11

THE STRUCTURE OF GENERALIZABLLITY THEORY

FOR

HIERARCHICALLY STRATIFIED TESTS

Hans-Magne Eikeland
University of Oslo

Oslo, February 1972

This report is a preliminary version issued for limited circulation. Corrections, criticisms, and suggestions for revision are solicited. The report should not be cited as a reference without the specific permission of the author.

Contents

1. Introduction.

Among other things, test theory may be said to be concerned with developing a rationale for making socalled psychometric inferences. In this type of inference making one intends to generalize to a universe of tests, rather than to a population of individuals which is a statistical inference problem (Kaiser & Caffrey 1965). The characteristics of universes of tests have been variously conceived. Classical test theory defined, syntactically, the universe of tests as composed of homogeneous items very restrictively, such that the universe consisted of what might be called fixed parallel tests, meaning that the universe could only include tests that were exactly like in certain statistical respects (Gulliksen 1950, Tryon 1957). A modern and liberalized view, generalizability theory, conceives of a universe of tests as being made up of random parallel tests. A random parallel test is construed to be a probabilistic sample from a defined universe of tests, each test being composed by randomly picking items from a homogeneously defined pool of items (Cronbach, Rajaratnam & Gleser 1963). Thus, random parallel tests can not be exactly like in statistical properties.

The generalizability problem in psychometric inference is to estimate for the random parallel test the squared correlation between an observed test score and the universe score, thus giving the proportion of observed test score variance that

is determined by the universe score.The universe score is defined as the average test score in the universe of tests.The generalizability coefficient can also be defined as the expected correlation among random parallel tests as distinct from the reliability coefficient in classical test theory which is the correlation between fixed parallel tests (Cronbach 1951, Eikeland 1970).

Similar for both classical test theory and generalizability is that theory development has been restricted to dealing with a presumed homogeneous universe of test items.Test theory has until recently been concerned with the simplest of all test designs,the person by item design,although practical test construction for a long time has been going along lines that implicitly presupposes a theory for a more complex conception of item universes as being multifacet in nature.Test theory undoubtedly has lagged far behind test construction.Test batteries are being used for which there is no theory available. Multiple score tests are perhaps more commonly applied in practical testing than single score tests,but even recent advanced textbooks in mental test theory,e.g.Lord & Novick(1968),are exclusively dealing with theoretical issues associated with the homogeneous test.

Certainly,interesting theory development lies ahead for making psychometric inferences to universes of tests that are constructed according to more complex sampling plans for universes of items conceived of as multifacet as compared to the simple sampling plan involved in the construction of single factor

tests.The psychometric problem at issue in the present monograph is to conceive of a structural theory on which to base generalizability estimates for test batteries that are constructed according to a particularly construed multifacet universe of items.

2.The concept of a multifacet measuring operation.

When more than one source of variance is associated with a measuring instrument,that instrument is said to be multifacet. A rating procedure involving raters,only one trait being rated, is a onefacet operation.Guilford's (1954) classical rating problem involving raters and traits is conceptually a twofacet operation in that the ratees will be given both rater scores and trait scores.This twofacet procedure could be extended to a threefacet operation by stratifying raters into groups of raters.By this procedure ratees could be given trait scores,group of raters scores and rater scores.

Medley and Mitzel (1963) have treated multifacet operations for measuring classroom behavior by systematic observation. Their cris study involving classes,recorders,items,and situations is a fourway analysis of variance design;however,it is a threefacet measuring operation.Only recorders,items,and situations are in this study identifying aspects of the measuring procedure.Thus,the homogeneous test is a onefacet instrument in that only items are identified as a source of variance tied to the measuring operation.Yet,the design for analyzing

observed data from such a test is a twoway analysis of variance design,involving persons in addition to items as sources of variance.

When Rajaratnam,Cronbach & Gleser (1965) estimated the generalizability coefficient for the stratified parallel test,they were involved in a twofacet study,and not a onefacet study as maintained by Gleser,Cronbach & Rajaratnam (1965).In the Rajaratnam,Cronbach & Gleser (1965) study items and strata are facets.

There should be no reason to regard multifacet studies as different from stratified studies,e.g.studies in which items are grouped into defined strata such that a hierarchical design is formed.This seems also to be the conclusion drawn by Cronbach, Gleser,Nanda & Rajaratnam (1967) in commenting on stratified test construction: "It appears advantageous to reinterpret this as a multifacet problem,especially as this then opens the way to considering simultaneously the sampling of items and the sampling of other conditions"(p59).

A simple rule of thumb for deciding on the number of facets in a measuring operation is to count the number of main effects directly connected with the operation.

Many classification schemes for stratifying measuring operations into facets are conceivable.In testing,content,format,and occasions are common facets.In Guilford's (1967) structure of intellect,content,product,and operations are facets.So are also the types of content within content,types of product with-

in product,and types of operation within operation.In fact, the types are facets on a lower level.

Horst (1965) has discussed the various modes or categories which are fundamental to the investigation of a system of variation.His concept of mode fits well in to a multifacet system where characteristics of persons or entities are assessed by multiple procedures:

> (Therefore,) some systems,to be satisfactorily and completely characterized,may well take into account observations or recordings for a number of different entities (persons) on a number of different attributes on a number of different occasions by a number of different evaluators with respect to a number of different conditions or instructions.(Horst 1965,10)

Horst's system constitutes within the conceptual framework of the present monograph a fourfacet measuring operation.Attributes,occasions,evaluators,and conditions are facets,while persons are the entities being assessed.

When measuring operations are made into systems of facets,very complex variance structures of observed individual differences will be the result.While classical test theory was able to distinguish conceptually among many types of variation that go into a test score,the models for that theory could handle only two sources of variance at a time,namely the universe score variance and one undifferentiated error variance (Thorndike 1951, Magnusson 1967).What is at issue in making efforts toward a theory development for complex test designs,is how to treat multiple sources of test score variance simultaneously and how to make a rational decision for how to interpret the various

sources as being signal or noise in the particular context a measuring operation is being used.Here is where the multifacet studies are extremely challenging both from a syntactical and a semantical point of view.

The multifacet measuring operation of concern in this report,is a threefacet[e/] test having strata,substrata,and items as identifying aspects.This particular test design may be said to originate in a structural conception of the item universe which calls for a more complex sampling plan than commonly met in unstratified test construction.The theory development will be especially concerned with defining universes of threefacet tests of this particular design to which one wants to generalize.For this purpose mathematical models have to be built to fit definitions and interpretations of the test scores determined by multiple sources of variance.

3.Previous work on stratified tests:Twofacet studies.

The reliability problem of stratified tests,or test batteries, has been of some concern for test theory for a long time. Outstanding references are:Jackson & Ferguson 1941,Cronbach 1951, Mosier 1951, Tryon 1957. The split-half and the test-retest approach to the reliability of a stratified test is not of any interest in the present context where the internal consistency approach is of concern.No satisfactory general solution to the internal consistency problem of stratified tests was obtained within classical test theory.The correlation of sums

approach to this problem, like the solution reached by Tryon (1957), is in principle a special case of a more general solution[1] to be reviewed shortly.

Rather than give a complete historical account of the internal consistency problem of the stratified test, emphasis will be put on some recent formulations.

Rajaratnam, Cronbach & Gleser (1965) reformulated the reliability problem of stratified tests to fit a generalizability theory. They conceived of a universe of items that had been identified and divided into strata. To make test construction follow a formal sampling plan, they construed a test battery to be made up of a predetermined number of randomly sampled items from within the identified strata. Such a test may be regarded as one of an indefinitely large number of tests that may be constructed according to the same sampling plan provided the sub-universes of items are regarded as infinite. These tests form a universe of stratified tests. It is to this universe one wants to generalize, i.e. to estimate the squared correlation between the observed score of a randomly picked test and the universe score, the average test score across the universe of tests. Characteristic for the development by Rajaratnam, Cronbach & Gleser (1965) is that they restricted their definition of the universe of tests to a fixed number of strata, those represented in the particular test at hand. This will often be a realistic restriction in that these strata are the very strata of interest, or they exhaust the possibility of obtaining strata.

Yet,one may start playing with a more general formulation of how to define such stratified tests.Rabinowitz & Eikeland (1964) made an extension of the classical Hoyt (1941) procedure for finding the reliability of a stratified test where strata could rationally be regarded as random.This means that the strata actually found in the test at hand,by no means could be conceived to exhaust the strata to which the test constructor wanted to generalize.Thus,in the Rabinowitz & Eikeland formulation,two models for estimating the generalizability of a stratified test were developed,a fixed and a random model.The random model regards both items within strata and strata as randomly sampled from subuniverses of items and from a universe of strata.The fixed model regards items as random samples from within[a/] fixed number of strata.

Surely,items in generalizability theory will always be considered random.In effect,this is the hallmark of the theory.Although random strata may be more difficult to imagine than fixed ones, it is interesting to make formulations that are so general that such a possibility is included.

In moving from the twofacet test to the threefacet test it is the intention to extend the general formulation made by Rabinowitz & Eikeland to fit a still more complex test design.As will be shown later,there is a relationship between the original Hoyt analysis of the unstratified,or onefacet,test via the twofacet test to the threefacet test.

4.The concept of the hierarchically stratified test.

One distinct characteristic of the stratified ,or hierarchical, test is the nesting of items within strata.This means that there is no rationally based one-to-one correspondence between items from stratum to stratum.If such a correspondence could be established,one would have a crossed twofacet test design in that all possible combinations of strata and items are present in data.Many multifacet operations are crossed.The Medley & Mitzel (1963) classroom observation design and the Guilford (1954) rater-trait design mentioned above,have crossed facets.

In the stratified test the nesting of items comes from the fact that strata are thougnt to contain distinguishable types of items.One can pernaps most easily see how such types of items can be distinguished by conceiving of a stratification of a universe of items on the basis of content.

Now,a further stratification procedure on a universe of items can be thought of taking place,generating new nesting on other levels in the hierarchical structure of items.One can stratify already grouped items into strata of a higher order,or one can make finer groupings of already grouped items,generating strata of a lower order.

For the present purpose a second-order stratification of an item universe will do to make clear what is meant by a hierarchically stratified test.The unstratified test implies a zero-order stratification.What is usually called the stratified test, the test design described by Rajaratnam,Cronbach & Gleser (1965)

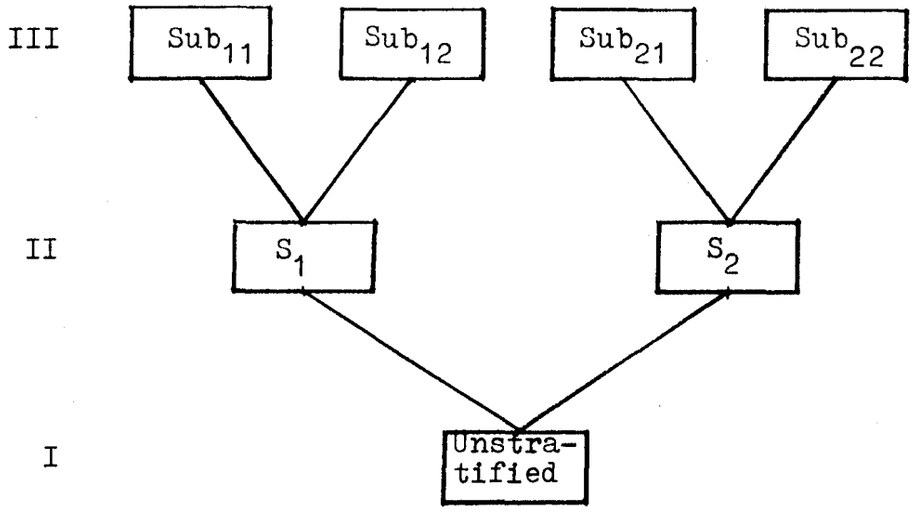

FIGURE 4-1. Hypothetical structure of a threefacet hierarchically stratified item universe.

Note.- I Zero-order stratification of item universe
II First-order stratification of item universe
III Second-order stratification of item universe

S = stratum Sub = substratum

and Rabinowitz and Eikeland (1964),implies a first-order stratification.A second-order stratification scheme means that items are grouped into substrata,which in turn are grouped into strata.This structure implies that substrata are nested within strata,and items are nested within substrata within strata.This is a typical hierarchical structure,also called a tree-structure.The principle of hierarchical stratification seems to justify calling the test of a second-order stratification a threefacet hierarchically stratified test.The facets are items,substrata,and strata. A tree-structure of a hierarchically stratified item universe of second order is presented

Insert FIGURE 4-1 about here

in FIGURE 4-1.It is the simplest conceivable structure of a balanced threefacet hierarchically stratified test.Burt (1954) uses another metaphor for the same structuring scheme.His simile is a sorting machine.

For the test constructor,if he is to adhere to a formalism in generating a hierarchically stratified test,the procedure should be to enter on a three-stage sampling plan.First,he should pick strata;second,substrata within strata;and third,he should pick items within substrata.Certainly,items have to be randomly sampled from the subpools of items.How the selection of conditions for strata and substrata is done,either by random sampling or by fixing on just those strata and substrata that are of substantive interest,or by a combination of random and fixed, is dependent on the test constructor's definition of the

universe of hierarchically stratified tests to which he wants to generalize.The generalizability problem can be formulated to imply how to find the expected correlation among tests that are constructed according to one of the particular sampling plans indicated above.

Building formal models often means idealizing conditions so much that there is a risk of finding no real world experiments fitting them.It is believed that one can find complex test designs in practical test construction approximately isomorphic to the hierarchically stratified test as here sketched,such that the model building is thought to be worth while as a means of being able to assess the properties of complex tests more adequately than before.The Primary Mental Abilities test and the California Test of Mental Maturity are examples of batteries that have been used for years,for which a proper theory has not been developed. Those tests,and several others can be mentioned,are fairly good fits to the formal models to be explicated in the subsequent discussion.

Stratum	1				2			
Substratum	1		2		3		4	
Item	1	2	3	4	5	6	7	8
P_1	X_{1111}	•	•	•	•	•	•	•
P_2	•	•	•	•	•	•	•	•
P_3	•	•	•	•	•	•	•	•
P_4	•	•	•	•	•	•	•	•
P_5	•	•	•	•	•	•	•	X_{5248}

FIGURE 5-1. A lay-out of a 5 x 2 x 2 x 2 hierarchically stratified test design.

Note. - P = persons, X_{5248} = item score for person 5 on item 8 within substratum 4 within stratum 2.

