.2453	0.2453
iv	0.0214	2.02		0.0214	0.0214
ov	0.0578	5.44			0.0578
iov	0.2243	21.10		0.2243	0.2243
Sum	1.0632		0.4438	0.4910	0.6194
Coefficient					
G				*0.4747*	
Phi					*0.4174*

gen5a.wk1

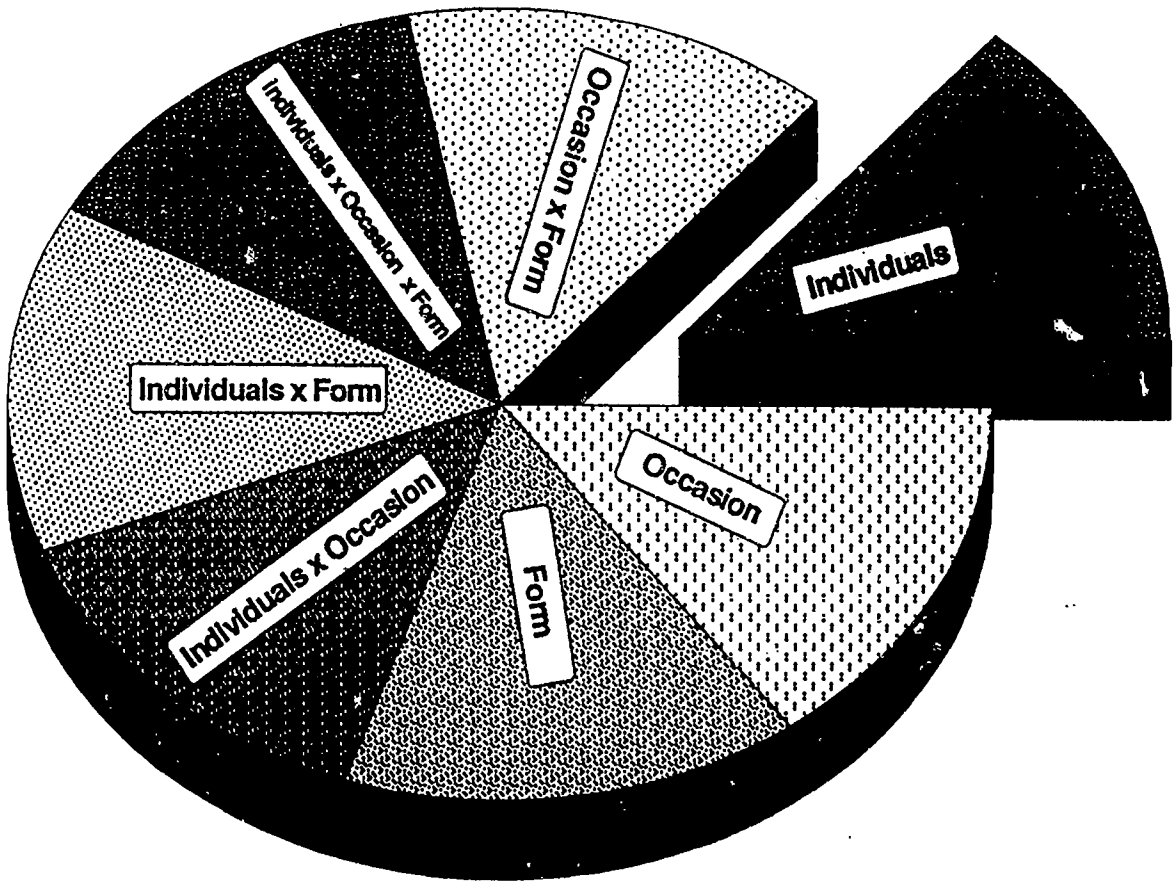
Note. The proportion of measurement error associated with *relative* decisions only evaluating rank-order score stability is evaluated by the *generalizability* coefficient, and equals systematic variance divided by total relevant variance ([0.4428 / (0.4438 + 0.4910)] = 0.4747). The proportion of measurement error associated with *absolute* decisions evaluating score stability in relation to a fixed score cut-off standard is evaluated by the *phi* coefficient, and equals systematic variance divided by total relevant variance ([0.4428 / (0.4438 + 0.6194)] = 0.4174).



In each analysis, only a single source of error is considered, items **OR** time **OR** form

Problems: Errors may be completely independent (any single coefficient overestimates the dependability of measure)
Errors may interact to create new sources of error

Figure 1: Classical Test Score Analyses



NOTE: If "Individuals" are the "object of measurement," the main effect for individuals is systematic variance, while all other partitions are error.

Figure 2: Partitioning of Error Variance