DOCUMENT RESUME

ED 204 374 TM 810 362

AUTHOR Stenner, A. Jackson; Rohlf, Richard J.

TITLE Construct Definition Methodology and Generalizability Theory Applied to Career Education Measurement.

PUB DATE [79]

NOTE 24p.

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Career Education; Definitions; Grade 9; *Measurement Techniques; Secondary Education; *Test Reliability; Vocational Maturity

IDENTIFIERS Career Maturity Inventory (Crites); *Constructs; *Generalizability Theory

ABSTRACT

The merits of generalizability theory in the formulation of construct definitions and in the determination of reliability estimates are discussed. The broadened conceptualization of reliability brought about by Cronbach's generalizability theory is reviewed. Career Maturity Inventory data from a sample of 60 ninth grade students is used to demonstrate the power of the technique to estimate reliability coefficients for a number of differing measurement procedures. It is concluded that researchers frequently use reliability coefficients that are inflated estimates of the precision with which their constructs are being measured. (Author/GK)

Construct Definition Methodology and Generalizability Theory Applied to Career Education Measurement

A. Jackson Stenner and Richard J. Rohlf

Introduction

The field of career education measurement is in disarray. Evidence mounts that today's career education instruments are verbal ability measures in disguise (See Westbrook's chapter in this volume). A plethora of trait names such as career maturity, career development, career planning, career awareness, and career decision making have, in the last decade, appeared as labels to scales comprised of multiple choice items. Many of these scales appear to be measuring similar underlying traits and certainly the labels have a similar sound or "jingle" to them. Other scale names are attached to clusters of items that appear to measure different traits and at first glance appear deserving of their unique trait names, e.g., occupational information, resources for exploration, work conditions, personal economics. The items of these scales look different and the labels correspondingly are dissimilar or have a different "jangle" to them.

As instrument developers and users we commit the "jingle" fallacy (Green, 1974) when we give the same or nearly the same name to clearly distinct underlying traits. Similarly, we commit the "jangle" fallacy when different labels are assigned to essentially the same underlying trait. When a trait label such as Career Maturity is assigned to a set of items which in fact measures verbal ability, we have committed the jangle fallacy. Whenever we find evidence that two similarly named scales are only moderately correlated, there exists the possibility of the jingle fallacy.

Whether or not a given scale is a measure of verbal ability as opposed to career maturity is, of course, a question of validity, i.e., is the scale actually a measure of "what it is intended to measure"---or is it? This chapter asserts that the current state of affairs in career education measurement exists because of the lack of carefully defined and operationalized career education constructs and will suggest a theory and methodology that researchers and practitioners will, hopefully, find useful in their continuing efforts to develop and refine measurement in the field of career education.

Construct Definition

Constructs are the means by which science orders observations. We take it on faith that the universe of our observations can be ordered and subsequently understood with a comparatively small number of constructs or inferred organizing influences. Observations are aggregated and constructs created through the mental processes of abstraction and induction. When we observe a group of children and describe some of the children as more aggressive than others, we employ a construct. We create the construct "aggression" by observing that certain behaviors tend to vary together and this pattern of covariation among observations we come to designate as aggression. In describing the differences in behavior among children, we might conclude that one child is much more aggressive than other children. We arrive at this conclusion informally by summing up the frequency of observed aggressive acts and we use the total score as an index of each child's level of aggression. These total scores are then compared and we arrive at decisions about each child.

This process of weighting individual observations, aggregating the observations into a total score and then checking the quality of the construct score by determining how well the total score can predict the original observations happens so fast and so frequently and works so well in our everyday lives, that there is seldom need to reflect critically on the process itself. The search for pattern or regularity among observations is, it seems, just as central to our daily lives as it is to scientific activity. Perhaps because the process of observation, abstraction and construct formation is so fundamental to daily functioning, it is taken for granted in behavioral science research. Often observations in the form of questionnaire items and test questions are aggregated without adequately examining the assumptions and implications inherent in the summation and averaging procedures. The simple fact that observations are combined and a total score computed means that we entertain a hypothesis that the observations are in some way related to one another. If the observations are uncorrelated, then combining them into a total score is a meaningless undertaking, since the total or construct score will carry no information about the original observations and consequently will be of no value in explaining anything else. If, however, the observations are correlated, then the construct score has meaning. Precisely what meaning depends upon the perceived nature of the organizing influences responsible for the correlations among observations.