5. A model for the hierarchically stratified test.

After having administered a hierarchically stratified test to a sample of persons, test data at hand would be a system in which persons are crossed with strata, substrata, and items. Items are nested within substrata, which in turn are nested within strata. This particularly constructed multifacet test can most appropriately be called a doubly nested test design. The double nesting refers to items which are nested within substrata within strata, and also to substrata which are nested within strata. The present design is different from a design described by Stanley (1961) as doubly nested. We would prefer to describe Stanley's design as a design with two nested variables, which implies two separate hierarchical structures.

In order to make clear how the hierarchically stratified test design looks, FIGURE 5-1 presents an exemplification with 5 persons, 2 strata, 2 substrata within each of the strata, and 2 items within each of the substrata. The nesting of substrata and

Insert FIGURE 5-1 about here

items is indicated by consecutively numbering substrata from 1 to 4, and items from 1 to 8. Here four different substrata are represented in the design and eight different items. In a completely crossed multifacet test design of the same order, there would be only two substrata, appearing under both of the two strata; and only two items, appearing under each of the substrata.

It seems sound to believe that the Rabinowitz & Eikeland (1964) development of a model for the stratified test can naturally be extended to the hierarchically stratified test. A variance components model most probably can serve as the structural framework for the test theory development needed for solving the generalizability problem at issue concerning the test design of interest here.

In developing the mathematical model for the hierarchically stratified test an equal number of substrata within strata, and an equal number of items within substrata is assumed. This is done in order not to complicate the formulation unnecessarily in an effort to present the principle features of the model. Modifications of the formultions are possible in cases where an unequal number of substrata within strata and items within substrata is employed.

Let n be the number of persons, k number of items within each of the substrata, m number of substrata within strata, and r the number of strata. The symbols P, I, H, and S, or p, i, h, and s, are used for persons, items, substrata, and strata, respectively. Capital letters are used in talking about the variables; when subscripts are needed, small letters are used.

As a symbol for nesting, a colon will be used. Substrata nested within strata is symbolized H:S, or n:s. The double nesting of items will be written I:H:S, or i:h:s, to be read items within substrata within strata. After this, the hierarchically stratified test design can be symbolized as a PxSxH:SxI:H:S design. For a similar notational system, see Millman & Glass (1967) and Cronbach, Gleser, Nanda & Rajaratnam (1967).

TABLE 5-1

Structural models for mean squares in an analysis of variance table of the threefacet hierarchically stratified test

Source	SS	df	MS	Component structure
P	SS_p	(n-1)	MS_p	$\sigma^2_{res}+k\sigma^2_{ph:s}+km\sigma^2_{ps}+kmr\sigma^2_p$
S	SS_s	(r-1)	MS_s	$\sigma^2_{res}+k\sigma^2_{ph:s}+km\sigma^2_{ps}+n\sigma^2_{i:h:s}+kn\sigma^2_{h:s}+kmn\sigma^2_s$
H:S	$SS_{h:s}$	(m-1)r	$MS_{h:s}$	$\sigma^2_{res}+k\sigma^2_{ph:s}+n\sigma^2_{i:h:s}+kn\sigma^2_{h:s}$
I:H:S	$SS_{i:h:s}$	(k-1)mr	$MS_{i:h:s}$	$\sigma^2_{res}+n\sigma^2_{i:h:s}$
P by S	SS_{ps}	(n-1)(r-1)	MS_{ps}	$\sigma^2_{res}+k\sigma^2_{ph:s}+km\sigma^2_{ps}$
P by H:S	$SS_{ph:s}$	(n-1)(m-1)r	$MS_{ph:s}$	$\sigma^2_{res}+k\sigma^2_{ph:s}$
P by I:H:S	$SS_{pi:h:s}$	(n-1)(k-1)mr	$MS_{pi:h:s}$	σ^2_{res}

Note. - P = persons, S = strata, H = substrata, I = items, n = number of oersons, r = number of strata, m = number of substrata within strata, and k = number of items within substrata.

An analysis of variance table of data for the hierarchically stratified test design is presented in TABLE 5-1. In all, seven sources of variance can be identified in this design. Not all of them are of concern in a problem where individual differences are at issue. Only the sources of variance associated with persons are of interest. These are the person main effect and the

Insert TABLE 5-1 about here

three interactions of persons with strata, substrata, and items. In testing, variances associated with facet main effects are most often not of any substantive interest as these sources reflect more or less arbitrary variances, for example difficulty levels of items. These sources make no contribution to the individual differences variance, which is the information of particular interest.

Another restriction will be made. There are several sources in the present design that are of considerable interest regarding the information contained on individual differences. We shall in the following pay attention only to the source of variance called persons. This source reflects the variance of the sum score for persons across the three facets. Most often this is the test score used in practical testing. The test scores in the present design to be ignored in the following discussion will be two types of difference scores, contained in the persons by strata interaction and the persons by substrata within strata interaction. These scores are of crucial importance if one is concerned with differential abilities, i.e. to what extent the various strata and substrata are measuring different abili-

ties.There are specific generalizability problems connected with these scores which can be more conveniently discussed in another context (Eikeland 1972).

In approaching the generalizability of the test score,the expected mean square for persons $E(MS_p)$,expressed as a weighted sum of variance components, is the key for unlocking what may be called the deep structure of the test.While the observed mean square for persons is the manifest test score variance, it should be clear that the variance structure as represented by the components,in effect is a theory of how the person variance is generated and composed by the particular measuring operation used.The structure can not be observed.The structure is imposed on data.It is an inferred latent structure that is thought to be of considerable help in trying to interpret the test score in terms of different types of variance that go into it.The latent variance structure can tell to what degree the test score is influenced by a common trait running through all the items of the test;by less common traits,common to each of the strata;and by specific traits,common only to items within the substrata.Particularly,the generalizability problem at issue as regards the present test design makes it urgent to be explicit as to which of these more or less common traits are of enough substantive interest to be included in our definition of the universe score.

The definition of the universe score is automatically given by a specification of the universe of tests to which one wants to generalize.This specification determines how the sampling

plan for constructing tests belonging to this universe should be conceived.The latent structure model for persons in TABLE 5-1 is developed under the assumption that strata,substrata, and items are randomly sampled to be representative of universe of strata,subuniverses of substrata within the strata universe, and subuniverses of items within the substrata universes.This completely random model is undoubtedly the least likely to be of practical interest.However,in defining more realworld universes of tests,the completely random model is syntactically so important that one is convinced that in just this model the components as structural components are also meaningfully defined for model[a] that consider strata[s] fixed and substrata random,and a model that considers both strata and substrata fixed.

It should be noted that this way of defining components is contrary to how components are defined in traditional experimental design textbooks where classical analysis of variance, as aiming at probability statements,is exclusively emphasized. Here components are defined differently for different models.

The conventional way of defining components can in the present design be illustrated by considering strata fixed.According to rules of thumb in writing an analysis of variance table (Winer 1962,Millman & Glass 1967,Kirk 1968),a term (a weighted component) in the random model containing a subscript that is extra to the source of variance naming the row in the table, should be deleted if this extra subscript represents a fixed factor.Deleting the person by stratum component for the person row in TABLE 5-1,according to the conventional rule,means in effect that the value of the deleted component is included

the value of
in/the person component.However,the coefficient for the person component (the prescript) will not be affected by considering strata as fixed.The result is that the person component defined for the case of fixed strata and random substrata will increase compared to the person component defined for the random model.

Instead of the traditional procedure described,we shall keep the term for the person by stratum interaction ($km\sigma^2_{ps}$) intact even in the case of considering strata as fixed. The strata fixed assumption implies that the universe of generalization is defined such that the person by stratum component will be considered part of the universe score variance and not part of error score variance (Eikeland 1971).

The difference of procedure in defining components,as here recommended,makes no difference for the generalizabilit/y coefficient for the sum score,although it makes quite a difference if one is interested in examining the variance structure of the observed test score.Another difference will become apparent: When the generalizability problem concerns finding the generalizability of one average item,one is in considerable trouble employing the traditional way of defining components,while the reformulation as given her/e will be congenial with the test theory development to be discussed in the following sections.

By thus tying the definition of components to the completely random model,or more correctly in view of the subsequent discussion,to define the components according to the inferred structural model for the observed test score variance,the next step should be to define the universe score variance in

keeping with the sampling plan decided on in constructing the test. As more than one sampling plan is possible, there are also several ways of defining universe score variance.

The sampling plans concern the various ways of combining random and fixed facets. The most convenient point of departure for this procedure is the structural model for person variance,

$$E(MS_p) = \sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_p \qquad F5\text{-}1$$

According to rules of thumb for writing expected mean squares, regarding strata and substrata as fixed would imply deleting the weighted components for the PS and the PH:S interactions from the model. Our way of defining components rules that these components should be kept in the model but interpreted as belonging to universe score variance, because one does not intend to generalize beyond a universe that contains other strata and substrata than those chosen for the test.

When both universe score variance and observed test score variance are defined, the generalizability coefficient is given as the ratio of universe score variance to observed score variance. The sampling plan presently under consideration prescribes a fixed model for the threefacet hierarchically stratified test design. Therefore, this model will be designated 3F. In developing a series of generalizability coefficients they will be named alpha coefficients. By this the intention is to point to the generic nature of the alpha construct. It should not be restricted

to its original domain, the unstratified test (Cronbach 1951); it will prove fruitful to extend its domain to any kind of test design where generalizability coefficients are sought.

$$\text{alpha}_{3F} = \frac{k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}}{\sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}} \qquad \text{F5-2}$$

F5-2 is here given as a defining formula for alpha_{3F} in terms of weighted variance components. Shortly, a more convenient computing formula for F5-2 will be given.

There are two options for choosing a mixed generalizability model for the hierarchically stratified test, either random substrata and fixed strata, or fixed substrata and random strata. Both mixed models may be useful, but the fixed strata, random substrata model seems to be the most realistic one. Especially when one is generalizing to a content universe, it does not seem likely that he can reasonably fix on substrata within random strata. On the other hand, if substrata were chosen on the basis of format, then certainly it is reasonable to use fixed formats within each of randomly sampled strata.

Only the fixed strata, random substrata model will be presented as a mixed model in the following. The rule for deciding how to define universe score variance when strata are fixed and substrata random is to allocate the random PH:S component to error variance and the fixed PS component to universe score variance. This model we shall call 3M, and the generalizability coefficient is defined by,

$$\text{alpha}_{3M} = \frac{km\sigma^2_{ps} + kmr\sigma^2_{p}}{\sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}} \qquad \text{F5-3}$$

Lastly,a random model can be defined,regarding the conditions for all three facets picked according to a completely random sampling plan.In this case,the PS component in the observed test score variance is a random component and will be allocated to the error score variance,thus leaving only the P component for the universe score variance.The generalizability coefficient for this model,designated 3R,should read,

$$\text{alpha}_{3R} = \frac{kmr\sigma^2_{p}}{\sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}} \qquad \text{F5-4}$$

It is apparent that the defining formulas for the estimation of the generalizability for the thre models developed are unwieldy computing formulas as they presuppose that components have been estimated.Convenient computing formulas can easily be established in terms of observed mean squares,as can be seen from TABLE 5-1.

$$\text{alpha}_{3F} = \frac{MS_p - MS_{pi:h:s}}{MS_p} \qquad \text{F5-5}$$

$$\text{alpha}_{3M} = \frac{MS_p - MS_{ph:s}}{MS_p} \qquad \text{F5-6}$$

$$\text{alpha}_{3R} = \frac{MS_p - MS_{ps}}{MS_p} \qquad \text{F5-7}$$

Others have shown that the estimates obtained by alpha for the unstratified test are lower bound estimates for the defined generalizability of tests, the definition being the squared correlation of a random test score with the universe score (Rajaratnam, Cronbach & Gleser 1965, Novick & Lewis 1967, Lord & Novick 1968). It is here assumed that the same will hold for alphas developed for more complex test designs. This means that the alphas for the differently defined threefacet hierarchically stratified test models are considered lower bound estimates of the squared correlation of an observed test score with a particularly defined universe score within this test design.

A test theory development for a complex test design in terms of a formalized language like analysis of variance will most likely be difficult to grasp unless the reader is well versed in this particular language. In order to get a deeper understanding of the thinking going into this formalized procedure, first a numerical example, as simple as possible, will be presented, emphasizing meaning. Later alternative conceptual approaches to the generalizability problem will be made. Hopefully, these explorations will make clear how the structure of the generalizability theory is generated.

TABLE 6-1

Hypothetical data for a 5 x 2 x 2 x 2 hierarchically stratified test design

	S_1				S_2				
	H_1		H_2		H_3		H_4		
	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	Sum
P_1	5	5	5	4	4	4	4	5	36
P_2	2	3	3	2	5	4	4	3	26
P_3	4	3	4	4	3	2	2	3	25
P_4	2	3	4	3	1	2	1	2	18
P_5	1	2	2	2	3	2	1	1	14
Sum	14	16	18	15	16	14	12	14	119

TABLE 6-2

Analysis of variance of hypothetical test data

Source	SS	df	MS	Component structure	VC
P	35,60	4	8,900	$\sigma^2_{res} + 2\sigma^2_{ph:s} + 4\sigma^2_{ps} + 8\sigma^2_{p}$	0,756
S	1,23	1			
H:S	1,25	2			
I:H:S	2,10	4			
PS	11,40	4	2,850	$\sigma^2_{res} + 2\sigma^2_{ph:s} + 4\sigma^2_{ps}$	0,588
PH:S	4,00	8	0,500	$\sigma^2_{res} + 2\sigma^2_{ph:s}$	0,081
PI:H:S	5,40	16	0,338		0,338
Total	60,98	39			1,763

Note. - VC = variance components σ^2_{p}, σ^2_{ps}, $\sigma^2_{ph:s}$, and $\sigma^2_{pi:h:s}$

6.Numerical example.

