chapter 6

Test Homogeneity, Reliability, and Generalizability

If two variables are correlated, there are at least three ways we can "explain" the presence of a relationship between them.

1. It may be that one variable (partly) determines the other, in a sense that has no converse. We say that one is a cause of the other. For example, observing a rat in an activity cage, we say that hunger causes activity. And after a good workout we say activity causes hunger. But we do not say that activity and hunger are merely contingently associated. Both the meaning and the verification of causal claims are deep and controversial matters. It is to be hoped that the reader, like the writer, makes a distinction between "teaching causes learning" and "teaching and learning are activities often found together."
2. It may be that the two variables are related effects of a common cause. For example, distinct stock prices vary together from the impact of political events on the psyches of market "players."
3. It may be that the two variables are correlated because they measure, or indicate, something in common. This can be literally true. Some tests contain items that can be scored for more than one trait, and the correlation between scores for the traits comes from shared score components. This would generally be regarded as a spurious correlation. But by *measuring in common*, nothing quite so literal is intended. The notion is that the variables are indicators, "symptoms" or manifestations of the same state of affairs. For example, extraversion is an abstract concept whose instances are the recognized extravert behaviors, and it is therefore circular to say that extraversion "causes" its manifestations. This explanation of relatedness has already been used earlier in a qualitative and intuitive way. At this point we use it to introduce a statistical model to refine our conception of homogeneous tests—of tests whose items are all of the same kind. This is a more general model, and the true-score model is just a special case of it. The model we need deals with relationships between the items, not just relationships between total test scores.

The next section shows how the Spearman single-factor model can be used to test the homogeneity of a set of items. The following section derives a measure of reliability from the parameters of the factor model, and the fourth gives two special cases. The fifth section gives results for binary data. The last section, which can be omitted without loss of continuity, gives an introduction to the theory of generalizability.

HOMOGENEITY AND THE SINGLE-FACTOR MODEL

In this chapter we do not need to distinguish between a test composed of m items that are not decomposable into smaller elements, and a test composed of m subtests—including *item bundles* or *testlets*, each of which are decomposable into their constituent items. We could also have a *test battery* composed of m tests, if there is any reason to form a total score over the entire set of tests. It is convenient to regard items and item scores here as including subtests and subtest scores. All of the results in this chapter apply to whatever measurements we regard as the basic sets of scores to be combined into a global test score.

Suppose m items $j = 1,\ldots, m$, with scores $X_1,\ldots, X_m$, have pairwise co-variances $_{jk}$ in a population of interest. These can be estimated from a sample by the formulas (3.12) and (3.13), correcting for bias in small samples. We wish to give a statistical meaning to the idea that these co-variances are nonzero because the items measure just one attribute in common. This requires an extension of the idea in the alternate-forms treatment of true-score theory. In this version, the statements

$$Y = T + E \quad \text{and} \quad Y' = T + E'$$

mean that T is the attribute common to the test forms; that it is measured equally well by either; and that E and E' are due to unique or idiosyncratic properties of the particular items in the separate forms.

If we apply this principle to the m items, we might consider the model

$$X_j = T + E_j \tag{6.1}$$

for each of them. However, the model (6.1) is a special case of the one that is needed, and it will be studied later for its special properties. For more than two items the model needs to allow three possibilities in explaining the covariances by a common attribute. First, it must be possible that indicators of the common attribute have different means in the population of interest. Items of different difficulties can measure the same thing. Second, the items are not equally good indicators in general. Some items may measure the attribute more sensitively than others. They may discriminate more clearly between levels of the attribute. Third, the items may have different amounts of unique variance—of variation due to their idiosyncratic properties. We symbolize the quantity measured in common by F, and we call it the *common factor* that ties the items together. The appropriate model is the single general factor model of Charles Spearman, namely:

$$X_j = \mu_j + \lambda_j F + E_j \qquad j = 1, \ldots, m. \tag{6.2}$$

In this model, X_j is a random examinee's score on the jth item, F is the examinee's measure of the common attribute, and E_j is the examinee's measure of the unique or idiosyncratic property of item j. More precisely, E_j is the amount by which the idiosyncratic property of item j shifts the response X_j, in a positive or negative direction, from the expected level of response to the attribute itself. The constants μ_j allow each item to have a distinct difficulty.

Following an old tradition, the coefficient λ_j is labeled the *factor loading* of item j. (Metaphorically, it is the extent to which the item is "loaded" with the attribute. Some "carry" more of it than others.) The factor loading measures how sensitively each item functions as an indicator of the common factor/attribute F. An item whose λ_j value is (relatively) large is a better indicator of F than one whose λ_j value is (relatively) small. It represents the amount of difference in the item score that corresponds to a unit difference in the attribute. It is therefore a measure of the ability of the item to discriminate between subjects with high and low values of F. It may be considered a measure of the *discriminating power* of the item. A (psychometrically) *homogeneous* test is one whose items measure just one attribute in common—a common factor. We can check on our judgment that the items are of the same kind—of homogeneous content—by seeing if the responses to them fit the single factor model.

As in the special case of alternate forms, the unique component E_j is independent of the common factor F (by definition), so it is uncorrelated with it. Also, any two unique components E_j, E_k are (by definition) independent of each other, so they are also uncorrelated. Equation (6.2) expresses the regression of X_j on F, in the usual meaning of regression theory.

Consequently, λ_j is the expected difference in X_j for a unit difference in F between subgroups of the population. An item with zero λ_j does not measure F at all.

Starting from the model (5.1), we were able to compute the variance of the true part and the variance of (either) error part, from the variances of the two measurements and their covariance using the derived equations. Now we need to generalize this. The conditions are that the common factor is uncorrelated with each unique component and the unique components are uncorrelated for all distinct items. (These conditions are not assumptions. They state what is meant by the *common factor F* and the *unique parts Ej.)* To determine a scale for F, we are free to consider it as a standard score, with mean zero and variance unity in the population studied. We use ψ_j^2 for the variance of E_j. This variance is referred to in the literature either as the *unique variance* of the item or as its *uniqueness*. In this model the covariance of any two item scores X_j, X_k is just the product of their factor loadings. That is,

$$\sigma_{jk} = \lambda_j \lambda_k. \tag{6.3}$$

Also, the variance jj of the jth item is expressed as

$$\sigma_{jj} = \lambda_j^2 + \psi_j^2, \tag{6.4}$$

the sum of the squared loading of the item and its unique variance.

***Proof is by the algebra of expectations, with terms in Cov{F,E_j} and Cov{E_j, E_k}becoming zero because of the assumptions. $$$

The quantities $\lambda_1, \ldots, \lambda_m, \psi_1^2, \ldots, \psi_m^2$ are the parameters of this model for the population. In practice we need to estimate them from samples. For the moment, suppose we know the numerical values of the parameters. It is then easy to use an ordinary calculator to compute the resulting variances and covariances by (6.3). Consider, for example, Table 6.1. (The student should check one or two further values.)

In applications we need the reverse procedure. Given the 5 variances and 10 covariances in Table 6.1(b), is it possible to compute the 10 parameters—five loadings, and five unique variances? The answer is yes. The factor loadings can be obtained from any three items, j, k, and l, by using

$$\lambda_j = \sqrt{\sigma_{jk}\sigma_{jl}/\sigma_{kl}}. \tag{6.5}$$

TABLE 6.1
Computation of Item Covariances From Spearman Parameters

(a) Suppose $\sigma_F^2 = 1$

and	$\lambda_1 = 1.8$	$\psi_1^2 = 1.0$		$\sigma_{21} = 2.7$
	$\lambda_2 = 1.5$	$\psi_2^2 = 1.2$		(1.8×1.5)
	$\lambda_3 = 1.2$	$\psi_3^2 = 1.4$	then	$\sigma_{11} = 4.24$
	$\lambda_4 = 1.0$	$\psi_4^2 = 1.6$		$(1.8^2 + 1.0)$
	$\lambda_5 = 0.8$	$\psi_5^2 = 2.0$		and so on

giving

(b) $\Sigma = 1$

	1	2	3	4	5
1	4.24	2.70	2.16	1.80	1.44
2	2.70	3.45	1.80	1.50	1.20
3	2.16	1.80	2.84	1.20	0.96
4	1.80	1.50	1.20	2.60	0.80
5	1.44	1.20	0.96	0.80	2.64

[If all the covariances are positive, all the loadings are positive. If some covariances are negative, a possible choice of the negative square root in (6.5) is resolved by consulting their signs. If we take *j* positive and *k* negative, *jk* is negative—and so on.] Then to get ψ_j^2 we use

$$\psi_j^2 = \sigma_{jj} - \lambda_j^2. \tag{6.6}$$

The important result is that item covariances fitting the model determine the loadings, and then the item variances determine the uniquenesses.

*** Proof of (6.5) and (6.6): From any three items *j*, *k*, *l*, we have

$$\sigma_{jk} = \lambda_j\lambda_k \qquad \sigma_{jl} = \lambda_j\lambda_l \qquad \sigma_{kl} = \lambda_k\lambda_l.$$

Then

$$\lambda_j^2 = (\sigma_{jk}\sigma_{jl})/\sigma_{kl}$$

which gives (6.5). $$$

For example, in Table 6.1,

$$\sigma_{12} = 2.70 \qquad \sigma_{13} = 2.16 \qquad \sigma_{23} = 1.80,$$

so

$$\lambda_1 = \sqrt{(2.70 \times 2.16)/1.80} = 1.80$$

and

$$\psi_1^2 = 4.24 - 1.80^2 = 1.0.$$

(The student might check some more of these results.) Notice that we can compute each loading in a number of ways. Here we can get λ_1 in six ways, from

$$\sigma_{12}\sigma_{13}/\sigma_{23} =$$
$$\sigma_{12}\sigma_{14}/\sigma_{24} =$$
$$\sigma_{12}\sigma_{15}/\sigma_{25} =$$
$$\sigma_{13}\sigma_{14}/\sigma_{34} =$$
$$\sigma_{13}\sigma_{15}/\sigma_{35} =$$
$$\sigma_{14}\sigma_{15}/\sigma_{45} = .$$

These results must be consistent because Table 6.1—an artificial example—contains exactly fitting "population" values.