A construct then is a theory which expresses how its inventor "construes" a set of interrelated observations. Construct labels (e.g., career maturity, occupational information, career decision making) serve as shorthand expressions for hypotheses regarding the nature of the predominant organizing influences responsible for correlations among observations.

What constitutes an observation? In career education measurement the most common "observation" would be a person's response to a test or rating item. Such observations provide information about a person's placement on a scale and serves as an indicant of the extent to which the subject possesses the attribute or trait being measured. A set of such indicants (items) comprises an instrument. The underlying structure or organizing influence operating on these observations is often determined by some combination of statistical structural analysis, e.g., factor analysis, and a logical analysis of the item content. Corroboration of the underlying structure is then frequently sought by confirmation via hypothesis testing and correlations with other instruments measuring conceptually similar and dissimilar constructs.

All observation whether made in service of the behavioral or physical sciences, is prone to error. Error is given more attention in behavioral sciences measurement probably because it exists in such abundance. Because of its abundance the process of construct definition must incorporate a theory of error. Various approaches to estimating the reliability of a measurement procedure rest on different assumptions about error and how it affects the observations we make.

Classical reliability theory is based on Spearman's model of an observed score (e.g., observation). Basically, an observed score is a function of two components, a true score and an error score. Within this framework, models of reliability have been formulated to assess the relative importance of each component. Campbell(1976) gives an excellent review of the historical development of reliability theory. All traditional measures of reliability (alpha, equivalent forms, retest) describe the agreement among repeated measurements of the same individauls. Although these reliability measures differ in their definition of error, they all assume a <u>single undifferentiated</u> source of error. Coefficient alpha attributes error to inconsistency in the extent to which individual items measure an attribute. Measures of stability such as test-retest or equivalent forms reliability coefficients, attribute error to changes in testing conditions; mood of examinee, etc.

In recent times authors such as Tryon (1957); Cronbach, et al. (1963), Cronbach, Glesar, Nanda, and Rajaratnam (1972); Nunally (1967); and Lord and Novick (1968) have departed from the classic concept of true vs. error scores and have instead incorporated what has become to be known as the domain sampling theory of reliability. The notion of a true score was replaced by a "domain" or "universe" score which is an individual's score if all observations in a domain or universe could be averaged. Measurement error in this framework is the extent to which a sample value differs from the population value.

This change in focus from a "true score" to "universe score" resulted in increased importance being placed on defining the "universe" from which a particular sample of items has been drawn and to which we want to generalize. Initially the concept of universe was restricted to thinking in terms of a universe of content, e.g., sampling of reading comprehension items from a universe of possible reading comprehension items. However, the work of Cronbach, et al. has broadened this original conceptualization. His work, referred to as generalizability theory, speaks to sampling of "conditions of measurement" which include additional sources of variation to that of just variation among samples of items, or components of content. This broadened conceptualization can be viewed as a change from a focus on the reliability of an instrument to a focus on the reliability of a measurement procedure.

For example, suppose the career maturity of a group of students is rated on a number of items by a number of different teachers on several different occasions. The traditional view of a content domain would focus on the items as a sample from the universe of all such similar items. However, Cronbach's generalizability theory forces us to acknowledge that there are probably systematic differences in item scores across occasions which do not reflect true change in level of career maturity; and, to recognize that there are systematic differences among students that are reflected in observed scores which are not necesarily due to difference in career maturity, e.g., socioeconomic status. Thus, from this perspective we are not only concerned with a universe of possible career maturity items, but, in addition, we need to think in terms of a universe of possible teachers and a universe of possible occasions, and a universe of possible respondents. Actually Cronbach does not talk in terms of different universes; but rather, each of the above would be considered a "facet" in the universe of measurement conditions. The more facets one chooses to include in defining a construct, the broader the universe of generalization. Cronbach also refers to facets as either "random" or "fixed." A fixed facet would be one that would not vary, i.e., would be a constant in the universe. For example, if raters were considered to be a fixed facet in a measurement procedure, the investigator would be planning to always use the same rater(s) whenever a measurement was taken. Given this condition, there would be no systematic differences in observed scores due to idiosyncratic differences in rating behavior among raters. However, if raters were considered to be a "random" facet, the investigator would be broadening the construct definition of career maturity such that a person's "universe score" would be an average score across the universe of career maturity items judged across all possible raters. In the "fixed" case, the "universe score" would be an average score across the universe of items as judged by a particular rater or set of raters.