The technique for estimating the generalizability of a hierarchically stratified test will be illustrated by hypothetical data containing 5 persons,2 strata,2 substrata within each of the strata,and 2 items within each of the substrata.Imagine that the test design is a kind of Wechsler scale.Let the two strata be[a/] verbal and a performance battery,with similarities and vocabulary as subtests within the verbal stratum,and picture completion and picture arrangement as subtests within the per-

Insert TABLE 6-1 about here

formance stratum.Within each of the subtests two items are picked.The data are presented in TABLE 6-1.It is the variance of the sum score for the 5 persons across all 8 items that is of most interest.The problem to solve is how to estimate the proportion of that variance that can be considered to be universe score variance.The basic data information for this purpose is contained in the intercorrelations among the 8 item columns.The analysis of variance result for the hypotethical test data is given in TABLE 6-2.Only those mean squares are

Insert TABLE 6-2 about here

presented that are of concern for the generalizability problem. These are the mean squares for the sources of variance which contribute to the test score variance.There are four sources

determining this variance,all of them having a P in the row symbol.The PS interaction assesses the lack of convergence between the two subscores for strata.In a way,it is the complement of a correlation measure.Thus,the more interaction,the less correlation between the two subscores.The PH:S interaction and the PI:H:S interaction can be interpreted the same way.The first interaction term is concerned with the discrepancy between the substrata scores within strata,the second with the discrepancy between item scores within the substrata.What is important to realize intuitively is that these interaction terms are influencing the test score variance.The more interaction,the less interindividual differences.Thus,by manipulating the data matrix by deliberately changing the correlation either between items within substrata,between substrata within strata,and between strata,the test score variance will be changed.

An insight into the mechanism at work here makes it somewhat more understandable why the interaction components should go into the model for the P variance.When the equations for the various components going into the observed test score variance are solved for,that variance can be written as a sum of weighted components according to the model for P in TABLE 6-2,

$$MS_p = 0,338+2.0,081+2.2.\ 0,588+2.2.2.\ 0,756 = 8,900$$
$$= 0,338+\ 0,162+\ \ 2,352+\ \ 6,048 = 8,900$$
$$1,000 = 0,038 + 0,018 + 0,264 + 0,680$$

In setting the P variance like 1,000,the contribution to total test score variance made by the weighted components can be

read as the proportion of variance accounted for.This is the structure of the total variance of individual differences.According to this 68 per cent of the variance is explained by a common trait running through all the test items,irrespective of whether they are verbal or performance items,similarities, vocabulary,picture completion,or picture arrangement items. About 26 per cent of the variance is accounted for by the fact that verbal and performance are tapping different traits.This is a reflection of the PS interaction.The contribution to variance made by the PH:S interaction is negligible,meaning that the substrata within strata are so highly correlated that they may be said to measure the same trait within their respective strata. The specificity component's contribution to variance is also negligible.This should be interpreted to mean that items within substrata to a very great extent are measuring the same thing. The structural properties of the test score variance as here presented are crucial for a meaningful interpretation of the battery score.

From the structure of the test score variance the generalizability estimates for the three models are found by allocating the components to universe score variance or to error score variance.How this allocation should be / done is determined by the definition of the universe of generalization.

In the present case it is reasonable to regard both strata and substrata as fixed.Probably the verbal and the performance domains as strata exhaust the universe of strata to which one wants to generalize.Also,the generalization intended is res-

stricted to the similarities and vocabulary tests within the verbal stratum and to the picture completion and picture arrangement tests within the performance stratum. In other words, if a parallel battery was to be constructed, a new sampling of items had to be undertaken within the same substrata within the same strata. For this fixed model, both components involving strata and substrata are included in the universe score variance together with the common component, the P component. Therefore, the generalizability coefficient for this model will be,

$$\text{alpha}_{3F} = \frac{0,162+2,352+6,048}{0,338+0,162+2,352+6,048} = 0,962$$

$$= \frac{MS_p - MS_{pi:h:s}}{MS_p} = \frac{8,900-0,338}{8,900} = 0,962$$

When verbal and performance are regarded as fixed, i.e. not sampled, and similarities and vocabulary, picture completion and picture arrangement as randomly sampled within verbal and performance strata, respectively, from subuniverses of tests, a mixed model is appropriate. Because substrata are regarded as sapled[m/], the random PH component is allocated to error variance, and the generalizability estimate will be,

$$\text{alpha}_{3M} = \frac{2,352+6,048}{0,338+0,162+2,352+6,048} = 0,944$$

$$= \frac{MS_p - MS_{ph:s}}{MS_p} = \frac{8,900-0,500}{8,900} = 0,944$$

In considering both strata and substrata as random, the least likely case for this Wechsler-like test battery, the PH:S and the PS components will as random components be ascribed to the error variance term. Only the common to all items component,

the P component,is allocated to universe score variance.Thus the proportion of universe score variance will be,

$$\text{alpha}_{3R} = \frac{6,048}{0,338+0,162+2,352+6,048} = 0,680$$

$$= \frac{MS_p - MS_{ps}}{MS_p} = \frac{8,900-2,850}{8,900} = 0,680$$

A most meaningful interpretation of the three alpha coefficients as obtained from the hypothetical test data, is that they are the estimated correlation of the test scores at hand with another set of test scores obtained from another test battery constructed according to the specific sampling plans defined for each of the models.It is also meaningful to see how the generalizability estimates are related to the proportional composition of the test score variance.As a matter of fact,the three estimates can be taken from that structure by simply adding components.

7. A covariance approach to the generalizability of hierarchically stratified tests.

The drawback by following a more or less rule of thumb procedure in developing the models for the generalizability of hierarchically stratified tests is apparent. By adhering to rules one can generate correct formulas, but no deep understanding necessarily follows. Particularly, the introduction of fixed and random facets in more complex test designs makes it difficult to see how the various generalizability formulas obtain under different sampling plans.

Fortunately, there are alternative approaches to generalizability estimates and the structural features of the generalizability theory that facilitate a more readily understandable rationale for how to obtain the generalizability coefficients presented in the discussion of the analysis of variance approach. Seemingly, the covariance procedure to be dealt with in the following is something quite different from the analysis of variance approach. Yet, as will most likely become clear in proceeding along a covariance line of thinking, there is not at all any difference between the two procedures. However, the covariance approach seems to be much more conducive to a fundamental understanding of what kind of structure one is imposing on data in order to arrive at the specific formulas for the different models.

As mentioned previously, the generalizability coefficient can also be defined as the expected correlation between random parallel tests. The ratio of the expected covariance between

TABLE 7-1

Variance-covariance matrix of hypothetical test data

			S_1				S_2			
			H_1		H_2		H_3		H_4	
			I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
S_1	H_1	I_1	2,70	1,55	5,00		12,10			
		I_2	1,55	1,20						
	H_2	I_3	5,00		1,30	1,00				
		I_4			1,00	1,00				
S_2	H_3	I_5	12,10				2,20	1,30	5,75	
		I_6					1,30	1,20		
	H_4	I_7					5,75		2,30	1,85
		I_8							1,85	2,20

Note. - $\bar{c}_{bb} = 0,756$, $\bar{c}_{bw} = 1,344$, $\bar{c}_{ww} = 1,425$, and $\bar{v}_i = 1,763$

two such random parallel tests,in our case two tests being constructed according to the same complex sampling plan,to the expected test variance (i.e.the product of the two tests' standard deviations) is the wanted correlation.Applying a covariance of sums rationale will serve our purpose for estimating the expected covariance between random parallel tests under different sampling plans.

First,consider the observed variance-covariance matrix of one test constructed according to the hierarchically stratified test design.For convenience,let the variance-covariance matrix be illustrated by the one generated from the hypothetical test data in TABLE 6-1.The variance-covariance matrix con-

Insert TABLE 7-1 about here

tains three distinguishable types of covariance among items. For the subsequent discussion it is important for the reader to be able to see this distinction clearly.One type of covariance is a monosubstratum-monostratum interitem covariance.(For a similar terminology,see Campbell & Fiske 1959.)This is a covariance among items within substrata within strata.Another type is a heterosubstratum-monostratum inter-item covariance. It is a covariance among items between substrata (among items from different substrata) within strata.Lastly,the third type of covariance is a heterosubstratum-heterostratum interitem covariance.It is a covariance among items between substrata between strata (among items from different substrata and dif-

TABLE 7-2

Covariance matrix for two random, hierarchically stratified tests. Fixed model.

Test 2

Test 1				S_1		S_2	
				H_1	H_2	H_3	H_4
				I_9 I_{10}	I_{11} I_{12}	I_{13} I_{14}	I_{15} I_{16}
S_1	H_1	I_1 I_2		c_{ww}	c_{bw}	c_{bb}	
	H_2	I_3 I_4		c_{bw}	c_{ww}		
S_2	H_3	I_5 I_6		c_{bb}		c_{ww}	c_{bw}
	H_4	I_7 I_8				c_{bw}	c_{ww}

Note.-Fixed model: strata fixed, substrata fixed, items random.

ferent strata). As a shorthand the three types of covariance will be called the covariance within-within, or ww; the covariance between-within, or bw; and the covariance between-between, or bb.

It is reasonable to believe, if a rational stratification plan is followed, that the average interitem covariance within-within is larger than the average interitem covariance between-within, which in turn is larger than the average interitem covariance between-between. If items belonging to different substrata and different strata are tapping the same trait, generally speaking, the three types of average interitem covariance are expected to be equal. In TABLE 7-1 the average covariances are, $\bar{c}_{bb} = 0{,}756$; $\bar{c}_{bw} = 1{,}344$; and $\bar{c}_{ww} = 1{,}425$. It should be noted that the covariance between-within is pooled for the four submatrices where this type of covariance is found, i.e. the covariance between substratum 1 and substratum 2 within stratum 1 is added to the covariance between substratum 3 and substratum 4 within stratum 2, and then averaged. The same pooling procedure is performed for the covariances within-within.

Next, let us construct a hypothetical covariance matrix between two random parallel hierarchically stratified tests, assuming a fixed model. Under this assumption both substrata and strata are fixed, implying that the same substrata and strata are used for the two tests. Under this particular sampling plan all of the three types of covariance defined above are established in the covariance matrix, as is hopefully evident from TABLE 7-2. In

Insert TABLE 7-2 about here

that table the same strata and substrata that appear in test 1 reappear in test 2,whereas a new sampling of items has been undertaken for test 2.In what is here called the fixed model, i.e.strata and substrata fixed,it should be quite clear that items within substrata within strata are still assumed to be randomly sampled from subpools of items.Therefore,it is legitimate to regard the tests in TABLE 7-2 to be random parallel. They are random parallel, fixed hierarchically stratified tests.

The expected covariance between two random parallel tests of the fixed model will be the sum of the different types of covariance in the matrix.An expected correlation between the tests can be defined by using the expectations for the different interitem covariances as a numerator and the product of two expected test standard deviations,i.e. the expected test variance,as a denominator.This definition of the correlation between two random parallel fixed hierarchically stratified tests is also the definition of coefficient alpha.

Let $\sigma_{ij}ww$, $\sigma_{ij}bw$, $\sigma_{ij}bb$,where $i \neq j$,symbolize the three expectations of the differently defined interitem covariances.Further, to make the formulations more general,let $\underline{k}$ be the number of items within substrata,$\underline{m}$ the number of substrata within strata, and $\underline{r}$ the number of strata.In such a matrix of covariances, there will be k^2mr covariances $\underline{ww}$,$k^2m(m-1)r$ covariances $\underline{bw}$, and $k^2m^2r(r-1)$ covariances $\underline{bb}$.The expected correlation between two random parallel tests of the threefacet hierarchically stratified test design,fixed model,can be defined

$$\text{alpha}_{3F} = \frac{k^2 m r \sigma_{ij}ww + k^2 m(m-1) r \sigma_{ij}bw + k^2 m^2 r(r-1) \sigma_{ij}bb}{E(V)} \qquad F7\text{-}1$$

There should be no problem estimating alpha_{3F} with data available from one test only. This can be done by using the average interitem covariances in the test as estimates for the three previously defined covariances. An estimation form of the defining formula F7-1 can therefore be written

$$\text{alpha}_{3F} = \frac{k^2 m r \frac{\Sigma C_{ij}ww}{k(k-1)mr} + k^2 m(m-1) r \frac{\Sigma C_{ij}bw}{k^2 m(m-1) r} + k^2 m^2 r(r-1) \frac{\Sigma C_{ij}bb}{k^2 m^2 r(r-1)}}{V} \qquad F7\text{-}2$$

By a little algebra, F7-2 reduces to

$$\text{alpha}_{3F} = \frac{(\frac{k}{k-1})\Sigma C_{ij}ww + \Sigma C_{ij}bw + \Sigma C_{ij}bb}{V} \qquad F7\text{-}3$$

Inserting the covariances and the test variance from the hypothetical test data in TABLE 7-1, the following result is obtained,

$$\text{alpha}_{3F} = \frac{2 \cdot 11{,}40 + 21{,}50 + 24{,}20}{71{,}20} = 0{,}962$$

It is important to note that this result is identical to that obtained in the analysis of variance approach. Thus F7-3 is equal to F5-5, although they are seemingly quite different formulas. The relationship between the two approaches will be discussed in a subsequent section.

TABLE 7-3

Covariance matrix for two random, hierarchically stratified tests. Mixed model.

				Test 2			
				S_1		S_2	
				H_5	H_6	H_7	H_8
				I_9 I_{10}	I_{11} I_{12}	I_{13} I_{14}	I_{15} I_{16}
Test 1	S_1	H_1	I_1 I_2	c_{bw}	c_{bw}	c_{bb}	
		H_2	I_3 I_4	c_{bw}	c_{bw}		
	S_2	H_3	I_5 I_6	c_{bb}		c_{bw}	c_{bw}
		H_4	I_7 I_8			c_{bw}	c_{bw}

Note.-Mixed model: strata fixed, substrata random. items random.

Next, consider a hypothetical covariance matrix between two hierarchically stratified tests, assuming a mixed model. Under this substrata random, strata fixed assumption, the covariance matrix will be somewhat different from the covariance matrix under the fixed model in TABLE 7-2. What is noteworthy about this

Insert TABLE 7-3 about here

modified covariance matrix is that no substratum reappears in the first test in the second test. Thus no covariance can be established among items from the same substrata. This is a result of the random sampling of substrata. Consequently, in the covariance matrix of this particular model there will be no covariance of the ww type. Under the mixed model only two types of covariance can be established, the bw and the bb type. What is interesting to note is that the k^2mr covariances ww in the fixed model have to be substituted by the same number of covariances bw.