The object of this demonstration was to show that the parameters of the single-factor model are determined by the covariance matrix they generate. Formally, we say the parameters are *identified*—they have a unique identity as functions of the variances and covariances. That is, (6.3) and (6.4) have unique solutions (6.5) and (6.6). Some readers will know that a set of simultaneous linear equations may or may not have a consistent or a unique solution. Equations (6.3) and (6.4) are simultaneous nonlinear equations. More general common factor models—with more than one factor—can fail to have unique solutions. (See Chap. 9.) The Spearman model is a possibly false statistical hypothesis, which may or may not fit a given set of items. Table 6.2 contains a covariance matrix resembling the one in Table 6.1, which does not fit the model (6.2) for homogeneity. Using covariances of variables 1, 2, 3 gives

$$\lambda_1 = \sqrt{(2.70 \times 2.16)/1.80} = 1.8$$

as before, but using variables 1, 4, 5 gives

$$\lambda_1 = \sqrt{(.18 \times .144)/.96} = .164.$$

Table 6.2 corresponds to a case where items 1, 2, 3 are homogeneous, and items 4 and 5 are homogeneous, but each group measures different things, so jointly they need more than one common factor—more than one attribute—to explain the relations between the items. This is the topic of Chapter 9.

TABLE 6.2
Covariance Matrix—Nonhomogeneous Case

	1	2	3	4	5
1	4.24	2.70	2.16	0.18	0.144
2	2.70	3.45	1.80	0.15	0.12
3	2.16	1.80	2.84	0.12	0.96
4	0.18	0.15	0.12	2.60	0.80
5	0.144	0.12	0.96	0.80	2.64

For decades—from 1907 to 1967—there were only crude devices for estimating the parameters of a common factor model. Consider the co-variance matrix in Table 6.3. This is a sample covariance matrix from items marked A, B, C, D, E in Table 5.1, scored 1–7 as indicated there (which explains the order of magnitude of the item variances). The sample size is $n = 215$. In an example like this, with a small number of items, we can use the earliest device for fitting the Spearman—one-factor—model. This device is due to Spearman himself. We apply the expression (6.5) for every k,l pair to the sample covariances, getting somewhat varying estimates of j, and then average these estimates. For the estimated factor loading of item 1, the expressions listed

$$\hat{\lambda}_1 = [(1.560 \times 1.487)/1.283]^{1/2} = 1.345$$
$$\hat{\lambda}_1 = [(1.560 \times 1.195)/.845]^{1/2} = 1.485$$
$$\hat{\lambda}_1 = [(1.560 \times 1.425)/1.313]^{1/2} = 1.301$$
$$\hat{\lambda}_1 = [(1.487 \times 1.195)/1.127]^{1/2} = 1.256$$
$$\hat{\lambda}_1 = [(1.487 \times 1.425)/1.313]^{1/2} = 1.270$$
$$\hat{\lambda}_1 = [(1.195 \times 1.425)/1.323]^{1/2} = 1.135$$

give an averaged estimate 1.299. Then the estimate of ψ_1^2 is just the variance, 2.566, minus the squared loading, 1.299^2, giving .879. Similarly, we can get the remaining parameters. The reader will find it easy but very tedious to do this for the remaining four loadings.

TABLE 6.3
Covariance Matrix—Satisfaction With Life Scale

	1	2	3	4	5
1	2.566	1.560	1.487	1.195	1.425
2	1.560	2.493	1.283	0.845	1.313
3	1.487	1.283	2.462	1.127	1.313
4	1.195	0.845	1.127	2.769	1.323
5	1.425	1.313	1.313	1.323	3.356

Modern methods of estimation use a computer program that systematically searches—quite literally—for a set of parameter values that make a *discrepancy function*—a function of the discrepancies, $s_{jk} - _{jk}$ between the unbiased sample covariances and the fitted covariances—as small as possible. A simple and intuitively natural discrepancy function is

$$q_u = (1/m^2) \sum_j \sum_k (s_{jk} - \sigma_{jk})^2, \tag{6.7a}$$

the ordinary mean of the squared differences between sample and fitted values. (Other discrepancy functions are mentioned later.) This function is the unweighted least squares function—the ULS function. For the single-common-factor model, this discrepancy function can be broken down into two parts, one for the diagonal elements, $s_{jj} - _{jj}$, and one for the off-diagonal elements, $s_{jk} - _{jk}$, by writing

$$q_u = (1/m^2) \left[\sum_{j \neq k} \sum (s_{jk} - \lambda_j \lambda_k)^2 + \sum_j (s_{jj} - \lambda_j^2 - \psi_j^2)^2\right]. \tag{6.7b}$$

Unweighted least squares is so called because the discrepancy function q_u gives equal "value" or "weight" to each discrepancy. The computer searches for values $\hat{\lambda}_j, \hat{\psi}_j^2$ of λ_j, ψ_j^2 that make this mean-square discrepancy as small as possible. It is not necessary for the general reader to know the systematic process—the *algorithm* by which a computer finds the minimum. Having found it, the computer program prints out the estimates and the entire matrix of discrepancies, $s_{jk} - \hat{\sigma}_{jk}$, if the user asks for it. (This should always be examined by the program user.) Estimation using ULS does not give a statistical test of the hypothesis, and it does not tell us how accurately we have estimated the parameters. We can look at the discrepancies and judge that they are "negligible." A good measure of closeness of fit of the model takes into account the magnitude of the sample covariances themselves, by computing

$$c = (1/m^2) \sum_j \sum_k s_{jk}^2 \tag{6.8}$$

and then defining the *goodness of fit index*

$$\text{GFI} = 1 - q_u / c. \tag{6.9}$$

If the fit is good, the GFI is close to unity, which would be perfect fit. ULS makes no assumptions about the distribution of the item scores (although the results will be affected by that distribution).

Fitting the single-factor model to the covariance matrix of Table 6.3 by program CONFA1 using the ULS function gives the fitted parameter values in Table 6.4(a) and the *fitted* matrix—reproduced from those values—in Table 6.4(b). The discrepancies $s_{jk} - \hat{\sigma}_{jk}$ are given in Table 6.4(c). We do not usually study the fitted covariances, as it is hard to judge how close they are to the sample values. We should always look at the discrepancies.

Using (6.8) we obtain for the mean of squares of the sample covariances calculated from Table 6.3, $c = 2.866$. For the mean of squares of the discrepancies in Table 6.4, we obtain $q_u = .00915$, so the goodness of fit index

$$\text{GFI} = .9968.$$

Experience suggests this is a very good fit. As a guide to the inexperienced user, but not a firm decision rule, I suggest that the fit is "good" when GFI is greater than .95, and "acceptable" when it is greater than .9. The discrepancy matrix will tell us if any misfit is due to just one or two large discrepancies, or a general spread of discrepancies over the matrix. Generally, the discrepancy matrix allows a much more informed judgment than any simple fit index. Unfortunately, most investigators appear to rely entirely on a fit index and a rule of thumb for its application.

TABLE 6.4
Satisfaction With Life Scale—Spearman Analysis

	(a)		(b)				
	λ	ψ^2	1	2	3	4	5
1	1.290	0.901	2.565	1.424	1.481	1.328	1.529
2	1.104	1.274	1.424	2.493	1.267	1.051	1.308
3	1.148	1.144	1.481	1.267	2.462	1.093	1.360
4	0.952	1.863	1.328	1.051	1.093	2.769	1.128
5	1.185	1.951	1.529	1.308	1.360	1.128	3.355

	(c)				
	1	2	3	4	5
1	.0	.135	.006	−.033	−.104
2	.135	.0	.015	−.206	.004
3	.006	.015	.0	.035	−.048
4	−.033	−.206	.035	.0	.195
5	−.104	.004	−.048	.195	.0

Next we consider two special cases of the single-factor model, resembling (6.1). To repeat, the model for m homogeneous items is

$$X_j = \lambda_j F + E_j + \mu_j,$$

where F and E_j are uncorrelated for every item j, and E_j and E_k are uncorrected for all distinct pairs. There are two important special cases of this model.

True-score equivalent items—also known as *essentially tau-equivalent* items—can be defined by the property

$$\lambda_1 = \lambda_2 = \cdots = \lambda_m = \lambda.$$

For then

$$X_j = \lambda F + E_j + \mu_j, \tag{6.10}$$

which becomes

$$X_j = T_x + E_j + \mu_j \tag{6.11}$$

if we write Tx for F, that is, if we rescale the common attribute so that it is measured in the same units as the items. That is, the *item true-score* Tx is just the common factor multiplied by the factor loading. It then has variance 2. To describe the items as true-score equivalent is to declare that they measure the common property equally sensitively—with equal discrimination. Then the common factor can be rescaled so as to be considered the "true" part of each item score. Note that in this chapter we have a number of "true scores," which are distinguished by a subscript indicating the variable of which each is the "true part."

The *true-score equivalence* model (6.10) or (6.11) gives, for the elements of the covariance matrix,

$$\sigma_{jk} = \lambda^2 = \sigma_{Tx}{}^2 \tag{6.12}$$

for all j not equal to k, and

$$\sigma_{jj} = \lambda^2 + \psi_j^2 = \sigma_{Tx}^2 + \sigma_j^2 \tag{6.13}$$

for all j. This is a very restrictive hypothesis. The $m(m-1)/2$ covariances of distinct items must all be equal.

Parallel items can be defined by the property that $\lambda_1 = \lambda_2 = \cdots = \lambda_m = \lambda$, plus the property that $\psi_1^2 = \psi_2^2 = \cdots = \psi_m^2 = \psi^2$. Accordingly, the *parallel items* model is again written as (6.10) or (6.11), and gives (6.12) for the covariances, but it also restricts the item variances to

$$\sigma_{jj} = \lambda^2 + \psi^2 = \sigma_{Tx}^2 + \psi^2. \tag{6.14}$$

That is, all covariances are equal to each other, and all variances are equal to each other. Some writers define parallel items so that they must also have equal means. To make a distinction that we need in the present treatment, we call items with this further restriction *strictly parallel* items. (This is also consistent with the definition of strictly parallel items used in item response theory.) Equation (6.14) includes the basic breakdown of variance in Chapter 5 for test scores from two alternate forms, when $m = 2$. But then the hypothesis of parallelism is not restrictive and not testable/falsifiable. Using only information from total test scores in this way is generally unsafe because it is uninformative.