As discussed above, the process of construct definition begins with the recognition that observed scores (observations) are determined by some set of underlying organizing influences. In addition to "wanted" influences causing variation among scores, we must also recognize that there are "unwanted" (error) influences exercising potentially biasing or misleading effects on observed scores. Generalizability theory enables us to specify these sources of variance in observed scores in terms of characteristics of the object of measurement, characteristics of the indicants (items), characteristics of the context of measurement, and the interactions both within and across those categories.

In addition to a conceptual model, generalizability theory, using analysis of variance procedures, provides the techniques by which we can specify the sources of variance (both wanted and unwanted) in observed scores and estimate the magnitude of their effects. The procedure also yields a generalizability coefficient(s) which can be interpreted in a manner similar to traditional reliability coefficients, e.g., in estimating the standard error of measurement. However, before these analysis of variance procedures can be applied, it is necessary to design a study in which sources of variance are systematically varied.

Generalizability theory makes a distinction between G and D studies. A G study is a study in which data is collected in order to examine a wide range of sources of variance affecting a measurement procedure whereas a D study (for <u>D</u>ecision) selects either the G study design or some modification of that design for use in estimating the generalizability coefficient that can be expected in some subsequent application of the measuremennt procedure. A D study does not involve the gathering of data but rather uses the variance estimates from the sources designed into the G study to estimate what the generalizability coefficient would be under alternative construct definitions and sampling specifications. For example, suppose that the authors of a career maturity scale employ a p:c x i x occ (persons nested within class crossed with items crossed with occasion) G study design. That is, the career maturity scale is administered to several classes on at least two occasions. Under this

Scenario #'s 3, 4 and 5 all employ the broadest permissible construct definition (items and moments random) but the sampling frequencies for items and/or moments differ. In Scenario #3 we estimate what the generalizability coefficient would be if 50 items were administered on one occasion (i.e., moment). Note that the generalizability coefficient under this construct is coincidentally the same as that observed under Scenario #1. As a rule when the construct definition is broadened and the sampling specifications are unchanged, the generalizability coefficient goes down. Similarly when the construct definition is narrowed and consequently the universe of generalization is narrowed, the generalizability coefficient is increased. The reasoning for this outcome is straightforward; if the universe under examination is quite broad, then a larger number of observations must be sampled to attain a specified level of precision, whereas a narrower universe permits a smaller number of observations to attain the same precision. Under Scenario #4 the item sample remains at N_i=50 but the number of testing sessions is increased, N_m=2, resulting in an improvement in the generalizability coefficient ($\Sigma\hat{\rho}^2$=.81). Finally, under Scenario #5 sampling frequencies are increased for both items (N_i=100) and moments (N_m=3) resulting in a substantial increase in precision of measurement.

Classical reliability theory, as practiced in the field of career education measurement, is unnecessarily restrictive. Disciples of classical theory compute a number of equivalence coefficients by correlating student performance on split halves of an instrument or by computing coefficient alpha (KR-20) for an instrument administered on a single occasion. Similarly, stability or retest coefficients are computed by correlating student performance on one occasion with performance, say, two weeks later. Finally, interrater reliability coefficients are computed by correlating the ratings of two or more raters of the same behavior. Each of these forms of reliability coefficient reflect a single undifferentiated source of error and, more importantly, derive from different construct definitions. Coefficient alpha accurately reflects an instrument's reliability under the highly restricted construct definition which treats items as the only random facet and, consequently, the p x i interaction (confounded with the residual) as the only source of error. The stability coefficient properly reflects an instrument's reliability under a construct definition that treats occasions as the only random source of variance and the persons x occasion interaction as the only source of error.

It is important to recognize that traditional forms of reliability permit error variance to be confounded with true score variance because they do not differentiate among the many possible sources of error. For example, the test-retest reliability coefficient will not "break out" a p x i interaction and thus that variance will be a "hidden" component of the true score variance. Likewise, the p x occ interaction will be a hidden component of the true score variance when calculating coefficient alpha. This can result in an artificially inflated estimate of true score variance which, in turn, results in an inflated reliability coefficient. The attractiveness of generalizability theory is that it permits simultaneous consideration of items, occasions and other facets as random sources of error.