By finding the correct number of the interitem covariances of the bw and the bb types, the ratio of common variance to test variance, or the expected correlation between tests of the mixed model can be defined. In changing from the fixed model assumption to the mixed model assumption it should be noted that the expected test variance does not change.

$$\text{alpha}_{3M} = \frac{(k^2mr + k^2m(m-1)r)\sigma_{ij}bw + k^2m^2r(r-1)\sigma_{ij}bb}{E(V)}$$

$$= \frac{k^2m^2r\sigma_{ij}bw + k^2m^2r(r-1)\sigma_{ij}bb}{E(V)} \qquad \text{F7-4}$$

The estimation form of F7-4 can be obtained by substituting average interitem covariances from one test for the expectations in the defining formula,

$$\text{alpha}_{3M} = \frac{k^2m^2r\,\dfrac{\Sigma C_{ij}bw}{k^2m(m-1)r} + k^2m^2r(r-1)\dfrac{\Sigma C_{ij}bb}{k^2m^2r(r-1)}}{V} \qquad \text{F7-5}$$

By a little algebra F7-5 reduces to

$$\text{alpha}_{3M} = \frac{(\frac{m}{m-1})\Sigma C_{ij}bw + \Sigma C_{ij}bb}{V} \qquad \text{F7-6}$$

Inserting the covariances and the variance from the hypothetical test data in TABLE 7-1 in F7-6, the following alpha coefficient is obtained,

$$\text{alpha}_{3M} = \frac{2 \cdot 21{,}50 + 24{,}20}{71{,}20} = 0{,}944$$

Again, the covariance approach gives the same result as the analysis of variance approach. The equivalence of F5-6 to F7-6 should be noted.

Lastly, the random model will be considered in terms of the covariance approach. In the random model both substrata and strata are assumed to be randomly sampled. The hypothetical covariance matrix between two hierarchically stratified tests constructed according to the same sampling plan defined for the random model, will be different from the two preceding covariance matrices under the fixed and mixed models, in TABLE 7-2 and TABLE 7-3, respectively.

TABLE 7-4

Covariance matrix for two random, hierarchically stratified tests. Random model.

				Test 2			
				S_3		S_4	
				H_5	H_6	H_7	H_8
				I_9 I_{10}	I_{11} I_{12}	I_{13} I_{14}	I_{15} I_{16}
Test 1	S_1	H_1	I_1 I_2	c_{bb}	c_{bb}	c_{bb}	
		H_2	I_3 I_4	c_{bb}	c_{bb}		
	S_2	H_3	I_5 I_6	c_{bb}		c_{bb}	c_{bb}
		H_4	I_7 I_8			c_{bb}	c_{bb}

Note.- Random model: strata random, substrata random, items random.

What is different in the covariance matrix for the random model as presented in TABLE 7-4 compared to the covariance matrix for the mixed model in TABLE 7-3, is that the same stratum will not appear two times in the matrix for the random model. While the strata fixed assumption in the mixed model implied that the same strata would be used for all random parallel tests, the strata random assumption in the random model prescribes a new sampling of strata for every new test to be constructed.

Insert TABLE 7-4 about here

Therefore, in the covariance matrix under consideration now, neither the interitem covariance of the ww type, nor the covariance of the bw type can be established. All the interitem covariances are of one type, namely the bb type. They will be covariances between different items from different substrata and from different strata. Consequently, the expected correlation between random parallel tests of the random model will have a relatively simple form,

$$\text{alpha}_{3R} = \frac{k^2 m^2 r^2 \sigma_{ij}\text{bb}}{E(V)} \qquad \text{F7-7}$$

The estimation form of F7-7 can be obtained by substituting the average interitem covariance bb for the covariance parameter and taking the observed test variance as an estimate of E(V).

$$\text{alpha}_{3..} = \frac{k^2m^2r^2\frac{\Sigma C_{ij}bb}{k^2m^2r(r-1)}}{V} = \frac{(\frac{r}{r-1})\Sigma C_{ij}bb}{V} \qquad \text{F7-8}$$

An interesting structural similarity between F7-8 and traditional coefficient alpha will become apparant when $\Sigma C_{ij}bb$ is substituted for $V - \Sigma V_s$, i.e. the total test variance minus the sum of the strata variances,

$$\text{alpha}_{3R} = (\frac{r}{r-1})(V - \frac{\Sigma V_s}{V}) = (\frac{r}{r-1})(1 - \frac{\Sigma V_s}{V}) \qquad \text{F7-9}$$

Evidently, under the random assumption model, strata are regarded as items in a homogeneous test. The alpha_{3R} is concerned with the internal consistency of randomly sampled strata.

Inserting the covariance and the variance from the hypothetical test data in TABLE 7-1 in F7-8, the following alpha coefficient is obtained,

$$\text{alpha}_{3R} = \frac{2 \cdot 24,20}{71,20} = 0,680$$

Exactly the same result is obtained here by the covariance approach as was obtained by the analysis of variance approach.

The most important feature to pay attention to in the covariance approach is the rationale established for defining the different sum/s of covariances to go into the alpha formula for the various models. It should be understood how the different covariances obtain under the three specifications made for the sampling plan for each of the models.

The convergence of the analysis of variance approach and the covariance approach to the generalizability of a hierarchically stratified test as established in terms of exactly the same results, is at this moment not easily explained by reference to an underlying, more basic, common conceptual framework. This fundamental model will hopefully become clearer as we proceed to another way of looking at the structure of the generalizability problem involved in the hierarchically stratified test.

8. Generalizability estimates in terms of the expected variance-covariance matrix of a random parallel hierarchically stratified test.

The covariance approach to the generalizability of hierarchically stratified tests estimates the expected covariance between two random parallel tests constructed according to a particularly defined sampling plan,reflecting the universe of tests to which one wants to generalize.The three categories of covariance defined above are expected observed covariances in the universe of tests.The covariance structures conceived in TABLE 7-2,TABLE 7-3,and TABLE 7-4 are manifest covariance structures for the different models of the hierarchically stratified test design.

Instead of hypothetically correlating random parallel tests of the design at issue,as was done above,one can think of an alternative approach that is concerned with an inferred variance structure of one /random parallel hierarchically stratified test. The intuitive logic of this approach has been described by Eikeland (1970) for the random parallel,unstratified test.The same logic seems also to be sound for stratified tests.In the following this rationale will be extended,first,to the twofacet hierarchical test;next,to the threefacet hierarchically stratified test.

As regards the unstratified test,one can conceive of a latent structure of the variance-covariance matrix of a random parallel test consisting of two components,a covariance component and a variance component.In the universe of items this covariance

TABLE 8-1

Latent structure of the variance-covariance matrix for a 4-items unstratified test

	I_1	I_2	I_3	I_4
I_1	$\sigma^2_{pi}+\sigma^2_p$	σ^2_p	σ^2_p	σ^2_p
I_2	σ^2_p	$\sigma^2_{pi}+\sigma^2_p$	σ^2_p	σ^2_p
I_3	σ^2_p	σ^2_p	$\sigma^2_{pi}+\sigma^2_p$	σ^2_p
I_4	σ^2_p	σ^2_p	σ^2_p	$\sigma^2_{pi}+\sigma^2_p$

component is the common variance shared by items in the defined universe.It is an expected value.Under certain assumptions the observed covariance among items is equal to the expected universe score variance.When all items are pooled,the observed-score variance equals the universe score variance plus error score variance (see Lord & Novick 1968,Chapter 8).The inference made in constructing the latent variance-covariance matrix for a random parallel composite is to impose on the expected item variance the covariance component plus a residual component,the error component,which is the difference between the expected item variance and the imposed covariance component.Thus the

Insert TABLE 8-1 about here

latent variance-covariance matrix of a random parallel unstratified test will be conceptually composed of k^2 covariance components and k error components,or residuals,as seen from TABLE 8-1.

The generalizability estimate for the test is the ratio of the universe score variance,the sum of the covariance components,to the test variance which is the sum of all components in the matrix.On the basis of this expected variance-covariance matrix coefficient alpha can be given a fairly well known form,

$$\text{alpha} = \frac{k^2\sigma_{ij}}{k\sigma_e^2 + k^2\sigma_{ij}} = \frac{k\sigma_{ij}}{\sigma_e^2 + k\sigma_{ij}} = \frac{k\sigma_p^2}{\sigma_{pi}^2 + k\sigma_p^2} \qquad \text{F8-1}$$

Eikeland (1970) has shown that the reconstruction of the gene-

ralizability for the unstratified composite in terms of this intuitive logic is identical to the formal analysis of variance approach as first developed by Hoyt (1941).What is called the universe score component,or true score component, in the analysis of variance approach (σ_p^2), is just another name for the expected covariance among items (σ_{ij}). This identity, $\sigma_{ij} = \sigma_p^2$, explains the interchangeability of formulas in F8-1 and the particular symbols used in TABLE 8-1,where traditional analysis of variance symbols are adhered to.F8-1 should make it clear that the more abstract,and for many a somewhat obscure,analysis of variance approach can be conceived in terms of a latent variance-covariance matrix of items.

The intuitive logic as developed for the latent variance-covariance matrix of the unstratifie test will next be extended to the twofacet stratified test,in order to make a still further extension to the hierarchically stratified test more easy to grasp. The formal approach to the generalizability of the stratified,or hierarchical,test design can be found in Rabinowitz & Eikeland (1964) and Rajaratnam,Cronbach & Gleser (1965).

In a test constructed according to the twofacet,hierarchical design with items nested within strata,two types of covariance among items are conceivable. First,a covariance among items within strata,called the within covariance, is defined, $\sigma_{ij}w$. Next,a covariance among items between strata,called the between covariance,can be defined, $\sigma_{ij}b$.

The inferred variance structure of the test scores revealed by the construction of a latent variance-covariance matrix for the twofacet,hierarchical test will be somewhat more complex than for the unstratified test.The covariance among items between strata,the between covariance,represents the common variance across strata. These covariances reflect the most general of the traits tapped by a multifacet measuring procedure. The covariance between strata accounts for the common-to-all-items variance,regardless of strata. This component of the variance structure is the $\sigma_{ij}b$ covariance,as defined above. In keeping with what was found for the unstratified test,the covariance component,$\sigma_{ij}b$, for the hierarchical test,is equal to the person component,σ_p^2,as defined in the analysis of variance approach.This identity, $\sigma_{ij}b = \sigma_p^2$,should be kept in mind for the subsequent discussion.

In the stratified universe of generalization,the covariance $\sigma_{ij}w$, among items within strata,is construed to be composed of two covariance components.First,the common-to-all-items variance component,σ_p^2,or $\sigma_{ij}b$, is naturally defined into the covariance within. Second, in addition to the more general trait measured by σ_p^2,the covariance within strata is thought to measure also a trait that is specific for each of the strata.This less generally conceived component of the covariance structure,reflects the common-to-groups-of-items variance,the groups being defined by the stratification plan for the item universe.While the common-to-all-items component is dependent upon the inter-individual differences in the sum scores across all strata when allowance is made for the less general effects,the common-to--groups-of-items component reflects the interaction between

TABLE 8-2

Latent structure of the variance-covariance matrix for a 2x2 hierarchical test

		S_1		S_2	
		I_1	I_2	I_3	I_4
S_1	I_1	$\sigma^2_{res}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p
	I_2	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{res}+\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p
S_2	I_3	σ^2_p	σ^2_p	$\sigma^2_{res}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$
	I_4	σ^2_p	σ^2_p		$\sigma^2_{res}+\sigma^2_{ps}+\sigma^2_p$

Note.- $\sigma^2_{res} = \sigma^2_{pi:s}$

persons and strata. This means that profiles of strata scores are different for different persons. It is therefore reasonable that the common-to-grous-of-items component has been denoted as an interaction component, as is customary in the analysis of variance approach. In the construction of a conceptual framework for an inferred, latent variance structure of hierarchical test scores, the expected covariance among items within strata is conceived to be composed of two additive covariance components, $\sigma_{ij}b$ and $\sigma_{ij}w - \sigma_{ij}b$. The first can also be designated σ^2_p, the second will be called σ^2_{ps}. Thus, the structure of the expected covariance among items within strata can be written, $\sigma_{ij}w = \sigma^2_{ps} + \sigma^2_p$.

The expected item variance from such a stratified universe can now be conceived of as consisting of the two covariance components defined above, and a residual component, σ^2_{res}. This component is technically an interaction component. It is the person by item interaction within strata. Thus the residual component will also be called $\sigma^2_{pi:s}$. By now, having established the conceptual frame-

Insert TABLE 8-2 about here

work for a latent variance structure of the hierarchical test score, the inferred structure of the variance-covariance matrix of these scores can be seen from TABLE 8-2. For convenience, TABLE 8-2 is based on a 2-strata-by-2-items design. In generalizing to a twofacet, hierarchical test with $\underline{r}$ strata and $\underline{k}$ items within each stratum, the sum of such a latent variance-covariance matrix will be a sum of weighted components. In a $\underline{kr}$ x $\underline{kr}$ matrix there will be $\underline{kr}$ residual components, $\underline{k^2r}$

interaction components, and $\underline{k}^2\underline{r}^2$ common components. Thus the expected test score variance as an inferred structure can be written,

$$E(V) = kr\sigma^2_{res} + k^2r\sigma^2_{ps} + k^2r^2\sigma^2_p \qquad F8\text{-}2$$

The generalizability problem at issue, having established F8-2, is to find the ratio of universe score variance to total test score variance. In order to do this one has to define which of the covariance components should go into the universe score variance. This is a question of deciding on the universe of generalization of substantive interest for a particular testing purpose. In the present case, there are two possibilities of defining a universe of generalization, either to regard both components, σ^2_{ps} and σ^2_p, as belonging to the universe, or only the common component, the σ^2_p component.

By regarding strata as fixed, one is interested in generalizing to just those strata which are found in the test at hand. Therefore, it is reasonable to consider the within covariance as replicable covariance in that the same strata will reappear in the construction of another random parallel test. Consequently, for the fixed model, the universe score variance should include both covariance terms. This conclusion can be made still more convincing by referring to the logic established in TABLE 7-2, TABLE 7-3, and TABLE 7-4. While those tables illustrate the threefacet hierarchically stratified test design, one could by the same reasoning construct covariance matrices for random parallel hierarchical tests, showing that the present conclusion is correct.

According to the reasoning established for the definition of universe score variance for the fixed model, the generalizability estimate should be,

$$\text{alpha}_{2F} = \frac{k^2 r \sigma^2_{ps} + k^2 r^2 \sigma^2_p}{k r \sigma^2_{res} + k^2 r \sigma^2_{ps} + k^2 r^2 \sigma^2_p}$$

$$= \frac{k \sigma^2_{ps} + k r \sigma^2_p}{\sigma^2_{res} + k \sigma^2_{ps} + k r \sigma^2_p} \qquad \text{F8-3}$$

The eventual form of F8-3 is identical to the reliability form for the fixed model as developed by Rabinowitz & Eikeland (1964) for the same test design by an analysis of variance approach.