In the case of true-score equivalent and parallel items, the ULS estimators are precisely what intuition would guess them to be. They can easily be calculated from the sample covariance matrix, without using a computer search algorithm. (In practice we would still find it easier to use a computer program.) In the true-score equivalence model, the ULS estimate $\hat{\sigma}_T^2$ of σ_T^2 in (6.12) is the average of all covariances of distinct items, i.e.,

$$\hat{\sigma}_T^2 = [1/m(m-1)] \sum\sum_{j \neq k} s_{jk}, \tag{6.15}$$

and the estimate of ψ_j^2 in (6.13) is

$$\hat{\psi}_j^2 = s_{jj} - \hat{\sigma}_T^2, \tag{6.16a}$$

for each item. (In small samples, we would use unbiased estimates of the variances and covariances.) In the parallel-items model, the estimate of σ_T^2 is (still) (6.15), and the estimate of the constant ψ^2 is

$$\hat{\psi}^2 = (1/m) \sum_j s_{jj} - \hat{\sigma}_{Tx}^2. \tag{6.16b}$$

From Table 6.3, the reader may easily verify that

TABLE 6.5
Discrepancy Matrices—Restricted Models

(a) True-Score Equivalence Model

	1	2	3	4	5
1	.0	.273	.200	−.092	.138
2	.273	.0	−.004	−.442	.026
3	.200	−.004	.0	−.160	.026
4	−.092	−.442	−.160	.0	.036
5	.138	.026	.026	.036	.0

(b) Parallel Items Model

	1	2	3	4	5
1	−.163	.273	.200	−.092	.138
2	.273	−.236	−.004	−.442	.026
3	.200	−.004	−.268	−.160	.026
4	−.092	−.442	−.160	.040	.036
5	.138	.026	.026	.036	.627

$$\hat{\sigma}^2_{Tx} = 1.287,$$

which is the average of the elements below the diagonal. In the true-score-equivalence model, $\hat{\psi}^2_1 = 1.279$, $\hat{\psi}^2_2 = 1.206$, $\hat{\psi}^2_3 = 1.175$, $\hat{\psi}^2_4 = 1.482$, and $\hat{\psi}^2_5 = 2.069$. The parallel-items model gives the same estimate of true score variance, of course, and, from Table 6.3,

$$\hat{\psi}^2 = (1/5)(2.566 + \cdots + 3.356) - 1.287 = 1.442.$$

The true-score equivalence model gives the discrepancy matrix in Table 6.5(a), and GFI = .991. The parallel-items model gives the discrepancy matrix in Table 6.5(b), and GFI = .949. Both these models give acceptable approximations by the usually accepted criteria.

THE RELIABILITY OF A HOMOGENEOUS TEST

From the single-factor model (6.2) for homogeneous items, the total test score is

$$Y = \sum_j X_j = (\sum_j \lambda_j)F + \sum_j E_j + \sum_j \mu_j. \qquad (6.17a)$$

If we write $C = (\Sigma\ \lambda_j)F$ for the part of Y due to the common factor, and $U_j = \Sigma\ E_j$ for the part due to unique properties, then

$$Y = C + U + \mu_Y, \qquad (6.17b)$$

which is the sum of the *common part* of Y and its *unique part* (plus the mean of *Y)*. The common part C is the measure of the attribute given by Y, and the unique part is the error of measurement, so we are back to the decomposition into true and error parts, as in Chapter 5. But now we have a clear meaning for the decomposition and a clear method for estimating the variances of these. That is, the factor model gives us an interpretation of the total true score T_Y as the common part C of Y, and of the error E as the unique part U. So in another notation, (6.17) can be written as

$$Y = T_Y + E_Y + \mu_Y. \qquad (6.17c)$$

By the algebra of expectations,

$$\sigma_Y^2 = (\sum \lambda_j)^2 + \sum \psi_j^2, \tag{6.18a}$$

or

$$\sigma_Y^2 = \sigma_C^2 + \sigma_U^2, \tag{6.18b}$$

that is,

$$\sigma_Y^2 = \sigma_{T_Y}^2 + \sigma_{E_Y}^2. \tag{6.18c}$$

The variance of the total score is made up of two parts. The part $\sigma_{T_Y}^2 = \sigma_C^2 = (\Sigma \lambda_j)^2$, the square of the sum of the factor loadings, is the true-score variance of the total test score—variance due to the attribute of which the items are indicators. The part $\sigma_{E_Y}^2 = \sigma_U^2 = \Sigma \psi_j^2$, the sum of the *m* unique variances of the indicators, is the error variance of the total test score—a sum of variances due to the individual, idiosyncratic properties of each of the *m* indicators. (The beginning student is warned to be careful about the difference between summing loadings and squaring the sum for estimating σ_C^2, and simply summing the unique variances for σ_U^2.)

On this interpretation, the reliability coefficient defined in (5.4) becomes

$$\rho_r = \sigma_C^2/(\sigma_C^2 + \sigma_U^2). \tag{6.19}$$

This gives us a reliability coefficient based on the factor model—coefficient omega.[2] The coefficient is defined by

$$\omega = \sigma_C^2/\sigma_Y^2 = (\sum \lambda_j)^2/\sigma_Y^2, \tag{6.20a}$$

or

$$\omega = (\sum \lambda_j)^2/[(\sum \lambda_j)^2 + \sum \psi_j^2]. \tag{6.20b}$$

Omega is the ratio of the true-score variance of *Y* to the total variance of *Y* Here the true-score variance is interpreted as the variance due to the (common) attribute. The variance of *Y* is the sum of the true-score variance (i.e., common variance), and the error variance (i.e., unique variance). In a homogeneous set of items—with a single common factor—an equivalent expression is given by

$$\omega = 1 - (\sum \psi_j^2)/\sigma_Y^2. \tag{6.21}$$

The first form of the expression—(6.20)—has a version that applies to nonhomogeneous sets of items (see Chap. 9), whereas the second—(6.21)—requires homogeneity.

The coefficient omega—the reliability coefficient based on the parameters of the items in the factor model—can easily be estimated in applications. We just substitute estimates of the loadings $_1,\ldots,_m$ and the unique variances $\psi_1^2, \ldots, \psi_m^2$ in (6.20b). In estimating from a sample, the expressions (6.20a) and (6.21) will give identical results if Gy is estimated from the fitted covariances. A (slightly) different answer will be obtained if the sample variance of *Y*—computed from the total scores or by summing the elements of the sample covariance matrix—is mixed in with the estimated parameters of the model. This is acceptable and more convenient unless the fit is poor, in which case we should not be using the coefficient anyway.

Coefficient omega has been defined as the ratio of the variance due to the common attribute to the total variance of *Y*. It may be shown also that:

1. Omega is the square of the correlation between *Y* and the common factor *F*, or between *Y* and *TY* (or *C*), which is just *F* rescaled to be in the same units as F.
2. Omega is the correlation—not the squared correlation—between two test scores *Y* and *Y'* that have the same sum (or average) of their loadings and the same sum (or average) of their unique variances, and jointly fit the single common factor model, that is, are jointly homogeneous.
3. Omega is the square of the correlation between the total (or mean) score on the given *m* items and the mean score on an infinite set of items from a homogeneous domain of items of which the *m* items used in the test are a subset (see final section of this chapter).

Property 3 is consistent with a conception of the true score of the set of *m* items as the mean score on a test of infinite length.

Some accounts assume that the *m* items, and the infinitely many more, are parallel or true-score equivalent.[3] This is not necessary. The conclusion at this point is that coefficient omega captures the notion of the reliability of a test score. It measures the precision with which a homogeneous test measures the common attribute of its items.

For the SWLS data, from Table 6.3, the ULS estimates in Table 6.4 give

$$(\sum \hat{\lambda}_j)^2 = 1.290 + 1.104 + 1.148 + .952 + 1.185)^2 = 32.251$$

and

$$\sum \hat{\psi}_j^2 = .901 + 1.274 + 1.144 + 1.863 + 1.951 = 7.133.$$

Then using (6.20b) gives

$$\hat{\omega} = 32.251/(32.25 + 7.133) = 32.25/39.383 = .8189.$$

The sum of the elements in Table 6.3 gives $s_Y^2 = 39.388$ for the sample variance of *Y*. This is hardly different from the fitted value 39.383, obtained by summing estimated common variance and unique variances, so (6.20a) gives

$$\hat{\omega} = 32.251/39.388 = .8188.$$

COEFFICIENT ALPHA AND THE SPEARMAN-BROWN FORMULA

In the case in which the *m* items satisfy the true-score equivalence model, (6.2) becomes (6.11), on rescaling the common factor so that it has variance 2. This gives the simple decomposition of item variance into true variance and error variance in (6.13)—with distinct error variances σ_{Ej}^2 for the items. The reliability coefficient can then be written as a function of the covariance between any two items and the variance of the total score. That is, when σ_{jk} has a single value $\lambda^2 = \sigma_T^2$, or, equivalently, all items have equal factor loadings, omega in (6.20a) takes the simple form

$$\omega = m^2\lambda^2/\sigma_Y^2. \tag{6.22}$$

This can be expressed as

$$\omega = \sigma_{Ty}^2/\sigma_Y^2 = m^2\sigma_T^2/\sigma_y^2. \tag{6.23}$$

We can estimate it by substituting $\hat{\sigma}_{Tx}^2, \hat{\sigma}_Y^2$ in (6.23). In the SWLS example in the last section, $\hat{\sigma}_{Tx}^2 = 1.287$, $\hat{\sigma}_Y^2 = 39.382$, so

$$\hat{\omega} = 25 \times 1.287/39.382 = .8170,$$

which is slightly lower than the value—.8189—estimated under the more general hypothesis of homogeneity.

If we use the ULS estimator of σ_{Tx}^2, and use s_y^2 as the estimator of σ_y^2, then

$$\hat{\omega} = m^2\hat{\sigma}_{Tx}^2/s_Y^2 \tag{6.24}$$

where

$$\hat{\sigma}_T^2 = [1/m(m-1)]\sum\sum_{j \neq k} s_{jk}, \tag{6.25}$$

so

$$\hat{\omega} = [m/(m-1)][\sum\sum_{j \neq k} s_{jk}/s_Y^2. \tag{6.26}$$

Recalling that

$$s_Y^2 = \sum_j \sum_k s_{jk},$$

so

$$\sum_{j \neq k} \sum s_{jk} = s_Y^2 - \sum_j s_{jj},$$

we see that

$$\hat{\omega} = [m/(m-1)][1 - \sum_j s_{jj}/s_Y^2]. \tag{6.27}$$

For our example, Table 6.3 gives

$$\begin{aligned}\hat{\omega} &= (5/4)[1 - (2.566 + 2.493 + \cdots + 3.356)/39.382] \\ &= (5/4)(1 - 13.646/39.384) \\ &= .8170,\end{aligned}$$

as in the previous computation.

The estimate of reliability given in (6.27) is very well known. We can define an analogue for it in the population, namely,

$$\alpha \equiv [m/(m-1)][1 - \sum_j \sigma_{jj}/\sigma_Y^2]. \tag{6.28}$$

The identity sign draws attention to the fact that the expression on the right actually defines what is universally referred to as coefficient alpha. A number of different ways to motivate the estimation of this coefficient have been given in the literature. Coefficient alpha was first given (denoted L3) by Louis Guttman in 1945. He showed that it is a lower bound in the population to the reliability coefficient of the test score. Coefficient alpha is often incorrectly attributed to a paper by Cronbach in 1951.[4] In view of Cronbach's contributions to our understanding of this coefficient, it is referred to here as the *Guttman-Cronbach alpha* or *G-C alpha*.