Generalizability theory provides a framework that forces us to better conceptualize the constructs we use. Unfortunately many investigators do not invest sufficient time in construct definition activities. Time and again researchers move ahead to answer substantive research questions without carefully defining the constructs that figure in their theories or program evaluations. The most common mistake found in the behavior sciences is that investigators will state a conceptually broad construct definition, but will use a reliability estimate that is based on a much narrower definition, thus yielding a coefficient which exaggerates the precision with which the construct can be measured. Elsewhere we have argued that career education will go the way of many previous fads unless, as a field, it can stake out a set of well-defined constructs and related instrumentation (Stenner, Strang, Baker, 1978). So far, efforts in this regard have been disappointing.

Some Special Applications

Many seemingly diverse issues in measurement can be accommodated within generalizability theory. Cronbach (1972) states:

> What appears today to be most important in G Theory is not what the book gave greater space to. In 1972 G Theory appeared as an elaborate technical apparatus. Today the machinery looms less large than the questions the theory enables us to pose. G Theory has a protean quality. The procedures and even the issues take a new form in every contest. G Theory enables you to ask your questions better; what is most significant for you cannot be supplied from the outside (p.199).

In the discussion to follow we attempt to illustrate the range of applications for which generalizability theory, coupled with construct definition methodology, can be useful.

Toward a Theory of the Indicant

A historical convention for which we can find no rational explanation has contributed to avoidance of a potentially fruitful type of construct validation study. The convention is to report person scores as number of items (or indicants) correct and item scores as proportion of respondents answering an item correctly. Thus person scores and item scores are expressed in different metrics, leading some investigators to assume that there is some fundamental difference in the way people and items can be analyzed. For example, construct validity studies often emphasize relationships between theoretically relevant valirbles and the construct under study and use the person as the unit of analysis. On the other hand, little work has been done in explaining variance in item scores. Some authors, including ourselves, contend that many career development scales containing items of the multiple choice variety in fact measure verbal ability and not career maturity (see Westbrook's chapter in this volume). One approach to investigating this contention which, on the face, seems more direct than focusing on person score correlations, would involve predicting item scores using a set of item readability and syntax measures as well as theoretically derived ratings of the extent of career maturity called for by each item. If the readability and syntax measures explain a large proportion of the variance in item scores and the theory-based ratings explain little of the variance, then it is likely that the so-called career development items are really verbal reasoning items in disguise. If a construct is really well defined, then it should be possible to explain the behavior of indicants of that construct, i.e., explain variance in item or indicant scores. Unfortunately this is a test that few constructs in the behavioral sciences, let alone career education, have passed.

Generalizability of Ratings

Many outcomes in career education do not readily lend themselves to paper-pencil testing. For example, outcomes such as employability skills, personal work habits and job interview behavior are better measured by trained observers in either real or simulated settings. Generalizability theory provides a framework for estimating the dependability of these ratings.

Suppose a career education program sets about to improve the job interview behavior of a group of students. Five employers from the local community are called upon to interview each student and complete a rating scale. One highly informative design for examining the generalizability of these ratings would be p x i x r (persons crossed with items crossed with raters). Thus each student would be rated by each employer on all items. Under this design the broadest permissible construct definition generalizes over items and raters.

Separate estimates of alpha or interrater agreement would overestimate the precision with which the construct as defined can be measured. The generalizability coefficient more accurately reflects our measurement precision and provides information on how the precision can be increased to an acceptable level. One excellent illustration of this type of analysis is provided by Gilmore, Kane and Naccarato (1978). Note that this design does not include the "occasion" facet. If a sizeable p x occ interaction exits, our estimate of measurement precision may be inflated if we have defined our construct to be stable over time.

Competency Testing

Some career education programs have objectives which state that all students will attain a particular mastery level in reading and mathematics. In assessing this objective, instrumentation is needed that has a special kind of reliability. Discussion above focused on developing instruemnts that would maximally differentiate among objects (e.g., students). In competency or mastery testing, the objective is to differentiate among two groups of students, those that have attained the minimal performance level and those that have not. Generalizability theory provides a framework for studying the dependability of mastery or competency decisions. The most thorough treatment of this application of generalizability theory is provided by Brennan and Kane (1977).