At this point it should be noted that the test variances as estimated by the covariance approach for the unstratified test and the hierarchical test, are different from the test variances as estimated by the MS_p in the analysis of variance approach. However, they bear a functional relationship to each other. While the E(V) for the unstratified test is the sum of the components in TABLE 8-1, $k\sigma^2_{pi} + k^2\sigma^2_p$, the $E(MS_p)$ for the same test design in an analysis of variance approach is $\sigma^2_{pi} + k\sigma^2_p$. Correspondingly, for the hierarchical test the E(V) as seen from TABLE 8-2 is $kr\sigma^2_{pi:s} + k^2 r \sigma^2_{ps} + k^2 r^2 \sigma^2_p$, and the $E(MS_p)$ in an analysis of variance approach would be $\sigma^2_{pi:s} + k\sigma^2_{ps} + kr\sigma^2_p$. The relation ship between E(V) and $E(MS_p)$ obviously is the following, $kE(MS_p) = E(V)$ for the unstratified test, and $krE(MS_p) = E(V)$ for the hierarchical test. Actually, the difference noted can be seen as a difference in the conventions established in estimating the test score variance. According to these conventions

the sum score variance of an unstratified two-item test would be computed the following ways,

$$V_t = 1/(N-1)\ \Sigma(x_1 + x_2)^2 = v_1 + v_2 + 2cov_{12} \qquad F8-4$$

$$MS_p = 1/(N-1)\ 2\ \Sigma\left(\frac{x_1 + x_2}{2}\right)^2 = (1/2)(v_1 + v_2 + 2cov_{12}) \qquad F8-5$$

What is shown in F8-4 and F8-5 can easily be generalized to $\underline{k}$ items for an unstratified test and to $\underline{kr}$ items for a hierarchical test.

The relationship established between E(V) and E(MS_p) clearly implies that the basic reasoning in the analysis of variance approach is concerned with a latent variance-covariance matrix as developed above.However,this convergence of the analysis of variance approach on the deep covariance structure conceived in the present monograph has never been explicated in the literature,as far as the author knows.

Returning now to the generalizability estimates for the two-facet hierarchical test,a random model regards strata as randomly sampled from a pool of defined strata.Compared to the fixed model developed above,one has to reinterpret the universe score variance such as to match a differently conceived universe of generalization.In the case of the random model one intends to generalize to a universe of tests where there can be no room for resampling of items within the same strata.As a matter of fact,in the covariance matrix of two random parallel hierarchical tests,constructed in accordance with the prescrip-

tions of the random model, there will be no covariance among items within strata, only a covariance among items between strata. Thus the component for the covariance among items within strata in the variance-covariance matrix of such a test has to be reinterpreted as belonging to the error score variance, because it is not a replicable variance component. The thinking going into this conclusion may become more convincingly clear if the reader can be able to modify TABLE 7-4 to fit the random model of the twofacet, hierarchical test design.

According to the rationale developed for the random model, the alpha coefficient as a generalizability estimate should be,

$$\text{alpha}_{2R} = \frac{k^2 r^2 \sigma_p^2}{kr\sigma_{res}^2 + k^2 r\sigma_{ps}^2 + k^2 r^2 \sigma_p^2}$$

$$= \frac{kr\sigma_p^2}{\sigma_{res}^2 + k\sigma_{ps}^2 + kr\sigma_p^2} \qquad \text{F8-6}$$

The defintion of alpha_{2R} reached in F8-6 is equal to the definition of the reliability for the random model of the hierarchical test design as developed by Rabinowitz & Eikeland (1964) in their analysis of variance approach. Again, this result is a new corroboration of the convergence of the covariance approach and the analysis of variance approach.

The generalizability estimates developed so far for the unstratified test and the hierarchical test in terms of the latent structure of the expected variance-covariance matrices should facilitate the next extension of the conceptual framework. In going to the threefacet hierarchically stratified test design the structural conception of the complex test score will be further complicated by another covariance component compared to the twofacet case just considered. The previous discussion of the threefacet test in Section 7 made it clear that one can define into the variance-covariance matrix three types of covariance: (1) A covariance among items between strata between substrata, called between-between, or bb. (2) A covariance among items between substrata within strata, called between-within, or bw. (3) A covariance among items within strata within substrata, called within-within. The theoretical construction that lies ahead for the threefacet test design is to incorporate a third covariance component into the inferred structure of the variance-covariance matrix of the hierarchically stratified test.

The most general trait measured by the test battery of this design is reflected in the covariance between-between, since this is a covariance among items that are maximally dissimilar. It is the covariance among different items from different substrata and from different strata. This common trait is thought to run through all of the items, so that the component due to the common factor is built into the covariance between-within and also into the covariance within-within. Lastly, because the items belong to a defined family of items, it is reasonable to impose the bb component also on the item variances.

TABLE 8-3

Latent structure of the variance-covariance matrix for a 2x2x2 hierarchically stratified test

			S_1				S_2			
			H_1		H_2		H_3		H_4	
			I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
S_1	H_1	I_1	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p	σ^2_p	σ^2_p
		I_2	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p	σ^2_p	σ^2_p
	H_2	I_3	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p	σ^2_p	σ^2_p
		I_4	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	σ^2_p	σ^2_p	σ^2_p	σ^2_p
S_2	H_3	I_5	σ^2_p	σ^2_p	σ^2_p	σ^2_p	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$
		I_6	σ^2_p	σ^2_p	σ^2_p	σ^2_p	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$
	H_4	I_7	σ^2_p	σ^2_p	σ^2_p	σ^2_p	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$
			σ^2_p	σ^2_p	σ^2_p	σ^2_p	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$	$\sigma^2_r+\sigma^2_{ph:s}+\sigma^2_{ps}+\sigma^2_p$

Note σ^2 σ^2 σ^2

Less general traits can be assumed to be measured by the covariance among items between substrata within strata. This type of covariance should reflect the common-to-each-stratum variance in addition to the common-to-all-items variance which has already been imposed on it. Consequently, the structure of the covariance between-within can be conceived as a sum of the common component and a stratum-specific component. This more specific component reflects the person by stratum interaction and will be called σ^2_{ps}. Thus one defines $\sigma_{ij}bw = \sigma^2_{ps} + \sigma^2_{p}$.

Still less general traits can be assumed to be measured by the covariance among items within strata within substrata. This type of covariance should reflect the common-to-each-substratum variance in addition to the common-to-all-items variance, σ^2_{p}, already imposed. However, also the common-to-each-stratum component should be imposed on the within-within covariance, since what is common-to-each-stratum variance must also be common to the substrata within each stratum. It seems therefore reasonable to define a covariance component that accounts for the specific traits tied to the different substrata. This component will be the residual within-within covariance when the σ^2_{p} and the σ^2_{ps} components have been accounted for. Thus one defines $\sigma_{ij}ww = \sigma^2_{ph:s} + \sigma^2_{ps} + \sigma^2_{p}$, where the new component is conceived as a person by substratum interaction within each stratum.

Insert TABLE 8-3 about here

The item variance structure can reasonably be conceived to contain all three covariance components. In the completely

hierarchical structure of the defined item universe each item should tap common-to-all-items variance,common-to-its-stratum variance,and common-to-its-substratum variance.In addition each item will measure something wholly specific which goes as the person by item interaction within substrata.This specific component is called $\sigma^2_{pi:h:s}$,or σ^2_{res}.After this,the expected item variance can be written as a sum of four components,three covariance components and one residual variance component, $E(V_i) = \sigma^2_{pi:h:s} + \sigma^2_{ph:s} + \sigma^2_{ps} + \sigma^2_{p}$.

The latent variance-covariance matrix for a hierarchically stratified test can according to the theoretical construction above be illustrated by a 2-strata-2-substrata-2-items design as presented in TABLE 8-3.In generalizing to a threefacet hierarchically stratified test with $\underline{r}$ strata,$\underline{m}$ substrata within each stratum,and $\underline{k}$ items within each substratum,the sum of a latent variance-covariance matrix will be a sum of weighted components.In a $\underline{kmr}$ x $\underline{kmr}$ matrix there will be $\underline{kmr}$ residual components,$\underline{k^2mr}$ ph:s interaction components,$\underline{k^2m^2r}$ ps interaction components,and $\underline{k^2m^2r^2}$ p components.Thus the expected test score variance as an inferred structure can be written,

$$E(V) = kmr\sigma^2_{pi:h:s} + k^2mr\sigma^2_{ph:s} + k^2m^2r\sigma^2_{ps} + k^2m^2r^2\sigma^2_{p} \qquad F8\text{-}7$$

Which of the covariance components in F8-7 to consider universe score variance in estimating generalizability can only be decided after having made clear what kind of family of tests one is interested in generalizing to.Once again the generalizability problem involves whether strata and substrata are defined as random or fixed,or as an admixture of both.

The most restricted universe of generalization will result by defining both strata and substrata as fixed. This means that all tests belonging to the defined family of tests have to be constructed by random sampling of items from within the fixed substrata within the fixed strata. The same strata and substrata have to provide items for the class of tests to which one wants to generalize. For the fixed model all of the three covariance components defined above will be part of the systematic variance in the test. These will under fixed assumptions be replicable variances, while only that part of the test variance attributable to random sampling of items within substrata, the person by item interaction,will naturally be considered error variance.The reasoning going into this discussion may be made considerably clearer by examining once again TABLE 7-2, which shows which of the covariances to expect in a covariance matrix of two random parallel,fixed,hierarchically stratified tests.

According to the definition of the universe score variance for the fixed model as reached above,the generalizability estimate will be,

$$\text{alpha}_{3F} = \frac{k^2mr\sigma^2_{ph:s} + k^2m^2r\sigma^2_{ps} + k^2m^2r^2\sigma^2_{p}}{kmr\sigma^2_{pi:h:s} + k^2mr\sigma^2_{ph:s} + k^2m^2r\sigma^2_{ps} + k^2m^2r^2\sigma^2_{p}}$$

$$= \frac{k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}}{\sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_{p}} \qquad \text{F8-8}$$

The reduced form of F8-8 is identical to the alpha form for model 3F as developed formally by following rules of thumb in an analysis of variance approach (see F5-2).

For the two other generalizability models under the threefacet test design here considered, the mixed and the random model, a similar line of reasoning as used for F8-8, can be adopted. When strata are considered fixed and substrata random, the covariances of the between-between and the between-within types will be defined as belonging to universe score variance. The covariance within-within has to be allocated to error variance, since that type of covariance for this particular sampling plan will represent non-replicable variance. Therefore, this source of variance has to be regarded as error. This argument can be more convincing by referring to TABLE 7-3, which shows the covariance matrix for two tests constructed according to the sampling plan for the mixed model. The generalizability estimate for the mixed model will read,

$$\text{alpha}_{3M} = \frac{k^2m^2r\sigma^2_{ps} + k^2m^2r^2\sigma^2_p}{kmr\sigma^2_{pi:h:s} + k^2mr\sigma^2_{ph:s} + k^2m^2r\sigma^2_{ps} + k^2m^2r^2\sigma^2_p}$$

$$= \frac{km\sigma^2_{ps} + kmr\sigma^2_p}{\sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_p} \qquad \text{F8-9}$$

When both strata and substrata are considered random, only the covariance between-between can be defined into universe score variance. This can most easily be made clear by the reasoning

established in TABLE 7-4,where the covariance matrix for two hierarchically stratified tests are shown.These tests have been constructed according to the sampling plan prescribed for the random model.The generalizability estimate for this model will be,

$$\text{alpha}_{3R} = \frac{k^2m^2r^2\sigma_p^2}{kmr\sigma_{pi:h:s}^2 + k^2mr\sigma_{ph:s}^2 + k^2m^2r\sigma_{ps}^2 + k^2m^2r^2\sigma_p^2}$$

$$= \frac{kmr\sigma_p^2}{\sigma_{pi:h:s}^2 + k\sigma_{ph:s}^2 + km\sigma_{ps}^2 + kmr\sigma_p^2} \qquad \text{F8-10}$$

The three alpha coefficients for the hierarchically stratified test design,including the fixed,mixed,and random models,have now been derived by three different methods: (1) By an analysis of variance approach (F5-2, F5-3,F5-4). (2) By a covariance approach (F7-1, F7-4, F7-7). (3) By conceiving of a latent variance-covariance matrix of a random parallel test of this particular test design (F8-8, F8-9, F8,10).

It bears repeating that the different approaches converge.The seeming difference is not a real difference. What is of considerable interest to note is that the abstract and formal analysis of variance approach,more often used as a mechanical technique rather than as a tool for thought,can be reinterpreted in terms of a conceptual framework of covariance constructs.By seeing this convergence,analysis of variance as a technical device for most users can be made much more intuitively understandable,such that the generalizability estimates can be derived as logical and meaningful constructs.

9. The family of hierarchical alpha coefficients.

Traditionally coefficient alpha has been associated with the unstratified test design. Yet it seems quite reasonable to believe that the logic of alpha as an internal consistency construct naturally applies to more complex test designs. Also, alpha conceived as the expected correlation among random parallel tests, seems to apply to the different sampling plans within different test designs,like the fixed and random models for the hierarchical test design and the fixed,mixed,and random models for the hierarchically stratified test design. There should be no reason to doubt that the alphas developed for complex test designs are equally suited as lower bound estimates for the defined generalizability coefficients as is traditional alpha. It should be recalled that the defined generalizability is the squared correlation between observed test score and the universe score. No attempt will be made in this monograph to prove that the inequality demonstrated for onefacet alpha also holds for multifacet alpha. The proof for traditional alpha can be found in Rajaratnam,Cronbach & Gleser (1965), Novick & Lewis (1967), and Lord & Novick (1968).