It may be shown that G-C alpha is a lower bound to coefficient omega—generally underestimating it somewhat. They are equal if and only if the population of interest can be described by the true-score equivalence model, that is, if and only if the items fit the single-factor model with equal factor loadings. That is,

$$\omega \geq \alpha, \tag{6.29}$$

and alpha becomes the same as omega if and only if σ_{jk} is constant for all $j \neq k$.

***Proof:

$$\omega - \alpha = \left(1/\sigma_Y^2\right)\left[\left(\sum_j \lambda_j\right)^2 - \left\{m/(m-1)\right\}\left\{\left(\sum_j \lambda_j\right)^2 - \sum_j \lambda_j^2\right\}\right].$$

$$\omega - \alpha = \left[m/(m-1)\right]\left(1/\sigma_Y^2\right)\left[\sum_j \lambda_j^2 - \left(\sum_j \lambda_j\right)^2 / m\right]$$

$$= \left[m/(m-1)\right]\left(1/\sigma_Y^2\right)\left[Var\left(\lambda_j\right)\right],$$

which is positive, and zero if and only if $j = k$ for all j,k.[5] $$$

It is, in fact, difficult to invent a homogeneous population structure in which alpha is a very poor lower bound to omega, or to find empirical examples in which the estimate of alpha is very much lower than that of omega. Part of the case for

estimating omega itself rather than bounding it—"underestimating" it—by G-C alpha is that omega comes as a by-product of the factor analysis, which checks whether the items form a homogeneous set. If they do not, at least to a good approximation, it is not appropriate to form a total test score. If the evidence shows that the items are not only homogeneous but also true-score equivalent, then G-C alpha is an estimate of omega. But at this point we can easily calculate omega anyway.

Attempts have been made to give meaning to coefficient alpha, and to its estimate from a sample, when the items are not true-score equivalent. For example, Cronbach showed that coefficient alpha gives the average of split-half reliability coefficients, computed over all possible splits, supposing they have equal probability of being chosen by the investigator. This may be of some theoretical interest. It is often referred to as the "internal consistency" reliability of a test, but no clear meaning has ever been given for the notion of "internal consistency," and the terminology is not recommended here.

In the further special case of parallel items, (a) the factor loadings are equal and (b) the unique variances are equal. In this case, all the interitem correlations (as well as the item covariances) are equal to a constant value in the population, namely,

$$\rho_1 = \sigma_{Tx}^2/(\sigma_{Tx}^2 + \sigma_{Ex}^2). \tag{6.30}$$

We can think of ρ_1 as the reliability of one item—the same for any one of them. Then in this case

$$\omega = \alpha = \rho_m, \tag{6.31}$$

where

$$\rho_m = m\sigma_{Tx}^2/(m\sigma_{Tx}^2 + \psi^2). \tag{6.32}$$

This expression shows that when every item has the same true and error variance, the reliability of a test of *m* items is a simple increasing function of the number of items. We can rewrite (6.32) as

$$\text{S-B} \quad \rho_m = m\rho_1/[(m-1)\rho_1 + 1]. \tag{6.33}$$

This is an expression for the reliability of a test of *m* parallel items or subtests, from the reliability p of just one, given by (6.30).

*** To see this, note that

$$\rho_m = m\sigma_{Tx}^2/[(m-1)\sigma_{Tx}^2 + \sigma_{Tx}^2 + \psi]$$
$$= \frac{m[\sigma_{Tx}^2/(\sigma_{Tx} + \psi^2)]}{(m-1)[\sigma_{Tx}^2/(\sigma_{Tx}^2 + \psi)] + 1}. \quad \$\$\$$$

The expression (6.33) is traditionally known as the *Spearman-Brown prophecy formula* (hence the added "S-B").6 For the special case where $m = 2$, it was derived, independently, by Spearman and by Brown in 1910, as a formula to "prophesy" the increase in reliability that could be obtained if one were able to double the length of a (sub) test of known reliability p. If we have just one subtest, with a reliability coefficient ρ_1, then the reliability of a test of double length is

$$\rho_2 = 2\rho_1/(\rho_1 + 1). \tag{6.34}$$

This is the formula obtained by Spearman and by Brown for the obtainable reliability if we add a parallel subtest. This was then generalized to (6.33). The subscript *m* is attached in (6.33) to draw attention to the fact that this is the reliability coefficient of a test of *m* items, or, possibly, of *m* subtests. The expression (6.33) gives the reliability that we could get by adding *m* 1 more (parallel) subtests to a subtest of known reliability ρ_1.

Recall from the preceding section that the SWLS data have a lower GFI if we suppose the items are parallel in the population sampled. With this assumption, (6.32) gives us

$$\hat{\omega} = \hat{\rho}_m = (25 \times 1.287)/(25 \times 1.287 + 5 \times 1.442) = .8169,$$

on substituting the ULS estimates $\hat{\sigma}_{Tx}^2 = 1.287$, $\hat{\psi}^2 = 1.442$ given in the preceding section. Alternatively, to use (6.33), we can get the average, .476, of the sample correlations of the five items, as given in Table 6.6. Then (6.33) gives

TABLE 6.6
Satisfaction With Life—*SD*s and Correlations

SD	*1*	*2*	*3*	*4*	*5*
1.602	1.	.617	.592	.448	.486
1.579	.617	1.	.518	.322	.454
1.569	.592	.518	1.	.432	.457
1.664	.448	.322	.432	1.	.434
1.832	.486	.454	.457	.434	1.

$$\hat{\omega} = \hat{\rho}_m = (5 \times .476)/(4 \times .476 + 1) = .8196.$$

[At least slight differences between estimators using sample analogues of (6.32) and (6.33) can be expected because they average the information in the data in different ways.] Generally, there is no good reason to estimate reliability using the sample S-B $\hat{\rho}_m$ instead of G-C alpha. And if the single-factor model has been fitted, as I recommend, omega may as well be used.

A common application of the S-B 2 in (6.34) has been to a technique for estimating reliability known as *split-half* methods. A single test of *m* items—suppose *m* an even number—is administered, once only, and the items are split into two subtests, each of *m*/2 items, in some way. Items can be assigned at random, or odd-numbered items may be assigned to one subtest and even-numbered items assigned to the other. (Care is needed to prevent any resulting systematic pattern of assignment, such that the halves are not equivalent. In time-limited cognitive tests, for example, there must not be a split placing earlier items in one and later items in the other, because later items may be failed by not being reached.) The correlation between the half-test scores is the reliability of either. Then we can use the original Spearman-Brown prophecy formula (6.34) to approximate the reliability of the total score on the *m* items. This is generally not a safe, well-motivated procedure, as it does not make full use of the information in the data. It is still to be found in texts and possibly in applications, but cannot be recommended. There will certainly be variability in the reliability coefficient so estimated with the choice of split. If a split-half reliability is reported in the literature, and no better information is offered, we may reluctantly accept it, with some caution.

The reliability of the sum score of *m* items is given by if the items are homogeneous, equally by a if they are true-score equivalent, and equally by S-B *m* if they are parallel. From here on we reserve for the quantity given by (6.20), a for the quantity given by (6.28), and S-B *m* for the quantity given by (6.33), and similarly for the estimates obtained from the sample versions of these expressions.

It should be noted that omega and G-C alpha and their estimates from samples—but not the Spearman-Brown formula—will be altered by alteration of the scale of individual items/subtests. This is one reason why, with that one exception, we have used the covariance matrix of the items, not their correlation matrix. If we fit the single-factor model to the item correlation matrix instead of the covariance matrix, this is equivalent to standardizing each item score, substituting

$$Z_j = (X_j - \mu_{xj})/\sigma_{xj}$$

where *xj* and *xj* are the mean and *SD* of each item. The corresponding total score

$$S = \sum_j z_j$$

is a sum of standardized item scores. So implicitly, if we estimate omega or G-C alpha from a sample correlation matrix, we are estimating the reliability coefficient of a sum of standardized items. It would then be inconsistent to employ raw sum scores of the items. It is appropriate to mention this, because some readers may already be acquainted with common factor analysis as a psychometric technique. They would know that there is a sense in which, in general, the common factor model is independent of scale, and may be fitted to a sample correlation matrix—but not if we want coefficient omega.

The work in this chapter so far has rested on the rather restrictive model that requires all the items in the test to measure just one thing in common—to be *strictly homogeneous*, as we now call this property. In terms of item content, this requires that the items share one common attribute and that any further property is unique to each. In practice, this is a difficult ideal to attain. Fortunately, in applications, the requirement of strict homogeneity is indeed unnecessarily strict. Recall the example in Chapter 5 of two content-parallel test forms for measuring "intelligence," where it was remarked that the paired items jointly measured something they do not share with other items in their own set. Or note that the five SWLS items in Table 5.1 have been analyzed and found to be consistent with the hypothesis that they are homogeneous. But the fit is not perfect, and an examination of content suggests a division into three items measuring present satisfaction and two measuring satisfaction with the past. A set of items that share one general attribute, but form a number of small groups of items sharing further common properties, can often be treated as *essentially homogeneous*. This is because the effects of grouping are easily dominated by

the general attribute shared by all items, so the groups make a negligible contribution to the common part of the variance of the total score. This kind of case is illustrated by the content-parallel example. Chapter 9 shows how to treat the possibility that the general attribute we intend to measure yields a subclassification into a few groups, as in the case of the SWLS.

BINARY DATA

In the special case of binary items, coefficient alpha can be written as

$$\alpha = [m/(m-1)][1 - \{\sum_j \pi_j(1-\pi_j)\}/\sigma_Y^2] \tag{6.35}$$

[because $\pi_j(1-\pi_j)$ is the variance of a binary item]. As before, we have $_j$ for the probability of the keyed response to item j—the proportion of examinees in the population giving it. The corresponding estimator of alpha was given by Kuder and Richardson in 1937 as their equation (20) and it is conventionally referred to as KR-20. It will be written here as

$$\mathrm{KR}_{20} = [m/(m-1)][1 - \{\sum_j p_j(1-p_j)\}/s_Y^2]. \tag{6.36}$$

The Guttman-Cronbach alpha, both as an estimator and as a population coefficient, was actually a later generalization on KR 20. The coefficient KR20 was developed as a good estimator of the reliability coefficient under the assumption that all items are true-score equivalent. In the case of binary items, this is the hypothesis that

$$\sigma_{jk} = \pi_{jk} - \pi_j\pi_k$$

is constant for all pairs j, k, although the difficulty level j may vary.[7]

If it is also assumed that the items are strictly parallel (so that $_j =$ for all items), then KR20 reduces to the computationally convenient expression

$$\mathrm{KR}_{21} = [m/(m-1)][1 - \{\bar{Y}(m-\bar{Y}\}/(ms_Y^2)]. \tag{6.37}$$

This just needs the sample mean and sample variance of the total score. No item statistics are required. The computational convenience of KR21 may have been a motive for recommending it in the era before computers. It is now of only historical and perhaps some conceptual interest.[8]

A section of five items from the Law School Admission Test—items 11–15 of Section 6, to be referred to here as LSAT6—has been reanalyzed in a number of papers on psychometric theory. Unfortunately, the item stems are no longer available. The original data from 1,000 examinees can be summarized without any loss of information as in Table 6.7(a). The distribution of the total test scores is in Table 6.7(b). The distribution is skewed. Three examinees get a "perfect" 0, and 198 get a perfect 5.