Generalizability of Class Means

Some career education evaluations employ class or school rather than student as the unit of analysis (Stenner, Strang, Baker, 1978); Brennan (1975); Kane, Gilmore and Crooks (1976). Haney (1974) and Kane & Brennan (1977) have suggested that generalizability theory provides a conceptually and practically appealing approach to estimating the reliability of class means. The simplest design from which we can estimate the generalizability of class means is p:c x i (persons nested within class crossed with items). Note that this is the familiar persons x items design (from which coefficient alpha is computed) with the addition that knowledge is available on class membership. Under this design we can estimate the reliability of persons nested within classes and class means. In a more complex design such as p:c:s x i, the object of measurement might be persons (p), classes (c) or schools (s). In general, this type of split plot design can prove particularly useful in an evaluation in which multiple units of analysis (e.g., students, classes, schools) are employed (Hayman, Rayder, Stenner and Madey, 1979). As a rule, generalizability coefficients should be computed for each unit of analysis employed in a research or evaluation study.

In passing we should note that applications of generalizability theory in which class or school is the object of measurement, have focused exclusively on the mean or first moment of the distribution. Lohnes (1972), in an excellent but largely ignored paper, demonstarted that using the variance of a class or school distribution as an independent variable might also be useful in predicting outcomes. In such studies interest is centered on differentiating classrooms not in terms of their means, but rather in terms of their variances while generalizing over occasions or some other random facet.

Issues Of Test Bias

Much attention and controversy has surrounded the issues of race and sex bias in testing. Although there are many types of bias, perhaps the most pernicious is that which gives members of particular racial, ethnic or sex groups unfair advantage in responding to certain kinds of items. It is somewhat ironical that career eudcation has as one of its goals eradication of sex role stereotyping (Hoyt, 1975) and yet we were unable to find any studies of sex bias in career education measurement.

Some forms of item bias can be effectively studied within the framework of generalizability theory. For example, a simple p: s x i (persons nested within sex group crossed with items) will provide information on possible sex bias. In this design the component of variance related to sex bias is the s x i (sex by item) interaction. If this component of variance is large then items have a different meaning (i.e., measure something different) for males and females. Examination of the items contributing most heavily to the interaction can sometimes lead to explanations for the source of the bias (e.g., terminology unfamiliar to males or females). Students can also be nested within race groups to evaluate racial bias or nested within reading level groups to evaluate the extent to which item meaning is conditional on student reading level.

Although the literature on bias has focused almost exclusively on racial, ethnic and sex characteristics, the notion of bias is a generic concept. Any characteristic of the object of measurement which interacts with a random facet represents bias. For social and other reasons, item scores which are conditional on race and sex (i.e., interact with race and sex) have received the bulk of attention. From theoretical as well as practical perspectives, there are other types of bias that pose equally troublesome problems. For example, items that take on radically different meanings depending upon the examinee's reading level are just as invalid as indicators of career maturity as items that are conditional on the sex or race of the examinee.

design the broadest permissible construct definition generalizes over items and occasions with either person or class as the object of measurement. Suppose that an investigator has limited time available for student testing and wanted to know what the effect would be of reducing by 50% the number of items on the scale. This scenario could be set up and a generalizability coefficient estimated given specification of the object of measurement and construct definition (which facets are considered fixed and which random). This application of generalizability theory is analogous to power analysis (Cohen, 1977) in which different sampling scenarios are evaluated to determine the probability of detecting an effect. In D studies different measurement scenarios (alternative construct definitions coupled with alternative sampling frequencies) are evaluated to determine the precision with which objects of measurement can be differentiated.

An Illustration of Generalizability Theory

Before proceeding with an example, it may prove useful to reflect on the meaning of a generalizability coefficient as well as its general form. A generalizability coefficient is simply the ratio of true score variance (or universe score variance) to the sum of true score variance and error variance:

$$\hat{\Sigma}_{\rho}^{2} = \frac{\text{True score variance}}{\text{True score variance + error variance}} = \frac{\tau}{\tau + \delta}$$

The components that enter true score variance and error variance change as the construct definition changes, but the basic expression for a generalizability coefficient remains the same. Following are several descriptive comments about the generalizability coefficient that may help in gaining an intuitive grasp of what this ratio means:

- One task of measurement is to differentiate among objects (e.g., classrooms or children) on some scale while simultaneously generalizing over selected facets. The higher the generalizability coefficient, the better the differentiation or separation among objects.