In extending test designs from onefacet to multifacet ones, there are more and more possibilities for design versions. One aspect of the diversity of designs is whether facets are crossed or nested,or a combination of both. The concern in the present study is a threefacet test design with doubly nested items. Yet there are much more to tell about threefacet test designs,not of interest in this particular context. A threefacet measuring

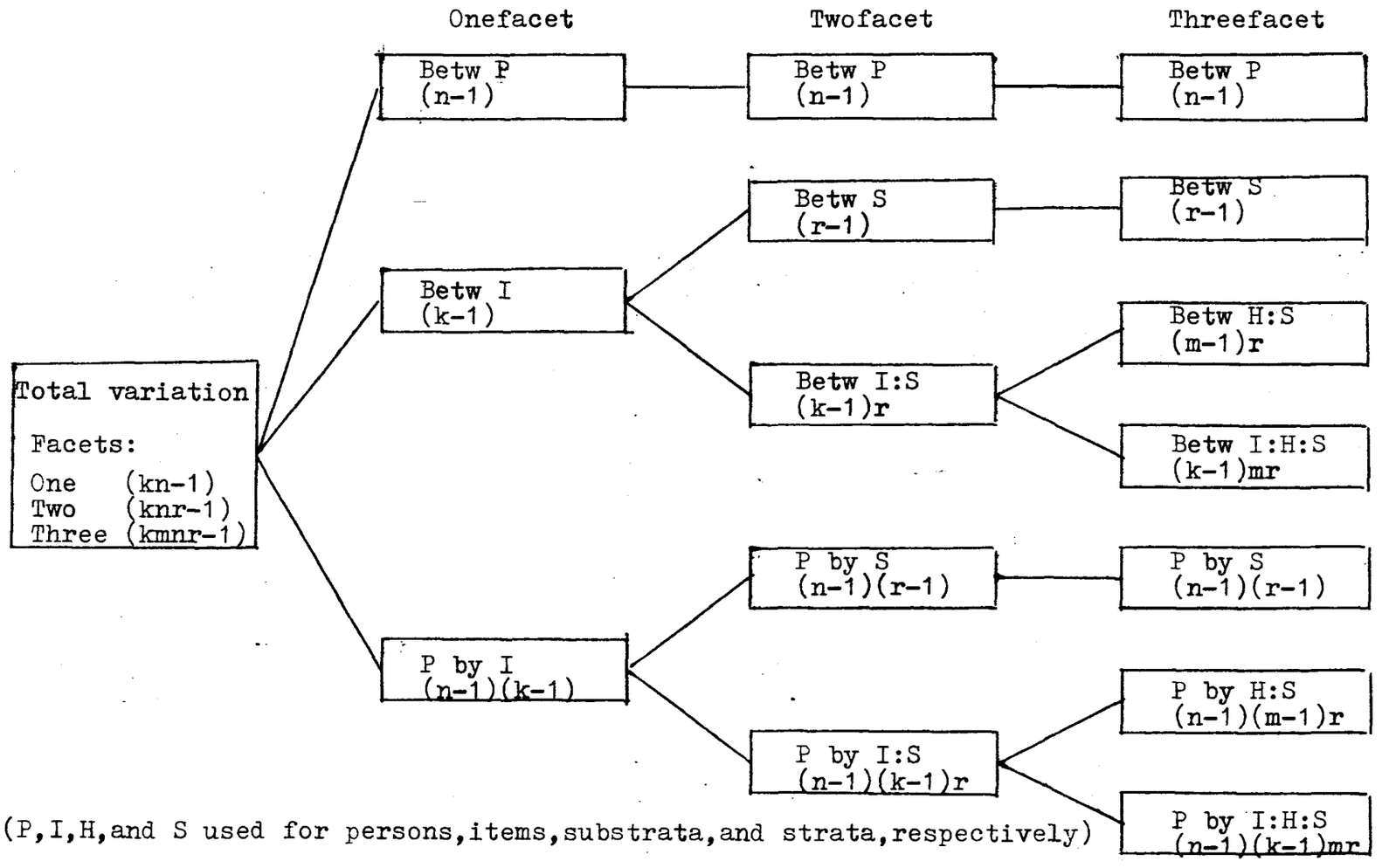

FIGURE 9-1. The relationship of sources of variation for the onefacet, twofacet hierarchical, and the threefacet hierarchically stratified test design, with associated degrees of freedom.

operation may well be of a doubly crossed,or a crossed-nested type (see,for example,Medley & Mitzel 1963). A twofacet operation may be either crossed or nested. The test design described by Rabinowitz & Eikeland (1964) is a twofacet nested design.

There is a relationship between the alphas developed for various test design. In order not to complicate unnecessarily this relatedness,we shall be concerned with establishing a family of alpha coefficients restricted to alphas connected with the unstratified test design,the twofacet nested,and the threefacet doubly nested test design.

These three $tes^{t}/$designs form a tightly knit structure. What is characteristic about the hierarchically stratified test design is that the lower order test designs are built into this more complex one. Within $th^{e}/$strata of the threfacet test one can find as many twofacet nested $deig^{s}/$ns as there are strata,consisting of substrata and items within substrata. Further,each substratum is an unstratified test,consisting of homogeneous items.

Insert FIGURE 9-1 about here

One way of conceiving the relationship between the three test design considered can be seen from FIGURE 9-1. The family tree can be regarded both as a generating and as a degenerating scheme in building item structures. In the case one thinks unstratified items to be heterogeneous, a stratification of items can be undertaken to take care of clustering effects in items.

If desireable,hierarchical clustering effects can be isolated by a second-order stratification,generating a hierarchically stratified test. Conversely, if a doubly nested design should prove too elaborate by showing negligible clustering effects, one can degenerate to less complex designs.

To bring the generalizability formulations for the three test designs more closely together,a little recapitulation may be in order. Although the logic of the various alpha coefficients may be more readily understood by emphasizing a conceptual framework of covariances,as was done in Sections 7 and 8,/the more technical development of the coefficients is most elegantly performed by the analysis of variance formulation.In developing the family of alpha coefficients by the analysis of variance technique,the reader should keep in mind that the covariance approach and the notion of the latent variance-covariance matrix are basically the same models as revealed by the analysis of variance technique (see Section 5).

The latent test score variance for the three test designs is a structure of weighted variance components. In effect,this amounts to focusing on/the inferred structure of the variance-covariance matrices of the different tests. The expected test score variance will be in terms of $E(MS_p)$.

Onefacet $E(MS_p)$ $\quad \sigma^2_{pi} + k\sigma^2_p$ F9-1

Twofacet $E(MS_p)$ $\quad \sigma^2_{pi:s} + k\sigma^2_{ps} + kr\sigma^2_p$ F9-2

Threefacet $E(MS_p)$ $\quad \sigma^2_{pi:h:s} + k\sigma^2_{ph:s} + km\sigma^2_{ps} + kmr\sigma^2_p$ F9-3

TABLE 9-1

Formulas for the family of alpha coefficients

	Defining formulas	Computing formulas
Alpha_{1}	$\dfrac{k\sigma_p^2}{\sigma_{res}^2+k\sigma_p^2}$	$\dfrac{MS_p-MS_{pi}}{MS_p}$
Alpha_{2R}	$\dfrac{kr\sigma_p^2}{\sigma_{res}^2+k\sigma_{ps}^2+kr\sigma_p^2}$	$\dfrac{MS_p-MS_{ps}}{MS_p}$
Alpha_{2F}	$\dfrac{k\sigma_{ps}^2+kr\sigma_p^2}{\sigma_{res}^2+k\sigma_{ps}^2+kr\sigma_p^2}$	$\dfrac{MS_p-MS_{pi:s}}{MS_p}$
Alpha_{3R}	$\dfrac{kmr\sigma_p^2}{\sigma_{res}^2+k\sigma_{ph:s}^2+km\sigma_{ps}^2+kmr\sigma_p^2}$	$\dfrac{MS_p- MS_{ps}}{MS_p}$
Alpha_{3M}	$\dfrac{km\sigma_{ps}^2+kmr\sigma_p^2}{\sigma_{res}^2+k\sigma_{ph:s}^2+km\sigma_{ps}^2+kmr\sigma_p^2}$	$\dfrac{MS_p-MS_{ph:s}}{MS_p}$
Alpha_{3F}	$\dfrac{k\sigma_{ph:s}^2+km\sigma_{ps}^2+kmr\sigma_p^2}{\sigma_{res}^2+k\sigma_{ph:s}^2+km\sigma_{ps}^2+kmr\sigma_p^2}$	$\dfrac{MS_p- MS_{pi:h:s}}{MS_p}$

Note.- $\sigma_{res}^2 = \sigma_{pi}^2, \sigma_{pi:s}^2, \sigma_{pi:h:s}^2$ for onefacet, twofacet and threefacet models, respectively.

From the conceptual structures of test scores in F9-1 to F9-3 one can generate altogether six distinct alpha coefficients when definitions of universe scores are considered by taking into account the different sampling plans that match the conception of the various universes of generalization.

Insert TABLE 9-1 about here

In TABLE 9-1 the alpha coefficients are given both as defining and computing formulas. The definitions are given in terms of weighted variance components, the computations in terms of observed mean squares.

In considering the defining formulas of TABLE 9-1, it may be useful to be reminded that the variance components are defined unconventionally in that a component has the same definition within a test design whatever the sampling plan. This means that whether strata and/or substrata are regarded as fixed or random in the threefacet case,or whether strata are regarded as fixed or random in the twofacet case, the components are uniquely defined as if the facets are all considered random. This ensures that the variance structures of defined universe scores and expected observed scores are maintained intact as structures even when facets are considered fixed. It should be understood that a conventional procedure,as prescribed in experimental design textbooks, where components are defined differently for different sampling plans,would give the same alphas, as the sums of weighted components are intact.

TABLE 9-2

The family of alpha models

Model	Items	Strata	Substrata
1	Random		
2R	Random	Random	
2F	Random	Fixed	
3R	Random	Random	Random
3M	Random	Fixed	Random
3F	Random	Fixed	Fixed

A summary table of the family of alpha coefficients in TABLE 9-2 shows the criteria for the classification of the six measurement models considered. It should be clear that model 1

Insert TABLE 9-2 about here

is the classical onefacet test. Except for the random parallel assumption as adopted by generalizability theory, model 1 is the one discussed by Hoyt (1941) within an analysis of variance framework. Model 2R has been discussed by Rabinowitz & Eikeland (1964). They also discuss model 2F,as do Rajaratnam,Cronbach & Gleser (1965). As far as the author knows, the thre models under the threefacet,doubly nested test design have not previously been discussed in the literature.

10. Describing test score variance in hypothetical data by the family of alpha coefficients.

The relationship established between the three test designs as diagrammed in FIGURE 9-1,makes it feasible to degenerate an originally hierarchically stratified test to a twofacet nested test,which in turn may be degenerated into an unstratified test. It is interesting to see how the alpha coefficients are changed in this degenerating process. It shows the effect of ignoring facets.

The hypothetical test data presented in TABLE 6-1 will be used to illustrate how alpha coefficients change in degenerating

TABLE 10-1

Alpha analysis of hypothetical test data as a threefacet, a twofacet, and a onefacet design by successively collapsing facets, first substrata, next strata.

Threefacet				Twofacet				Onefacet			
Source	SS	df	MS	Source	SS	df	MS	Source	SS	df	MS
P	35,60	4	8,900	P	35,60	4	8,90	P	35,60	4	8,90
S	1,23	1		S	1,23	1					
H:S	1,25	2		I:S	3,35	6		I	4,58	7	
I:H:S	2,10	4									
PS	11,40	4	2,850	PS	11,40	4	2,85				
PH:S	4,00	8	0,500	PI:S	9,40	24	0,39	PI	20,80	28	0,74
PI:H:S	5,40	16	0,338								
Total	60,98	39			60,98	39			60,98	39	
$alpha_{3R}$ = 0,680 $alpha_{3M}$ = 0,944 $alpha_{3F}$ = 0,962				$alpha_{2R}$ = 0,680 $alpha_{2F}$ = 0,956				$alpha_{1}$ = 0,917			

the threefacet design to a onefacet design. The analysis is shown in TABLE 10-1. What can be learnt from the threefacet

Insert TABLE 10-1 about here

alphas,simultaneously viewed,is that substrata within strata are on the average substantially correlated,while the strata are moderately correlated. This is reflected in the negligible difference between $alpha_{3M}$ and $alpha_{3F}$. From this result it is evident that almost no information on individual differences will be lost in degenerating the threefacet test to a twofacet one. This is confirmed by the $alpha_{2F}$ coefficient which is something in between the $alpha_{3M}$ and the $alpha_{3F}$ coefficients. By ignoring the substrata the two fixed alphas,3F and 2F, are practically the same magnitude. This amounts to saying that the correlation among items between substrata are almost equal to the correlation among items within substrata. By this result the substrata may be said to be nonexistent. The indication is that they do not serve any function in the test and can be ignored.

In considering the threefacet test as an unstratified composite, the $alpha_1$ gives a misleading information of how the internal structure of the complex test is constituted.In this analysis the differential traits measured by the strata,as evidenced by the moderate correlation between strata, are ignored.

Applying all alphas to successively more degenerate test designs, undoubtedly can tell which test design is most parsimonious in

accounting most economically for the information sought by the test user. In the present case it seems sound to regard the twofacet design as appropriate for a parsimonious description of the test score variance.

Strictly speaking, the analysis performed in TABLE 10-1 has here been commented upon as a description of test data beyond a more narrowly conceived[n/] generalizability study. In addition to give the expected correlation between random parallel tests according to specified designs, the alpha coefficients considered together can be exploited for the structural information they convey about the composition of the test score variance. Both ways of interpreting and drawing conclusions about test scores may be useful.

11. Traditional Spearman-Brown prophecy formula and the generalizability of hierarchically stratified tests.

For a complex test, say the hierarchically stratified test, it is not easily understandable how a traditional Spearman-Brown rationale is applicable in estimating[the] generalizability of the whole battery by knowing how different parts of the battery go together. Compared to the unstratified test where the Spearman-Brown prophecy formula takes advantage of the average interitem correlation, the situation in the case of the hierarchically stratified test is so much more complicated in that one has to take into account that different parts of the test may go together differently. One has to consider simultaneously the correlation among items within substrata, the correlation of

substrata within strata,and the correlation between strata. A further complicating feature is that in estimating the generalizability of a lengthened threefacet test,one has to consider the many possibilities in reaching a predetermined number of items for a test battery by a combination of number of items within substrata, number of substrata within strata, and number of strata. Still further, how can the notion of fixed and random substrata and strata be included in a traditional Spearman-Brown rationale?

Intuitively,one might think of a procedure that will be conceptually on a par with the Spearman-Brown rationale,and that will give approximate estimates of the generalizability of the threefacet tests,compared to the estimates obtained by the analysis of variance.

Let us be quite concrete about this problem by employing the hypothetical test data of TABLE 6-1 as processed in TABLE 7-1. If the test user is most interested in seeing to what extent the test battery is tapping one common trait,he certainly will pay attention to the beteen strata correlation. In doing this he ignores how substrata go together within strata and how items within substrata correlate. In effect what counts is to find how items from different strata go together.