We compute the proportions pj passing each item, as entered in Table 6.8, and the joint proportions pjk passing pairs of items, entered above the diagonal in the table. The sample item covariances are below the diagonal. The sample item variances are on the diagonal. Note that the mean of the total score is the sum of the pj values, and is equal to 3.818. (The student is encouraged to check some of the pj values from Table 6.7 by adding frequencies, and, similarly, to compute some joint frequencies of pairs, and a few covariances.)

From the item covariances, we have the total score variance

$$\sigma_{Ty}^2 = 1.0702,$$

from summing the 25 elements of the covariance matrix. (Or, alternatively, we may obtain this from the total test score distribution in Table 6.7.) Then G-C alpha—which is also KR20—is given by

$$\mathrm{KR}_{20} = (5/4)[1 - (.0702 + .2063 + \cdots + .1131)/1.0702] = .295.$$

Also,

$$\mathrm{KR}_{21} = (5/4)[1 - (3.818 \times 1.182)/(5 \times 1.0702)] = .196.$$

Fitting the single-factor model to the covariance matrix in Table 6.8 is a theoretically questionable procedure, but we use a computer program to do it anyway. This gives loadings, unique variances, and the discrepancy matrix in Table 6.9. The goodness of fit index

$$\text{GFI} = 1 - (.00019002/.158238) = .9988,$$

suggesting a satisfactory fit of the items to the single-factor model. Coefficient omega, by (6.20a), is

$$\omega = (.0605 + .1345 + \cdots + .0745)^2/1.0702 = .3068,$$

TABLE 6.7
LSAT Section 6 (LSAT6)

(a) Data

Index	*Test Score*	*Response Pattern Frequencies for Item* 1	2	3	4	5	*Observed Frequency*
1	0	0	0	0	0	0	3
2	1	0	0	0	0	1	6
3	1	0	0	0	1	0	2
4	2	0	0	0	1	1	11
5	1	0	0	1	0	0	1
6	2	0	0	1	0	1	1
7	2	0	0	1	1	0	3
8	3	0	0	1	1	1	4
9	1	0	1	0	0	0	1
10	2	0	1	0	0	1	8
11	2	0	1	0	1	0	0
12	3	0	1	0	1	1	16
13	2	0	1	1	0	0	0
14	3	0	1	1	0	1	3
15	3	0	1	1	1	0	2
16	4	0	1	1	1	1	15
17	1	1	0	0	0	0	10
18	2	1	0	0	0	1	29
19	2	1	0	0	1	0	14
20	3	1	0	0	1	1	81
21	2	1	0	1	0	0	3
22	3	1	0	1	0	1	28
23	3	1	0	1	1	0	15
24	4	1	0	1	1	1	80
25	2	1	1	0	0	0	16
26	3	1	1	0	0	1	56
27	3	1	1	0	1	0	21
28	4	1	1	0	1	1	173
29	3	1	1	1	0	0	11
30	4	1	1	1	0	1	61
31	4	1	1	1	1	0	28
32	5	1	1	1	1	1	298
Total							1,000

(b) Frequency Distribution

Total Test Score	*Frequency*
5	298
4	357
3	237
2	85
1	20
0	3

TABLE 6.8
LSAT6—Difficulties and Covariance Matrix

		Item				
Item	p_j	1	2	3	4	5
1	.924	.0702	.664	.524	.710	.806
2	.708	.0089	.2063	.418	.553	.630
3	.553	.0130	.0259	.2472	.445	.490
4	.763	.0050	.0120	.0231	.1808	.678
5	.870	.0021	.0132	.0089	.0142	.1131

TABLE 6.9
LSAT6—Spearman Analysis

	Loadings λ	Unique Variances ψ^2	Discrepancy Matrix (Sample-Fitted Covariance Matrix, $\mathbf{S} - \boldsymbol{\Sigma}$)				
1	.0605	.0665	.0	.0008	.0017	.0021	−.0024
2	.1345	.1882	.0008	.0	.0009	.0038	.0032
3	.1861	.2126	.0017	.0009	.0	.0012	.0050
4	.1174	.1670	.0021	.0038	.0012	.0	.0054
5	.0745	.1076	−.0024	.0032	.0050	.0054	.0

which is just slightly larger than KR20. Clearly the reliability is very low, and a longer test is needed. A sense of the limitations of this treatment of binary data—and the reason why it is admitted to be a "theoretically questionable procedure"—comes from the observation that the regression equations of the model are

$$\hat{x}_1 = .924 + .0605F$$
$$\hat{x}_2 = .709 + .1345F$$
$$\hat{x}_3 = .553 + .1861F$$
$$\hat{x}_4 = .763 + .1174F$$
$$\hat{x}_5 = .870 + .0745F.$$

The expected values $\hat{x}_j$ for the item scores are linear functions of the common factor F. Because the items are binary, these equations represent the probability of passing an item for an examinee with a given value of F. (Recall the introduction to Chap. 3.) One problem with the linear model is that for a small enough value of F the probability is negative, and for a large enough value it is greater than unity, which would be absurd. That is,

$$\text{if} \quad F < -p_j/\lambda_j \quad \text{then} \quad P\{X_j = 1|F = f\} < 0$$

and

$$\text{if} \quad F > (1 - p_j)/\lambda_j \quad \text{then} \quad P\{X_j = 1|F\} > 1.$$

In this data set, item 3 will have negative probabilities for $F < -.553/.186$, that is, $F < -2.97$, and item 1 will have probabilities greater than unity for $F > 1.256$. Using methods given in Chapter 7—as seen in equation (7.17)—we find that the lowest value of F in this data set is estimated as –3.795, which corresponds to the 0 total score, and the highest is 1.182, which corresponds to a total score of 5. Only the three examinees with zero total score yield an absurd probability value. We come back to this question later. A linear model is theoretically inappropriate for binary items, yet it can give an adequate overall

approximation to their behavior in many applications. Note that KR20 assumes a linear model, and is an approximation in precisely the same way.

If we fit the true-score equivalence model to these data, from Table 6.8 we get the ULS estimate of the true score variance, as the simple average of the covariances (below the diagonal), namely,

$$\begin{aligned}\hat{\sigma}^2 = (1/10)[&.0089\\ &+ .0130 + .0259\\ &+ .0050 + .0120 + .0231\\ &+ .0021 + .0132 + .0089 + .0142] = .01263.\end{aligned}$$

The fitted covariance and discrepancy matrices are given in Table 6.10, with a goodness of fit index

$$\text{GFI} = 1 - (.0009719/.158494) = .9939.$$

Fitting the parallel items model requires a common value of the diagonal elements of the fitted matrix. By ULS this is the average of the sample variances, (1/5) (.0702 + .2063 + .2472 + .1808 + .1131) = .1635, giving .1635 .0126 = .1509 for the error variance. Then, by (6.32),

$$\rho_m = (5 \times .0126)/(5 \times .0126 + .1509) = .2945.$$

TABLE 6.10
LSAT6—True-Score Model

(a) Fitted Covariance Matrix

.0702	.0126	.0126	.0126	.0126
.0126	.2063	.0126	.0126	.0126
.0126	.0126	.2472	.0126	.0126
.0126	.0126	.0126	.1808	.0126
.0126	.0126	.0126	.0126	.1131

(b) Discrepancies

.0	−.0037	.0004	−.0076	−.0105
−.0037	.0	.0133	−.0006	.0006
.0004	.0133	.0	.0105	−.0037
−.0076	−.0006	.0105	.0	.0016
−.0105	.0006	−.0037	.0016	.0

TABLE 6.11
LSAT6—Parallel Items Model

(a) Fitted Covariance Matrix

.1635	.0126	.0126	.0126	.0126
.0126	.1635	.0126	.0126	.0126
.0126	.0126	.1635	.0126	.0126
.0126	.0126	.0126	.1635	.0126
.0126	.0126	.0126	.0126	.1635

(b) Discrepancies

−.0933	−.0037	.0004	−.0076	−.0105
−.0037	.0428	.0133	−.0006	.0006
.0004	.0133	.0837	.0105	−.0037
−.0076	−.0006	.0105	.0173	.0016
−.0105	.0006	−.0037	.0016	−.0504

The fitted covariance and discrepancy matrices are given in Table 6.11. The GFI = 1 – (.02135/. 1584) = .8653. This suggests that the items are not parallel, though they closely approximate true-score equivalence, as well as the more general single-factor model.

The classical factor model (6.2) is a linear model. The item scores (i.e., their expected values) are linear functions of the factor attribute. We return to the problem, noted earlier, of fitting a linear factor model to binary variables, when we consider nonlinear factor models—item response models—in Chapter 12. Meanwhile, it is enough to note that although the classical linear model (6.2) can only approximate the behavior of binary items, summation as in (6.17a) tends to cancel the effects of nonlinearity. Consequently, coefficient omega (bounded by G-C alpha, i.e., KR20) still gives the reliability coefficient—defined as a ratio of common (attribute) variance to total variance. There is a more important limitation. In general the error variance (unique variance) of the total test score from a set of binary items cannot be independent of the test score. The error variance we obtain from the reliability coefficient is a general approximation to the error variances actually found along the scale of the test. An advantage of item response models is that they provide a functional relationship between the error variance and the true score (see Chaps. 12–15).