- Children or classrooms differ on a scale for many reasons (we usually refer to these reasons as sources). Some of these reasons are important to us and represent what we want to measure, and others are not of interest and represent noise. Differences among students that arise due to reasons we are interested in, we call true score differences whereas differences due to reasons we are not interested in, we call error differences. A generalizability coefficient is simply the ratio of average squared differences between objects that arise from wanted sources divided by the average squared differences between objects arising from wanted and unwanted sources.

- Observed score variance is the sum of true score variance and error variance. Thus the generalizability coefficient represents the proportion of observed score variance that is due to "wanted" sources of variance. If the generalizability coefficient is high, then a high proportion of the variance in observed scores is due to wanted sources of variation, whereas if the coefficient is low, it means that only a small proportion of differences among objects is due to wanted sources of variation.

- We can conceive of the generalizability coefficient as a heuristic that describes the confidence with which we can reject the null hypothesis that all objects' true scores are equal. Statisticians would use an F ratio for this purpose and, in fact, for the simple persons x items (p x i) design:

$$\hat{\Sigma}_\rho^2 = \frac{F-1}{F} \quad \text{or} \quad F = \frac{1}{1-\hat{\Sigma}_\rho^2}$$

- The generalizability coefficient is the squared correlation between observed scores and true scores. The true score is the average score we would obtain if all observations across the random facets of the universe of generalization could be exhaustively sampled. Errors of measurement (unwanted reasons that objects have different scores) contribute to ordering people differently on observed scores (which are samples) than they would be ordered if their scores could be averaged over all facets of interest (e.g., items or days during a two-week period). The generalizability coefficient provides an indication of how differently people are likely to be ordered if exhaustive sampling of all relevant observations was possible.

In summary, the generalizability coefficient provides an estimate of the precision of measurement given a construct definition. It is meaningless to refer to a reliability or generalizability coefficient without reference to the governing construct definition. What construct definition is most appropriate in a given situation is a substantive question that often cannot be answered by measurement specialists. What definition to employ is a complex question which takes us back to what we want our construct to mean. Ascribing meaning to constructs and increasing our understanding of variance arising from applications of our measurement procedures is what the process of construct definition is all about.

Table 1 provides estimated G study variance components for a p x i x m design, and Table 2 displays different D study designs or measurement scenarios. The data used in this illustration was graciously provided by Dr. Bert Westbrook and represents a subsample of the ninth grade data used in his chapter of this volume. The sample consists of 60 students responding to the 50 attitude items of the Career Maturity Inventory (CMI).

Examination of Table 1 reveals that a large proportion (60%) of the variance on this instrument is unexplained by facets of the measurement procedures. The second and third largest components of variance are the item (i) and person x item (p x i) interaction, respectively. The person (p) component explains four percent of the universe variance. The moment (m), person x moment (p x m) interaction and item by moment (i x m) interaction explain very small proportions of the variance.

A major advantage of generalizability theory is that the theory specifies which sources of variance are to be ignored, which contribute to true score (τ) and which contribute to error (δ) in estimating the generalizability of a measurement procedure under a particular construct definition. Whether stated or not, there are two essential aspects of a measurement procedure that must be made explicit before any reliability or generalizability coefficients can be interpreted. These are (1) the construct definition, i.e., which facets are to be considered random and which fixed, and (2) the sampling frequencies for each facet included in the construct definition.

Table 2 presents construct definitions and sampling specifications for five scenarios. Scenario #1 displays the generalizability coefficient under the classical reliability formulation in which moments (i.e., short-term occasions) are fixed (N_m=1) and items are random. The generalizability co-

Table 1

Illustrative Example of Generalizability Analyses for the Attitude Subscale of the CMI

Source	Notation	SS	df	MS	Estimated Variance Component	Estimated Proportion of universe varian attributable t to each source
Person	p	68.87	59	1.167	.00889	04
Item	i	270.41	49	5.519	.04328	18
Moment	m	1.20	1	1.204	.00030	00
Person x Item	pxi	661.62	2891	.229	.04111	17
Person x Moment	pxm	11.55	59	.196	.00098	00
Item x Moment	ixm	11.87	49	.242	.00159	01
Person x Item x Moment	$pxixm_e$	423.88	2891	.147	.14662	60

efficient ($\Sigma\hat{\rho}^2$ =.72) under this scenario accurately describes the precision of measurement only if our interest centers on how well students can be differentiated on a single occasion. This coefficient corresponds to the traditional coefficient alpha (or KR-20).