From the variance-covariance matrix of hypothetical test data, TABLE 7-1, the correlation between strata can easily be obtained by taking the ratio of the covariance between strata to the product of the standard deviations of the two strata. According to classical test theory the correlation between the two strata

would be the reliability for one of them,or for each of them. In going from the reliability of one stratum to the reliability of the sum of the two strata,it seems reasonable to apply the simple Spearman-Brown prophecy formula.

$$r_{s1/s2} = \frac{12,10}{(21,30)^{\frac{1}{2}}\ (25,70)^{\frac{1}{2}}} = 0,517$$

$$r_{tt} = \frac{2 \cdot 0,517}{1 + 0,517} = 0,682$$

The reliability of the whole test battery according to the Spearman-Brown procedure is 0,682. Indeed,one should not be too much surprised to find that this is approximately the generalizability for the random model, 0,680, as found by the previous approaches. By ignoring the correlations between substrata within strata and among items within substrata one has in effect allocated those common sources to the category of error variance as sources of no substantive interest for describing individual differences. It can not be expected that the value obtained by way of the Spearman-Brown procedure should equal the value obtained by the analysis of variance approach. The reason why is that the present approach is an interclass correlation procedure, while the estimate by analysis of variance is an intraclass correlation coefficient. In order for the two procedures to give exactly the same results,the variances of the two strata would have to be equal. A proof for this contention can be found in Haggard (1958), Appendix.

It might be that the test user is substantively interested in the common variance that is reflected in the correlation between substrata within strata in addition to the common variance reflected in the correlation between strata. In that case he intends to generalize to a universe of tests that is more narrow than the preceding one in that the less common variance between substrata included in the universe score variance means that generalization is restricted to fixed strata.

The traditional Spearman-Brown rationale as applied to the present case would involve correlating the two substrata within each of the two strata. The average correlation between substrata within strata is the reliability of one average substratum. In order to obtain the reliability of the full-length test one has to lengthen the substratum four times. To do this, one has to apply the Spearman-Brown prophecy formula once more.

The correlations of interest can be obtained by using the correct covariances and variances in the variance-covariance matrix of the whole test in TABLE 7-1.

$$r_{sub1/sub2} = \frac{5,00}{(7,00)^{\frac{1}{2}} \ (4,30)^{\frac{1}{2}}} = 0,911$$

$$r_{sub3/sub4} = \frac{5,75}{(6,00)^{\frac{1}{2}} \ (8,20)^{\frac{1}{2}}} = 0,820$$

Average substratum correlation: $\frac{0,911+0,820}{2} = 0,865$

$$r_{tt} = \frac{4 \cdot 0,865}{1 + (4-1)0,865} = 0,962$$

Conceptually, the reliability of 0,962 is equivalent to the generalizability estimate obtained for the mixed model. The estimate obtained by analysis of variance for the same model is 0,944. The discrepancy results from the differences in substratum variances.

Still another way of applying the Spearman-Brown rationale for finding the reliability of the whole test is possible. By also regarding the common variance for items within substrata as substantively interesting variance, the test user in effect considers the fixed model as the most appropriate for his purpose. In estimating the reliability of the whole test for this model by the Spearman-Brown procedure, the test user is best advised to find the average correlation between items within substrata. This correlation is taken as the reliability of one average item. As there are 8 items in the test, one has to lengthen the test 8 times in going from the item reliability to the reliability of the whole test. From the variance-covariance matrix for hypothetical test data, TABLE 7-1, the variances and covariances for computing the correlations can be found.

$$r_{i1/i2} = \frac{1,55}{(2,70)^{\frac{1}{2}} \ (1,20)^{\frac{1}{2}}} = 0,861$$

$$r_{i3/i4} = \frac{1,00}{(1,30)^{\frac{1}{2}} \ (1,00)^{\frac{1}{2}}} = 0,877$$

$$r_{i5/i6} = \frac{1,30}{(2,20)^{\frac{1}{2}}\ (1,20)^{\frac{1}{2}}} = 0,800$$

$$r_{i7/i8} = \frac{1,85}{(2,30)^{\frac{1}{2}}\ (2,20)^{\frac{1}{2}}} = 0,823$$

Average item correlation within substrata: $\frac{3,361}{4} = 0,840$

$$r_{tt} = \frac{8 \cdot 0,840}{1 + (8-1)0,840} = 0,977$$

The total test reliability of 0,977 as found by the Spearman-Brown prophecy formula for the fixed model is in conception equivalent to the result obtained for the same model by analysis of variance. That result was 0,962. Again, the discrepancy is a function of unequal item variances in the correlations computed above.

The reasoning underlying the application of the Spearman-Brown procedure for estimating the generalizability for the three models of the hierarchically stratified test seems to be sound, and is corroborated by the results obtained. However, the results are only approximate compared to the analysis of variance results, and the procedure is awkward. What is a desideratum is to be able to see all features of the generalizability problem for this complex test design included in one general formulation. This would be the aim for an extended Spearman-Brown rationale applicable to test batteries of complex structures, like the Primary Mental Abilities tests and the Wechsler scales.

In considering the/way the generalizability problem was solved by the analysis of variance approach,and also in terms of the latent variance-covariance matrix of the total test, there seems to be a fresh starting point for a reformulation of the Spearman-Brown rationale in terms of variance components. That approach will be general enough to take into account the varying number of conditions of each facet going into the/test,and differing sampling plans,simultaneously.

The clue to a completely general solution for a Spearman-Brown prophecy formula that also covers complex test designs, is the inferred structure imposed on the test score variance in the variance-covariance matrix of the test in terms of the variance (covariance) components. By reviewing the expected test score variance as given by F8-7 it should be clear that that formulation contains all that is needed for estimating generalizabilities both for same-length and lengthened tests conceived under different sampling plans. It is here maintained that it is sound reasoning to consider all of the six alpha coefficients,as defined in terms of variance components in TABLE 9-1, to be Spearman-Brown prophecy formulas adopted to particular designs,sampling plans,and number of conditions within each of the facets. Certainly,say for the threefacet test,by regarding the estimates of the parameters (components) as constants and the coefficients as variables,one is free to generalize to lengthened test of any kind of number-of-items, number-of-substrata,and number-of-strata combinations.

12. Analysis of real-world data.

There are quite a few notable test batteries currently in use that fit the hierarchically stratified test model. The Primary Mental Abilities Tests are constructed with abilities as strata and subtests within the strata. The six primary mental abilities are number,verbal meaning,space,word fluency,reasoning,and memory. Within each of the abilities are two subtests (except for memory which has only one). These substests are nested within the abilities, as there is no one-to-one correspondence between subtests for the different abilities. The California Test of Mental Maturity is principally a battery of the same structure. The Wechsler scale is also designed as a threefacet doubly nested test. In WISC,for example, verbal and performance tests constitute strata. Within the verbal stratum the subtests are information,comprehension,arithmetic,similarities,vocabulary, and digit span. Within the performance stratum are picture completion,picture arrangement,block design,object assembly, coding,and mazes. The items are certainly nested within the subtests.

As mentioned,for such complex test batteries, internal consistency analysis has lagged far behind construction. To be sure, the separate substests have been analyzed according to standard procedures for assessing internal consistency for homogeneous tests. But for the whole battery nothing else could be done than performing a split-half reliability study,or correlating strata,or substrata. A simultaneous analysis that can reveal the variance structure of the test scores for such complex designs by specifying the contribution made by each of the

sources to score variance has not been possible until models could be built that fit these designs. These models can only be formalized by exploiting complex mathematical structures that are capable of decomposing variance systems into component variances.

We think that much unexploited information on test score variance can be teased out of[a] hierarchically stratified test by applying the models conceived in the present monograph. This will be shown by analyzing real-world test data from a Norwegian test battery intended to measure mental maturity. Essentially, this battery is of the Thurstone type with strata composed of five abilities, subtests within abilities, and items within the subtests. The abilities are memory, verbal meaning, space, reasoning, and number. Within each of the abilities are are two subtests, except for the space factor which has three subtests.

There are three versions of this test battery for different age groups. From Series III, age group 12-15, data for 13-years old girls are arbitrarily chosen. From the relatively large group used for the standardizing of the test battery, 100 girls are randomly drawn from the larger sample.

The total test battery consists of 114 items. As last items in the subtests to a great extent seemed to be unattempted[n] items, only the first half of each subtest is analyzed in this illustrating study. As is well known, unattempted items scored zero will spuriously increase the internal consistency of a test.

In the present analysis the five-strata-eleven-substrata test battery is reduced to 65 items.

TABLE 12-1

A threefacet alpha analysis of real world test data

Source	SS	df	MS
P	102,617	99	1,037
S	32,781	4	8,195
H:S	21,389	6	3,565
I:H:S	174,008	54	3,222
PS	114,420	396	0,289
PH:S	152,834	594	0,257
PI:H:S	920,199	5346	0,172
Total	1518,248	6499	

$$Alpha_{3R} = \frac{MS_p - MS_{ps}}{MS_p} = \frac{1,037 - 0,289}{1,037} = 0,721$$

$$Alpha_{3M} = \frac{MS_p - MS_{ph:s}}{MS_p} = \frac{1,037-0,257}{1,037} = 0,752$$

$$Alpha_{3F} = \frac{MS_p - MS_{pi:h:s}}{MS_p} = \frac{1,037-0,172}{1,037} = 0,834$$

In the test to be analyzed there are unequal numbers of substrata within strata, and unequal numbers of items within substrata. The models developed in the sections above have for convenience assumed an equal number of substrata within each stratum, and an equal number of items within substrata. No complication will arise in the analysis as long as we keep to the mean squares in the analysis of variance approach. Complicating features arise when it comes to estimating the components.

Although the rationale for analyzing tests of this complexity may be more readily understood by going about the analysis in terms of a covariance approach, the most convenient and practical technique in performing the study is the analysis of variance approach, which will be used here.

Insert TABLE 12-1 about here

The analysis of test data is presented in TABLE 12-1, in which all of the three generalizability estimates are given. If one is solely interested in the generalizability of the test, only one of the estimates can be correct, depending on the definition of the universe for which a psychometric inference is thought to be valid. The choosing of the correct estimate follows the decision to regard strata as fixed or random, and substrata as fixed or random. Test batteries were most likely never constructed according to formal sampling plans like the ones presupposed for the models discussed in this monograph. Therefore, the test constructor will pobably not provide any information as to how the universe of generalization should be defined. Concerning

the test battery in question,it is reaonable to think that strata should be regarded as fixed. A battery constructed along the lines of reasoning done by Thurstone is not likely to have a random sample of abilities drawn form a universe of abilities. Therefore, $alpha_{3R}$ should not be considered the correct estimate of the correlation with another random parallel battery. It might be that the subtests could be regarded as random,as there should be ample possibilities to measure the abilities by choosing other types of subtests. Most likely,in spite of this, the subtests would be regarded as fixed. In that case the generalizability estimate is 0,834. From TABLE 12-1 it is evident that by considering both strata and substrata as fixed, one has gained in generalizability. However,the price to pay for this increase in generalizability is that the universe of generalization is a relatively narrow universe.

With no view to the definition of the universe of generalization, it should be clear that the three alpha coefficients given in TABLE 12-1 are all necessary in obtaining a picture of the structure of the test score variance, and they certainly tell a lot about the coherence among parts in the test battery.

According to the rationale established in the discussion of the models in terms of a covariance approach, the total test score variance is construed to be composed of several additive components. This structure of the test score can be extracted directly from the mean square column in TABLE 12-1 by a subtraction procedure. From the structural model of the hierarchically stratified test presented in TABLE 5-1 it can be seen how one

should proceed to find the weighted components reflecting the pure interaction effects going into the latent structure of the test score variance.

$$\sigma^2_{res} = 0,172$$

$$k\sigma^2_{ph:s} = MS_{ph:s} - MS_{pi:h:s} = 0,257-0,172=0,085$$

$$km\sigma^2_{ps} = MS_{ps} - MS_{ph:s} = 0,289-0,257=0,032$$

$$kmr\sigma^2_{p} = MS_{p} - MS_{ps} = 1,037-0,289=0,748$$

The sum of these weighted components makes up the total test score variance, $1,037 = 0,172 + 0,032 + 0,748$.

More often than being interested in components of absolute magnitudes, one prefers the relative contribution to test score variance made by the different components. Setting the total variance to unit variance, the following structure of proportions is obtained,

$$V_t = 1,000 = \underset{pi:h:s}{0,166} + \underset{ph:s}{0,082} + \underset{ps}{0,031} + \underset{p}{0,721}$$

What is evident from this variance structure is that the contributions are unevenly divided. Most of the variance, 72 %, is contributed by the person component, which is the source of variance representing the common variance running through the whole battery. This is a measure of the loading of the test by one common factor. The person by item component, to the left in the structure, is a measure of the inconsistency of items within substrata. As items in random parallel tests are always

regarded as random, this insconsistency will be a minimum definition of error variance in the test. The two other components are associated with the covariances previously called between-within (0,082) and between-between (0,031). Recalling that in the models constructed, the more general components are imposed on the less general sources of covariance, the components (except σ_p^2) are partial values, reflecting how much they contribute in addition to the more general components.

From the obtained variance structure it is obvious that there can not be much correlation between substrata within strata, and between items within substrata, that according to the model can not be explained by the common factor running through the whole battery. Slightly more than 10 % of the test score variance is explained by these more secific factors, tied to subtests within strata and to items within subtests.

From this description of the test score variance one has gained insight into the homogeneity of the test by how much of the variance can be attributed to one common trait tapped by the battery as a whole, and how much to more specific traits tied to strata and substrata, as parts of the battery. These considerations come close to a factor analytic conception of the hierarchically stratified test.

Without going into any detail in relating the present approach to factor analysis, it should be clear that the factoring in a hierarchically stratified test has been done <u>prior</u> to the analysis. Therefore it may be called an a priori factor analysis in that the factors are associated with strata and substrata

just by the rational stratification made of items. Thus factors as represented by strata and substrata are hypothetical until the analysis reveals whether the test constructor was right in his anticipation of differential abilities that might be measured by the parts of the battery.

A clear interpretation of an over all analysis of the test battery, like the analysis performed in TABLE 12-1, requires that certain assumptions about data are met. These assumptions concern the variances and covariances of the parts constituting the whole test. In the hierarchically stratified test the nesting of substrata within strata and items within substrata is a characteristic feature. As a consequence of nesting,several sources of variance within facets have to be pooled across the facets. For instance,within each stratum there is a person by substratum interaction,conveying information of how much substrata correlate within each of the strata. These interactions are pooled in the analysis to form an over all measure of the person by substratum interaction. It is obvious that such a measure to be meaningful should be based on approximately equal interactions within each of the strata. As is well known,analysis of variance is heavily involved in averaging procedures. The pooling of variances can be misleading if lack of homoscedasticy is apparent in the parts pooled.