SOME PRINCIPLES OF GENERALIZABILITY

[This entire section can be omitted without loss of continuity.] As already hinted, there is a close connection between reliability and certain conceptions of *generalizability* from the items we obtain to further items that we imagine constructing. It was remarked in an earlier section that coefficient omega is the squared correlation between the total (or mean) score from m items and the mean score of items forming a test of infinite length. In effect, the true score is the mean score of the infinitely long test. Such a hypothetical limiting score is known as the *domain* score. This conforms to general mathematical usage. The domain is the set of elements—"values" in a wide sense—to which a variable is limited. In the application here, the elements of the domain are items. The domain of items is sometimes referred to as the *behavior domain* or the *universe of content*, or the *universe of admissible measurements*. Thus, the set of all items under consideration is the *item domain*, and the mean score on that set is the *item domain score*. (Note that we must now work with the mean of the item scores because the variance of the total test score will increase unboundedly as items are added.)

One generally necessary condition for the treatment of generalizability is that the items of the domain measure just one attribute in common. The entire domain should fit the single-factor model (or a corresponding unidimensional item response model—Chaps. 12 and 13). It is also necessary in some models to assume that the given set of m items has been representatively sampled from the infinite set of items that could be written and given to the respondents. Random sampling of items is generally not possible in applications. These fundamental and untestable assumptions constitute an idealization that could only be approximated in any application, and in some cases could not be realized, even approximately. Under these idealizing conditions the domain mean score both defines and determines the true score. It defines it in the sense that the set of items defines the attribute being measured by the property they have in common. It determines it in the sense that an infinite set of items measures the attribute precisely. Then the errors of measurement in using the practically available set of items are due to having a limited sample of items to represent the attribute.

The best known treatment of generalizability is based on a linear variance components model. This will be familiar to readers who are well grounded in the standard applications of analysis of variance to mixed-and random-effect models. It is appropriate to refer to it as Cronbach generalizability theory, because Cronbach and his associates have given it its most systematic development. An alternative treatment rests on the factor analysis of covariance matrices as in an earlier section of

this chapter. Each of these treatments has advantages and limitations. We consider only some basic principles of Cronbach generalizability here, with a few remarks about the more extensive developments. The factor-analytic work will involve some deliberate recapitulation.

Suppose we have m measurements/item scores on n respondents, x_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, m$. We think of the items as coming from an infinite domain of items that would, if exhaustively tested, be homogeneous, fitting the Spearman factor model. We wish to obtain coefficients that quantify the relation between the actual m measurements and the conceivable domain measurements. In the Cronbach treatment the analysis of the data for this purpose is conventionally referred to as a *generalizability study*—commonly denoted "G-study." (In other contexts it would be a *calibration study*, and may be part of a test-development study.) For example, if we were using the factor-analytic work already covered, in the generalizability study we would compute coefficient omega or G-C alpha and would interpret it as the squared correlation between the item-sample mean score and the domain mean score. The results from the generalizability study can then be used in a *decision study*—commonly denoted "D-study"—to assess the expected measurement error in a new application. For example, having found that m items do not give a large enough G-C alpha, we may decide how many more to add to reduce measurement error sufficiently, by using the Spearman-Brown formula.

Before turning to the Cronbach treatment we review the factor-analytic treatment. To link the two we rewrite the factor model (6.2) as

$$X_{ij} = \mu_j + \lambda_j F_i + E_{ij}, \tag{6.38}$$

in which the subscript i has been added. As before, $_j$ and $_j$ are population parameters representing the mean (= difficulty) and the factor loading (= sensitivity, i.e., discriminating power) of item j. Attaching the subscript i to the variables complicates the previous account, but brings it into line with the conventional treatment of analysis of variance, to be given soon. Instead of just writing X_j as the score of a randomly sampled subject, we regard the n variables X_{ij} as the scores of n respondents to be independently randomly drawn from the population. This is to be distinguished from the values x_{ij} that we have after a particular sample has been drawn. [Formally trained students of mathematical statistics will recognize that this is just to state that X_{ij}, F_i, E_{ij}, $i = 1,\ldots, n$, are n independently and identically distributed random variables. Students whose background is in applied statistics may ignore this formal statement, but should accept that (6.38) is still a model for the population, not a description of a given sample.]

Accordingly, we recall (6.17)–(6.19), and we write corresponding expressions for means instead of total test scores. These are

$$M_i = (1/m)\sum X_j = \sum_j \mu_j + [(1/m)\sum_j \lambda_j)F_i + (1/m)\sum_j E_{ij} \tag{6.39}$$

or

$$M_i = \mu_\bullet + \lambda_\bullet F_i + E_{\bullet i} \tag{6.40}$$

with

$$\sigma^2_M = (\lambda_\bullet)^2 + \psi^2_\bullet, \tag{6.41}$$

where the dot replacing a subscript here indicates an average over that subscript. [Note that $(.)^2$ represents the mean factor loading squared, whereas $^2_.$ represents the mean of the unique variances.] Recalling also (6.20b), we have for coefficient omega

$$\omega = (m\lambda_\bullet)^2/(m\lambda_\bullet)^2 + \psi^2_\bullet). \tag{6.42}$$

This is the squared correlation between M_i and F_i.

The fundamental point is that F_i, the common factor defined by the homogeneous domain from which the items are taken, is a measure of the attribute that the infinity of items have in common, and that they measure precisely. The larger the number of items (with nonzero loadings) that we use, the more precisely we measure the attribute. In the limit the precision approaches perfection. Coefficient omega is the squared correlation between the observable mean score of m items and the domain mean score. It is also the proportion of variance of the mean score of the given items that is due to the domain score. Assuming homogeneity, we may call omega the *coefficient of generalizability* from the given item set to the domain.

Note that in this model we do not need to suppose that the m items are randomly or at least representatively sampled. The size of omega depends on the parameters of the items chosen. Remaining items could all have smaller or larger loadings or unique variances, without altering generalizability. As we have seen, if their required conditions on the item covariance matrix are satisfied, G-C alpha or the Spearman-Brown formula will yield the same reliability coefficient as coefficient omega, and accordingly these will serve as coefficients of generalizability. Otherwise, they are underestimates. It is not yet common practice in test development to check whether the items are true-score equivalent or parallel. It is recommended that this should always be done if possible. When their respective conditions are satisfied, these three coefficients measure

generalizability, under the assumption that the item domain is homogeneous. If the given items are true-score equivalent, we can use G-C alpha without supposing that the remaining items in the domain are also true-score equivalent. It is sufficient that the given items have equal factor loadings. The remaining items need not. Similarly, it is not necessary to assume that further items in the domain are parallel if the given items are.

We can check that the items we actually construct are homogeneous, but of course we cannot know that the entire item domain is homogeneous. In applications, great care is needed in conceptualizing the attribute that is the goal of measurement, so that we have a clear prescription for the items that will be indicators of it. It is also necessary to be able to imagine writing more items of the same kind. This is the question of *content validity*, to which we turn in Chapter 10. Some accounts unnecessarily restrict generalizability theory to the addition of further parallel items.

An advantage of the factor-analytic treatment of generalizability is that it allows the selection of subsets of items from the test development—calibration or generalizability—study to make a shorter test having optimal reliability/generalizability (see Chap. 7). Attempts to extrapolate from the given items to the behavior of a longer, but still finite, test must be conjectural, because they depend on the unknown loadings—sensitivities—of the items that may be added.

A good starting point for the introduction of Cronbach generalizability theory is the point of overlap with the factor analytic treatment—namely, the model for true-score equivalent items. The true-score equivalence model is obtained, as we have seen, by equating factor loadings in (6.2) to yield

$$X_{ij} = \mu_j + T_{Xi} + E_{ij}, \tag{6.43}$$

where $TXi = Fi$ is the true score measured in the units of the (true-score equivalent) items. Recall that we allow the items to have distinct error variances σ^2_{Ej}. To move toward applications of analysis of variance, we express each item mean as

$$\mu_j = \mu + \delta_j. \tag{6.44}$$

Here δ_j represents the deviation of the difficulty of the item from , which is the average of the m difficulties (so $\Sigma\, \delta_j = 0$), and the model becomes

$$X_{ij} = \mu + \delta_j + T_{Xi} + E_{ij}. \tag{6.45}$$

According to (6.45), four components contribute to the score of randomly drawn respondent i on item j:

1. A parameter represents the fixed contribution of the average difficulty of these m items.
2. A parameter δ_j represents the difficulty of item j as a deviation, positive or negative, from the mean difficulty.
3. The true score Ti corresponds to the domain score of the respondent.
4. The response of examinee i to item j is subject to a random "error" term Eij.

As before, we have the conditions that true scores and errors, and errors from different items, are uncorrelated, so we get the variance decomposition

$$\sigma^2_j = \sigma^2_T + \sigma^2_{Ej},$$

and, for item covariances,

$$\sigma_{jk} = \sigma^2_T.$$

This gives us the testable structure for the covariance matrix discussed earlier. The test mean score is given by

$$M_i = \mu + \delta_\bullet + M_{Ti} + E_{\bullet i} \tag{6.46}$$

where $\delta_\bullet$ are averages of item difficulties and errors respectively and $MTi = (1/m)Ti$ is the mean true score. The variance of M is

$$\sigma^2_M = \sigma^2_T + \sigma^2_{E\bullet} \tag{6.47}$$

where $\sigma^2_{E\bullet}$ is the average of the error variances and $\sigma^2_T = (1/m^2)\sigma^2_{Ty}$ is the variance of the mean true score. The covariance of M and T_X is σ^2_{MTx}, so the squared correlation between M and M_T is

$$\rho^2_{MTx} = \sigma^2_{Tx}/\sigma^2_M = (m\sigma^2_{Tx})/(m\sigma^2_{Tx} + \sigma^2_{E\bullet}). \qquad (6.48)$$

This coefficient is the reliability, as before. As now reinterpreted, it is the generalizability of the test mean score—or, indeed, of the total score, because this coefficient is independent of the scale. Given sample observations *xij* of *Xij*, we arrive—as in (6.24) through (6.27)—at the estimate

$$\hat{\rho}^2_{MTx} = [m/(m-1)][1 - (\sum_j s_{ij})/s^2_Y)]. \qquad (6.49)$$

This is just the conventional estimate of G-C alpha in (6.28).

Consider the constructed data set in Table 6.12, with five items and 20 respondents. These data could arise as Likert scores on an attitude scale, or as ratings of the 20 subjects on five named traits by an observer, or in a number of other ways. These data give the covariance matrix in Table 6.13. Summing the elements of the covariance matrix gives 10.516 for the variance of the sum score *Y* and 10.516/25 = .4206 for the variance of the mean score. Fitting the single-factor model by ULS gives the factor loadings and unique variances in Table 6.14(a), whereas fitting the true-score equivalence model gives the corresponding results in Table 6.14(b).