In scenario #2 moments are random and items are fixed. This construct definition corresponds to the traditional stability or retest coefficient. In other words, within the framework of generalizability theory the traditional retest coefficient may be computed under a generalizability design of the form p x i x m where items constitute a fixed facet and moments constitute a random facet. Note that in this case the retest coefficient is higher than the internal consistency coefficient because the p x m variance component accounts for virtually no variance whereas the p x i interaction (which contributes to the true score when items are fixed) accounts for 17% of the universe variance.

Table 2

Illustrative Scenario Table

Scenario	Construct Definition		Sampling	Sources of Variance							Generalizability
#	Random Facets	Fixed Facets	Specifications	p	i	m	pxi	pxm	mxi	$pxixm_e$	Coefficient
1	Items	Moments	$N_i = 50$ $N_m = 1$	τ	-	-	δ	τ	-	δ	.72
2	Moments	Items	$N_m = 1$ $N_i = 50$	τ	-	-	τ	δ	-	δ	.78
3	Items Moments		$N_i = 50$ $N_m = 1$	τ	-	-	δ	δ	-	δ	.72
4	Items Moments		$N_i = 50$ $N_m = 2$	τ	-	-	δ	δ	-	δ	.81
5	Items Moments		$N_i = 100$ $N_m = 3$	τ	-	-	δ	δ	-	δ	.92

REFERENCES

Brennan, R. L. and Kane, M.T. An Index of Dependability for Mastery Tests. Journal of Educational Measurement, 1977, Vol. 14, No. 3, pp. 277-289.

Campbell, J. P. Psychometric Theory. In M.D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally, 1976.

Cohen, J. Statistical Power Analysis For the Behavioral Sciences. New York: Academic Press, 1977, (Rev. Ed.).

Crites, John. Career Maturity Inventory. Monterey, California: CTB/McGraw Hill, 1973.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. The Dependability of Behavioral Measurements: Multifacet Studies of Generalizability. New York: John Wiley and Sons, 1972.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. Theory of Generalizability: A liberalization of Reliability Theory. British Journal of Statistical Psychology, 1963, 16, Part 2, 137-163.

Gilmore, G. M., Kane, M. T., and Naccarato, R. W. The Generalizability of Student Ratings of Construction: Estimation of the Teacher and Course Components. Journal of Educational Measurement, 1978, Vol. 15, No. 1, pp. 1-13.

Green, D. R. (Ed) The Aptitude-Achievement Distinction. Monterey: CTB/ McGraw-Hill, 1974.

Haney, W. The Dependability of Group Mean Scores. Unpublished special qualifying paper, Harvard Graduate School of Education, October 1974.

Hayman, John, Rayder, Nick, Stener, A. Jackson, and Madey, Doren L. On Aggregation, Generalization, and Utility In Educational Evaluation. Educational Evaluation and Policy Analysis, July-August, Vol. 1, No. 4, 1979.

Hoyt, Kenneth B. An Introduction to Career Education: A Policy Paper of the U. S. Office of Education. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1975.

Kane, M. T. and Brennan, R. L. The Generalizability of Class Means. The Review of Educational Research, 1977, Vol. 47, No. 1, pp. 267-292.

Kane, M. T., Gillmore, G. M., and Crooks, T. J. Student Evaluations of Teaching: The Generalizability of Class Means. Journal of Educational Measurement, 1976, 13, 171-183.

Lohnes, Paul. Statistical Descriptors of School Classes. American Educational Research Journal, 1972, Vol. 9, pp. 547-556.

Lord, F. M., and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Nunnally, J. C. Psychometric Theory. New York: McGraw Hill, 1967.

Stenner, A. Jackson, Strang, Ernest W., Baker, Robert F. Technical Assistance In Evaluating Career Education Projects: Final Report. Durham, N.C.: NTS Research Corporation, 1978.

Tryon, R. C. Reliability and Behavior Domain Validity: Reformulation and Historical Critique. Psychological Bulletin, 1957, 54, 229-249.