Next,an analysis of data by the degenerating procedure described above will be undertaken. This is done in order to see the effect of collapsing substrata as a facet. The sophisticated reader should have no difficulty interpreting the approximately equal values of $alpha_{3R}$ and $alpha_{3M}$ to indicate that the correlation

TABLE 12-2

Alpha analysis of real world data as a threefacet, a twofacet, and a onefacet design by successively collapsing facets, first substrata, next strata.

Threefacet				Twofacet				Onefacet			
Source	SS	df	MS	Source	SS	df	MS	Source	SS	df	MS
P	102,617	99	1,037	P	102,617	99	1,037	P	102,617	99	1,037
S	32,781	4		S	32,781	4					
H:S	21,389	6		I:S	195,397	60		I	228,178	64	
I:H:S	174,008	54									
PS	114,420	396	0,289	PS	114,420	396	0,289				
PH:S	152,834	594	0,257	PI:S	1073,033	5940	$0,18_1$	PI	1187,453	6336	0,187
PI:H:S	920,199	5346	0,172								
Total	1518,248	6499			1518,248	6499			1518,248	6499	
	$alpha_{3R}$ = 0,721				$alpha_{2R}$ = 0,721				$alpha_1$ = 0,820		
	$alpha_{3M}$ = 0,752				$alpha_{2F}$ = 0,825						
	$alpha_{3F}$ = 0,834										

between strata is almost as large as the correlation between substrata within strata. Therefore negligible information on

Insert TABLE 12-2 about here

individual differences will be lost by deleting substrata. This is brought out by analyzing data as a twofacet test design,as seen from TABLE 12-2.

Nevertheless,one might speculate whether this/the most correct is way of collapsing the design. Deleting substrata means that the PH:S interaction is pooled with the PI:H:S interaction. The difference between $alpha_{3F}$ and $alpha_{3M}$ indicates that something specific may be said to be measured by the substrata.It may be reasoned that because strata correlate about as much as substrata within strata, these sources should be pooled,rather than those pooled in TABLE 12-2. This alternative pooling would mean that strata are collapsed, leaving 11 substrata and the same number of items within substrata. In performing this alternative twofacet analysis the following results are obtained: $alpha_{2R}$ = 0,740 and $alpha_{2F}$ = 0,834. The practical result may seem to amount to the same, whatever strategy chosen. Yet the alternative twofacet analysis is logically to be preferred in the light of the alpha coefficients for the threefacet analysis.

The analysis of the hierarchically stratified test as an unstratified test, as performed in TABLE 12-2, has not much to recommend it. In collapsing both strata and substrata the clustering effects have been lost and mixed up in $alpha_1$, which has become conceptually obscure,despite the fact that the value of $alpha_1$ does not seem to be substantially lower than $alpha_{3F}$.

TABLE 12-3

A onefacet alpha analysis of substrata and a two-facet analysis of strata for real-world test data.

Stratum	Substratum	k	$alpha_1$	$alpha_{2R}$	$alpha_{2F}$
M	M_1	14	0,614	0,085	0,598
	M_2	5	0,465		
V	V_1	6	0,282	0,523	0,566
	V_2	5	0,533		
F	F_1	4	0,369	0,322	0,428
	F_2	4	0,340		
	F_3	4	0,005		
R	R_1	6	0,481	0,433	0,580
	R_2	5	0,432		
	Q_1	6	0,690	0,528	0,730
	Q_2	6	0,567		

Note. - M = memory, V = verbal, F = form, R = reasoning, Q = quantitative

The over all analysis of test data performed in TABLE 12-1 is concerned with decomposing the total test score variance.Such a battery is also a multiple score test.Each of these scores on lower levels in the battery may also be analyzed to get a more detailed information on internal consistency in the parts going into the battery. Without further analysis,these parts must be assumed to behave such that the over all analysis can be meaningfully interpreted.

It should be clear that each stratum in the battery is a two-facet nested test unit to which the Rabinowitz-Eikeland models can be applied to examine in more detail how the variance structure is for these lower units in the hierarchy. Further,each of the substrata are unstratified tests that can be analyzed by means of the Hoyt model. The suggested analyses of the twofacet and onefacet test units going into the whole threefacet test battery are shown in TABLE 12-3. First each substratum is ana-

Insert TABLE 12-3 about here

lyzed as a homogeneous test as indicated by the $alpha_1$ column. Next each stratum is analyzed according to the twofacet test models as shown by the $alpha_{2R}$ and $alpha_{2F}$ columns. In the two-facet analysis the coherence of substrata is of particular interest. The number of items going into each substratum after cutting down the tests beause of unattempted items is given in th k column.

The various analyses performed in exploring the internal consistency of the hierarchically stratified test design have demonstrated how the whole family of alpha coefficients in the hierarchy of designs considered in the present monograph can be brought to bear upon both the suprastructure of variance for the total threefacet test battery and the substructures of variance for the lower level designs as parts of the battery.

13. Concluding remarks.

The purpose of the present monograph has been to approach the problem of making psychometric inferences based on measuring operations of complex designs,and examing the composition of the variance structure of scores from such batteries. The hierarchically stratified test design that has been of particular concern is but one of many complex test designs in need of a structural theory. For a long time complex tests containing multiple scores have been lacking such a theory. The theory for the unstratified test is altogether an inadequate theory for multifacet tests.

Guttman saw this need for a structural theory in order to solve the inference problem in psychometrics:

> Conventional sampling problems concern the selection of people from a large population. Mental test theory faces also another type of sampling problem, that of selecting items from one or more indefinitely large universes of content. This is a basic problem of item analysis. To this reviewer it appears that there can be no solution without a structural theory. (Guttman 1953, 129)

Guttman said this in his review of Gulliksen's (1950) Theory of Mental Tests. In the almost twenty years that have passed since this review, some progress has been made in conceiving of such a structural theory. Guttman himself saw the implications for the building of more sophisticated mathematical models. In a later discussion he presented a conceptual framework of how such structures could be conceived in terms of a mathematical system (Guttman 1958).

It might be said that presently we are about to see some of Guttman's facet theory intuitions come true. The multifacet studies in the sixties all converge in that they are basically involved in structural descriptions of complex measuring operations (Medley & Mitzel 1963, Gleser,Cronbach & Rajaratnam 1965).

On the conceptual level,Thorndike (1951) made an excellent approach to classifying the manifold of possible systematic and error variance sources in testing, but no comprehensive theory emerged although complex test designs were in frequent use. At that time there also seemed to be a lack of techniques to analyze complex test data simultaneously to see how the contribution to test score variance by the diverse sources listed by Thorndike could be distinguished.While experimental designs had reached a sophisticated level by way of analysis of variance thinking, a similar sophistication for test designs lagged far behind. This situation was a regrettable result of the schisma that existed for so long between experimental and differential psychology (Cronbach 1957, Cattell 1966, Cronbach,Gleser,Nanda & Rajaratnam 1967).

By now we are about to bridge a gap between a sophisticated conception of the composition of complex test scores and a mathematical system that is considered isomorphic to that conception, emerging in a structural theory. When substantive theory and a formal relational system is brought to converge for complex test designs, a considerable step forward in theory development has been made.

The exploration of the hierarchically stratified test made in this monograph has been involved both in generalizability estimates and structural descriptions of test score variance for this design. There is a close connection between the two ways of considering test data. As shown, the structure imposed on test score variance is an inferred structure, applicable for a pure descriptive purpose. Yet this structure can be exploited in making inferences about how much of the test score variance can be attributed to universe score variance. Crucial for this mode of thinking is that one defines a family of hierarchically stratified tests, constructed according to a specified sampling plan. For a multifacet instrument a sampling plan prescribes what facets to regard as fixed and/or random. The construction of tests belonging to the same defined family of tests will have to follow the companion sampling plan.

It ought to be recalled that for a test to be random parallel, whatever the sampling plan, items at least have to be considered random. In the context of generalizability theory, items can never be fixed. For the interpretation of test scores in terms of generalizability the information needed is contained in the

composition of the test score variance, brought forth by the structural analysis. The generalizability part of the game is to reassemble the components of the test score variance into two categories of variance, the universe score and the error score variance. For each of the models under the hierarchically stratified test considered, this splitting up into two categories of variance will be different.

At present it might be difficult to find real world experiments that fit all of the three models for the hierarchically stratified.[test/] It is not difficult to find tests that most likely can be thought to fit the fixed model. There might be tests that are appropriately interpreted to fit the mixed model. For the random model, however, there seems to be no known existing real world experiment that applies. Yet it seems likely that tests could be conceived that match a practical testing situation in which all facets could reaonably[s/] be considered random.

It should be strongly emphasized that whether the three generalizability models fit or not, the structural analysis is still useful. As a matter of fact, it is here argued that the most interesting and informative analysis of complex test data is the description of test score variance. The structural analysis is a correlational approach that describes the relationship of the parts going into the hierarchy. The decomposing into variance components is the fundamental basis for making a meaningful interpretation of the observed test score in terms of the extent to which the battery is measuring one common trait running through all items and less common traits attributed to strata. Even specific traits can emerge, attributable to the substrata.

The idea of a latent covariance structure is the basis for the theory of the hierarchically stratified test as here developed. This inferred structure, imposed on data, makes it more easily understandable what the underlying rationale for the analysis of variance approach is. Yet the structural theory for the particular threefacet test design discussed in this monograph is in fact a very general conceptual framework that applies to other designs as well.

Actually, we think that this general structural theory is but an extension of the long-respected Spearman-Brown rationale. That rationale has so far been restricted to the lowest level in the hierarchy of test designs, the unstratified test. The Spearman-Brown rationale has been the cornerstone in mental test theory for more than sixty years. What seems to come out of multifacet studies conducted so far, is that the Spearman-Brown basic thinking in test theory is about to get a much more general formulation. The new perspective for this old formula covers a variety of complex measurement procedures, where the hierarchically stratified test design is but one.

References

Brown,W. 1910. Some experimental results in the correlation of mental abilities. British Journal of Psychology 3, 296-322.

Burt,C. 1954. The sign pattern of factor-matrices. The British Journal of Statistical Psychology 7, Part I, 15-29.

Campbell,D.T. & Fiske,D.W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin 56, 81-105.

Cattell,R.B. 1966 a. Psychological theory and scientific method. Chapter 1 in Cattel,R.B.(Editor), Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally.

Cattell,R.B. 1966 b. The principles of experimental design and analysis in relation to theory building. Chapter 2 in Cattell,R.B.(Editor), Handbook of Multivariate Experimental Psychology.Chicago: Rand McNally.

Cronbach,L.J. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297-334.

Cronbach,L.J. 1957. The two disciplines of scientific psychology. American Psychologist 12, 671-684.

Cronbach,L.J., Rajaratnam,N. & Gleser,G.C. 1963. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology 16, 137-163.

Cronbach,L.J.,Gleser,G.C.,Nanda,H. & Rajaratnam,N. 1967. The dependability of behavioral measurements: Multifacet studies of generalizability. Preliminary version. Stanford University, September 1967.

Eikeland,H.M. 1970. Coefficient alpha and the expected variance-covariance matrix of random composite measurements.Mimeographed. To appear in Scandinavian Journal of Psychology.

Eikeland,H.M. 1971. Correlational analyses of school marks influenced by multiple sources of variance. Explorations into internal structures of complex systems of variation. Paper presented at a symposium on evaluation in Uppsala, Sweden, November 1971. Mimeographed.

Eikeland,H.M. 1972. Components of reliability for intra-individual difference scores. Mimeographed.

Gleser,G.C., Cronbach,L.J. & Rajaratnam,N. 1965. Generalizability of scores influenced by multiple sources of variance. Psychometrika 30, 395-418.

Guilford,J.P. 1954. Psychometric Methods. New York: McGraw-Hill.

Guilford,J.P. 1967. The nature of human intelligence. New York: McGraw-Hill.

Gulliksen,H. 1950. Theory of mental tests. New York:Wiley.

Guttman,L. 1953. A special review of Harold Gulliksen's Theory of mental tests. Psychometrika 28, 123-130.

Guttman,L. 1958. What lies ahead for factor analysis ? Educational & Psychological Measurement 18, 497-515.

Haggard,E.A. 1958. Intraclass correlation and the analysis of variance. New York: Dryden.

Horst,P. 1965. Factor analysis of data matrices.New York: Holt,Rinehart & Winston.

Hoyt,C. 1941. Test reliability estimated by analysis of variance. Psychometrika 6, 153-160.

Jackson,R.W.B. & Ferguson,G.A. 1941. Studies on the reliabilities of tests. Bulletin 12. Department of Educational Research. Toronto: University of Toronto.

Kaiser,H.F. & Caffrey,J. 1965. Alpha factor analysis. Psychometrika 30, 1-14.

Kirk,R.E. 1968. Experimental design:Procedures for the behavioral sciences. Belmont,California: Brooks/Cole.

Lord,F.M. & Novick,M.R. 1968. Statistical theories of mental test scores. Reading,Mas: Addison-Wesley.

Magnusson,D. 1967. Test theory. Reading,Mass: Addison-Wesley.

Medley,D.M. & Mitzel,H.E. 1963. Measuring classroom behavior by systematic observation. In Gage,N.L.(Editor),Handbook of research on teaching. Chicago: Rand McNally.

Millman,J. & Glass,G.V. 1967. Rules of thumb for writing the ANOVA table. Journal of educational measurement 4, 41-51.

Mosier,C.I. 1951. Batteries and profiles. In Lindquist,E.F. (Editor), Educational measurement. Washington D.C.: American Council of Education.

Novick,M.R. & Lewis,C. 1967. Coefficient alpha and the reliability of composite measurements. Psychometrika 32, 1-13.

Rabinowitz,W. & Eikeland,H.M. 1964. Estimating the reliability of tests with clustered items. Pedagogisk Forskning (Scandinavian Journal of Educational Research) 8, 86-106.

Rajaratnam,N., Cronbach,L.J. & Gleser,G.C. 1965. Generalizability of stratified-parallel tests. Psychometrika 30, 39-56.

Spearman,C. 1910. Correlation calculated with faulty data. British Journal of Psychology 3, 271-295.

Stanley,J.C. 1961. Analysis of a doubly nested design. Educational & Psychological Measurement 21, 831-837.

Thorndike,R.L. 1951. Reliability. In Lindquist,E.F.(Editor) Educational measurement. Washington D.C.:American Council on Education.

Tryon,R.C. 1957. Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin 54, 229-249.

Winer,B.J. 1962. Statistical principles in experimental design. New York: McGraw-Hill.