TABLE 6.12
Constructed Quantitative Item Set

	Item					
Respondent	1	2	3	4	5	$x_{i\bullet}$
1	3	2	3	2	3	2.6
2	5	3	3	2	2	3.0
3	2	2	1	2	2	1.8
4	5	4	3	3	2	3.4
5	2	2	2	1	2	1.8
6	3	5	4	3	3	3.6
7	1	4	3	3	3	2.8
8	3	2	4	3	2	2.8
9	4	2	3	3	3	3.0
10	3	4	3	4	3	3.4
11	4	4	3	4	4	3.8
12	4	2	3	2	3	2.8
13	6	4	4	3	4	4.2
14	3	4	3	3	3	3.2
15	5	4	3	3	2	3.4
16	2	2	2	1	2	1.8
17	3	5	4	3	3	3.6
18	1	4	3	3	3	2.8
19	3	2	4	3	2	2.8
20	4	2	3	3	3	3.0
$x_{\bullet j}$	3.3	3.15	3.05	2.7	2.7	($x_{\bullet\bullet}$ = 2.98)

Both models give an acceptable fit to the data. Summing the loadings in Table 6.14(a), squaring, and dividing by the variance of *Y* gives a coefficient omega equal to 2.702V10.516 = .694. We can also use (6.20a) for the true-score equivalence analysis, giving $\hat{\omega} = (5 \times .538)^2/10.516 = .688$. Using the estimator for G-C alpha gives

$$\hat{\alpha} = (5/4)[1 - (1.8 + 1.292 + .576 + .642 + .432)/10.516] = .686,$$

agreeing well enough with the previous estimate.

We now reexamine the model (6.45). Readers well grounded in analysis of variance will recognize that it is, indeed, a linear

model for X_{ij} with four components—the grand mean, , the fixed effect of item difficulty, δ_j, the random effect of respondent attribute value, Tx_i, and a *residual*, E_{ij}. The residual consists of an interaction between the item and the respondent and anything we might further regard as "error." If we obtained repeated measurements for the respondents on each item, we might separate out an error of replication from the interaction, which is thought of as the effect of the unique properties of the item on a random examinee's response.

TABLE 6.13
Covariance Matrix—Constructed Item Set

1.800	.163	.353	.253	.147
.163	1.292	.360	.521	.310
.353	.360	.576	.332	.174
.253	.521	.332	.642	.274
.147	.310	.174	.274	.432

TABLE 6.14
Spearman and True-Score Equivalence Analyses—Constructed Quantitative Data Set

(a) Spearman		*(b) True-Score Equivalence*	
λ	ω	λ	ω
.382	1.654	.538	1.511
.688	.818	.538	.993
.541	.283	.538	.287
.694	.160	.538	.353
.397	.274	.538	.143
GFI = .990		GFI = .970	

The analysis of variance (ANOVA) can be set out as follows: Standard ANOVA algebra requires us to express each observed *xij* in terms of four components.

$$x_{ij} = x_{\cdot\cdot} + (x_{\cdot j} - x_{\cdot\cdot}) + (x_{i\cdot} - x_{\cdot\cdot}) + (x_{ij} - x_{\cdot j} - x_{i\cdot} + x_{\cdot\cdot}), \qquad (6.50)$$

which are sample counterparts of the four components in the model. (Again we use a dot for a mean over a subscript.) From (6.50) it may be shown by the standard algebraic substitutions of analysis of variance theory that

$$SS_{Tot} = SS_{\delta} + SS_{Tx} + SS_{E} \qquad (6.51)$$

where

$$SS_{Tot} = \sum_i \sum_j (x_{ij} - x_{\cdot\cdot})^2 = \sum_i \sum_j x_{ij}^2 - nmx_{\cdot\cdot}^2 \qquad (6.52a)$$

$$SS_{D} = n\sum_j (x_{\cdot j} - x_{\cdot\cdot})^2 = n\sum_j x_{\cdot j}^2 - nmx_{\cdot\cdot}^2 \qquad (6.52b)$$

$$SS_{Tx} = m\sum_i (x_{i\cdot} - x_{\cdot\cdot})^2 = m\sum_i x_{i\cdot}^2 - nmx_{\cdot\cdot}^2 \qquad (6.52c)$$

$$SS_{E} = \sum_i \sum_j (x_{ij} - x_{\cdot j} - x_{i\cdot} + x_{\cdot\cdot})^2$$

$$= SS_{tot} - SS_{\delta} - SS_{Tx}. \quad (6.52d)$$

By the usual convention, SS denotes a *sum of squares* of deviations about the means that appear in the expressions. This is the fundamental analysis of sums of squares that underlies the "analysis of variance." Conventional ANOVA then defines corresponding *mean squares* (MS), which are estimates of combinations of variances, and it uses the algebra of expectations to obtain their expected values. In this case these are:

$$MS_{\delta} = SS_{\delta}/(m-1) \Rightarrow \sigma_E^2 + n[(\sum_i \delta_j^2)/(m-1)], \quad (6.53a)$$

$$MS_{Tx} = SS_{Tx}/(n-1) \Rightarrow \sigma_E^2 + m\sigma_{Tx}^2, \quad (6.53b)$$

$$MS_E = SS_E/[(m-1)(n-1)] \Rightarrow \sigma_E^2. \quad (6.53c)$$

[The symbol "$\Rightarrow$" is to be read here as "is an unbiased estimate of." Note that because item difficulty has fixed effects, in (6.53) we do not have a population variance σ_{δ}^2 to be estimated.] This is a *mixed model*, with a random factor T_X—the random effect of respondent attribute—and a fixed factor δ_j, the fixed effect of the difficulty of each of the chosen items. We are not able to get a separate estimate of the error variance of each of the m items. But if we interpret σ_E^2 as the average of these error variances, the generalizability coefficient in (6.48) becomes

$$\rho_{MT}^2 = (m\sigma_{Tx}^2)/(m\sigma_{Tx}^2 + \sigma_E^2). \quad (6.54)$$

Then from (6.53), MS_{Tx} is an unbiased estimate of $m\sigma_{Tx}^2 + \sigma_E^2$, and MS_{Tx} MS_E is an unbiased estimate of $m\sigma_{Tx}^2$. So a satisfactory estimate of ρ_{MTx}^2, the generalizability coefficient, is given by

$$r_{MT(A)}^2 = (MS_{Tx} - MS_E)/MS_{Tx}. \quad (6.55)$$

[The subscript (A) is intended to denote "by ANOVA."] This is a classical formula for reliability by ANOVA, with a very long history.[9] It may be shown that $r_{MT(A)}^2$ and G-C alpha in (6.49) are algebraically equivalent, and must give the same answers. (The latter is based on a slightly more general model, as it allows the error variances of the items to be distinct.) The practical implication of this equivalence is that the researcher who has access to a computer program for ANOVA can obtain the sums of squares and mean squares as standard output, and use (6.55) instead of (6.49), as employed in specialized computer programs for G-C alpha. Whatever choice is made, it is again strongly recommended that the sample item covariances be at least examined for possible departures from the expected structure, which requires equal item covariances, with possibly unequal variances.

A conventional analysis of variance of the data in Table 6.12, set out in the standard form of presentation, is given in Table 6.15. From the mean squares for T_X and for E, we obtain $r_{MT(A)}^2$ in agreement, as we expect, with G-C alpha and with omega estimated earlier under true-score equivalence.

Experienced ANOVA users may be puzzled by the absence of the usual F ratios and tests of significance. In truly doubtful cases, we could obtain these, and could test to see if there is significant variance due to the attribute, or if the items vary significantly in difficulty. It is to be hoped that the first significance test will not be needed. It would be unfortunate if we have absolutely failed to measure anything. The second significance test may become of interest in an alternative model for randomly selected items.

TABLE 6.15
ANOVA—Constructed Quantitative Data Set

Source	*SS*	*df*	*MS*
δ	5.86	4	1.465
T_x	39.96	19	2.103
$T_x\delta$	50.14	76	0.660
Total	95.96		

Model (6.45) yields a coefficient of generalizability that we can now estimate in three ways, namely, as omega, as alpha, and by the ANOVA formula (6.55) if the covariances have the appropriate structure. We now ask what this tells us about the item domain. It is important to note that this coefficient is the squared correlation between the domain mean score and the

score on the given items. It is not the expected squared correlation between the domain score and the mean score from a sample of *m* items drawn at random from the domain. It measures generalizability without the assumption that the remaining items in the domain are also true-score equivalent. It is enough that the domain is homogeneous.

An advantage of the structural (factor analysis) treatment over the ANOVA treatment of generalizability is that even when the items are true-score equivalent, we get separate estimates of the error variances of the *m* items. We might wish, for future applications, to choose a subset of the *m* items to make a shortened form of the test on the basis of the test development phase—the generalizability study. Given separate error variances, we can choose subsets with optimal generalizability (see Chap. 7).

On the other hand, if we wish to extrapolate, so to speak, from the generalizability of the given *m* items to the generalizability we would expect to get by adding a number of further items, we need more restrictive assumptions. This is because we cannot know the error variances or factor loadings of items we do not yet have. In practice, as we face the actual task of writing further items, we hope to have a clear enough conceptualization of the attribute to write more items measuring the same thing. But we certainly do not know how to write them to be equally sensitive to the attribute or to have equal unique variability. In applying the formulas in the direction of extrapolation to unwritten items, it is necessary to assume that the average factor loading and average unique variance of the added items equal those of the *m* given items. Intuitively, one sees that the larger the number of added items, the safer this assumption becomes. Even so, in applications, there could be an unpleasant surprise if, as may often be the case, the best indicators of the attribute have been written first, and writing more good items proves difficult.

*** Proof of equivalence of ANOVA formula to G-C alpha: (I cannot trace lines of proof in the original literature on this. I believe the following is necessary.)

$$\begin{aligned} SS_{Tot} - SS_D &= \sum\sum (x_{ij} - x_{\bullet\bullet})^2 - n\sum (x_{\bullet j} - x_{\bullet\bullet})^2 \\ &= \sum\sum x_{ij}^2 - n\sum x_{\bullet j})^2 \\ &= \sum\sum (x_{ij} - x_{\bullet j})^2 \\ &= SS_E + SS_{Tx}. \end{aligned}$$

It follows that

$$\sum_j s_j^2 = (m-1)MS_E + MS_{Tx},$$

yielding equivalence. $$$

So far we have done little more than reexamine classical reliability theory from the point of view that a true score is a domain score and therefore "the" reliability coefficient is a generalizability coefficient. Now we consider an alternative model—the *random items model* for the observations *xij*. The random items model requires the strong assumption that the *m* items are in effect randomly sampled from the item domain—that every item in the domain has an equal probability of being chosen. In this case the model for the score on the *j*th randomly sampled item is

$$X_{ij} = \mu + D_j + T_i + E_{ij}, \tag{6.56}$$

where *Dj* is the random difficulty of item *j* as a deviation from . In this model is the mean difficulty of the entire item domain, whereas in the fixed items model it was the mean difficulty of the *m* given items. Here we must suppose that the variance σ_E^2 of the error-term *Eij* (which again includes interaction) is independent of the item sampled. The *m* items are drawn with independent random difficulties whose variance is σ_D^2.

Then the variance of the item score *Xij* of a randomly drawn respondent given a randomly selected item consists of three components—one due to item difficulty, one due to respondent domain score, and one comprising error, that is,

$$\sigma_X^2 = \sigma_D^2 + \sigma_T^2 + \sigma_E^2. \tag{6.57}$$

The important consequence is that under the model of random item selection, and random item difficulty, the variance of the item difficulty is, in effect, included in the variance of the error of measurement. Part of the error with which the respondent's mean score on m items determines that person's domain mean score depends on the difficulties of the sampled items. The squared correlation between a randomly drawn item and the domain mean score is given by

$$\rho_{XT}^2 = \sigma_T^2 / [\sigma_T^2 + (\sigma_E^2 + \sigma_D^2)], \tag{6.58}$$

and the squared correlation between the mean test score M_i on a test of m items and the domain mean score is

$$\rho^2_{MT} = (m\sigma^2_T)/[m\sigma^2_T + (\sigma^2_E + \sigma^2_D)]. \tag{6.59}$$

The expression (6.59) is actually for the expected squared correlation between the domain score and the score from a random sample of m items.

Standard ANOVA algebra gives the same results as in the fixed items model, (6.50) through (6.53), except that (6.53a) is replaced by

$$\mathrm{MS}_D = \mathrm{SS}_D/(m-1) \Rightarrow \sigma^2_E + n\sigma^2_D. \tag{6.60}$$

Then

$$(\mathrm{MS}_T - \mathrm{MS}_E)/m \Rightarrow \sigma^2_T, \tag{6.61a}$$

$$\mathrm{MS}_E \Rightarrow \sigma^2_E, \tag{6.61b}$$

as before, and

$$(\mathrm{MS}_D - \mathrm{MS}_E)/N \Rightarrow \sigma^2_D. \tag{6.61c}$$

Accordingly, good estimates of the reliability/generalizability of a randomly drawn item and of the mean score of m such items are obtained by substituting these estimates in (6.58) and (6.59). In the example of Table 6.12, with the ANOVA in Table 6.15, we have $\hat{\sigma}^2_E = .660$ and $\hat{\sigma}^2_T = .289$ as before, and $\hat{\sigma}^2_D = (1.465 - .660)/20 = .040$, giving .289/(.289 + .660 + .040) = .292 for the generalizability of a random item, and 1.443/(1.443 + .660 + .040) = .673 for the generalizability of the mean score from m randomly drawn items. In this example, the items do not vary much in difficulty. In other cases, widely dispersed difficulties could add a great deal to the error of measurement in this model.

The application of the random items model and the resulting variances, and of the generalizability coefficients and their ANOVA estimates, would be motivated by a situation in which we actually intend to give independently drawn samples of items to different respondents. In a subpopulation of examinees having the same domain score, a set of m items with random mean difficulty $D_.$ can take the observed mean score above or below the domain mean score, as a particularly easy or difficult set of items happens to be assigned to an examinee by chance sampling of items. In the practice of test construction and measurement, it would not be common practice to assign distinct samples of items to different examinees under the assumptions of the random items model. One may question the motivation of such a procedure. In the commonest situation, m items of known structure are given to all examinees in any population of interest. In major testing programs, a usual situation is one in which we have a large *pool* or *bank* of items, and for reasons of security of the testing process against unethical access to or use of the items, distinct subsets of items from the pool are given to examinees and they need to be comparably scored. In such a case, by methods described in Chapter 16, it will usually be possible to obtain parameters characterizing each item in the entire set, corresponding to difficulties and factor loadings. These can then be used to obtain comparable estimates of the attribute from different sets of items—a process known as *test equating*. It is also possible by these methods to assess the variance of an error of measurement that will not be inflated by the inclusion of chance effects of receiving items of varying difficulty.

A more likely application of the random items model for assessing generalizability is the case where the "items" are raters. Expert or possibly inexpert persons are employed to make judgments about a number of examinees. These are expressed in the form of, or scored as, numerical ratings. Such judgments are generally uncertain and in that sense "subjective." They might be m examiners' judgments of the quality of an expository paragraph or an essay, as written by each of n examinees. They might be ratings of a named characteristic—say, a defined personality trait—of n examinees, by m judges—professionals or peers. Raters may show, as their own stable characteristics, individual differences in the severity of their examination marks, or in the variability they are willing to allow between the essays judged best and worst. Similar remarks apply to ratings of named traits. It is hardly conceivable that we might have a large pool of raters, whose individual severities and variabilities in judgment are accurately known. If m raters in a generalizability study can be considered representatively sampled from the set of raters that might be freshly sampled in a future decision study, the random "items" model would be appropriate for describing the reliability/generalizability of one newly drawn rater's score, by (6.58) or of the mean of m of them, by (6.59). This might be well motivated if fewer raters are used in the later decision study than in the initial generalizability study. If there were more of them, the decision study could be used to assess the parameters—means and variances of ratings—characterizing the raters actually used. Then we could return to the true-score equivalence model or to the factor model for the generalizability of the ratings of the given judges to the domain of further judges. For all we then know or care, further judges might all be more or all less severe, or all more or all less sensitive to the characteristic rated. Thus, the most likely application of the random model (6.56) is in a generalizability study with multiple raters used to assess the error with which one fresh judge will approximate the mean rating that would be obtained from the available population of judges. The model does not allow the judges to vary in sensitivity or in error variance; therefore, this application will be of

interest only if the absolute value of the rating is of importance. We would use this model if, for example, a new rater's essay mark is compared to a fixed criterion for a passing grade, or if a newly recruited psychiatrist's rating of psychosis is to be compared to a criterion of severity used to determine hospitalization. If ratings are used for relative judgments, the difficulty variance should not be included in the error variance, and we again return to the fixed models (6.2) or (6.45).

The remaining comments in this section are clearest to students familiar with multifactorial designs. Cronbach generalizability theory includes a wider range of measurement designs than the simple respondents by items/measurements/ratings case considered so far. In principle, one can have a generalizability design for just about every major multifactorial ANOVA design. No attempt is made to present the general form of the theory here. We briefly consider some examples, to illustrate the general nature of these developments. One possibility is that we have responses of n examinees to m objective test items, on r occasions of repeated measurement, giving item scores x_{ijk}, $i = 1, \ldots, n$; $j = 1, \ldots, m$; $k = 1, \ldots, r$. Another is that we have ratings on n individuals of m behaviors thought to be indicators of a single attribute, by r raters. Either of these cases might be represented by the model

$$X_{ijk} = \mu + T_i + \delta_j + \gamma_k + (T\delta)_{ij} + (T\gamma)_{ik} + (\delta\gamma)_{jk} + E_{ijk}. \qquad (6.62)$$

In this model, with the notation used, the effect T_i of the attribute of the examinee would be regarded as random, whereas the effects j of items/rated behaviors and k due to raters/occasions of measurement are fixed. But further models can be obtained by replacing by D, and/or replacing by G, to denote random sampling of items/rated behaviors and/or of occasions/raters. Note that the additional terms in parentheses are interactions, allowing effects of combinations of conditions, as well as the main effects of items and occasions/raters.

Each of the resulting models yields a decomposition of the variance of X_{ijk} into components,

$$\sigma_X^2 = \sigma_T^2 + \sigma_D^2 + \sigma_G^2 + \sigma_{TD}^2 + \sigma_{TG}^2 + \sigma_{DG}^2 + \sigma_E^2, \qquad (6.63)$$

in which σ_D^2, σ_G^2, and σ_{DG}^2 are replaced by mean squares of effects if these are fixed. Students well grounded in ANOVA theory will easily see how the previous discussion can be applied to this class of design, to yield sums of squares, mean squares, and expected values of mean squares. These are combinations of the variance components in (6.63), from which they can be estimated. From these we can get generalizability coefficients in which some or all of the variance components are included with σ_E^2 in the "error term," depending on whether we regard the items/rated behaviors or the occasions/raters as random or fixed.

More generally, in these developments, we suppose that in addition to the sampled population of examinees, we have one or more distinct *facets*. In Cronbach generalizability theory this term replaces the word *factors* as used in multifactorial ANOVA designs. Facets are principles of classification of the measurements. Examples are items, rated behaviors, occasions, and raters. In Cronbach theory, the sampling units—respondents, examinees on which the measurements are made—are not regarded as forming a facet but are referred to as the *objects of measurement*. (In most applications, these units will be individual examinees. They could include other sampling units whose properties are to be measured—for example, classrooms.) The case on which the earlier discussion has been concentrated—respondents by items—is a single-facet design with respondents as objects of measurement and items as the only facet. The cases of respondents by items by occasions and of persons by ratings by raters are two-facet designs. These have, respectively, items and occasions, and ratings and raters, as their facets. A design with n persons by m ratings by r raters on t occasions is an example of a three-facet design. For example, we might have ratings of aggressive behaviors by a number of raters over a number of time intervals on preschool children in a play situation. At present, we do not seem to have more general versions of the factor theory of generalizability to analyze such data sets. (But see the model for "multitrait multimethod" data discussed in Chap. 10, as an example of what might be needed).

The interested student will pursue the reference note.[10] A few more remarks must serve to guide the reader in that pursuit and to close this chapter. All of these models require stringent assumptions of representativeness. This means fair sampling, and broad exchangeability of the facet elements—raters, occasions, items—in the domain. If we feel that the first m items we write are good ones—with a medium level of difficulty, and satisfactory sensitivity to the attribute—and that it will be hard to keep finding/writing items of similar quality, it is safer to regard the items as fixed and use the factor model for generalizability, forgoing any attempt to guess the behavior of further items. The same may be true of the characteristics of readily available raters in some situations. Not all their peers know people about equally well. It is particularly difficult to imagine that occasions of measurement constitute a representative sample of an infinite domain of distinct times. Most time-dependent behavior is best analyzed for systematic trends, possibly with sophisticated probabilistic (*stochastic*) models for time series—autoregressive/moving average models.[11] The further apart in time repeated measurements are made, the less similar they are, and the simple ANOVA treatments of generalizability do not model this. (That is one difficulty already noted with "retest reliability.") All of the models involving a random facet—random items, random raters—are well motivated only in cases where we wish to know the error of measurement of the domain score from distinct items or distinct raters. These will mostly be cases where the decision study (a) is distinct from the generalizability study, (b) uses fresh items, raters, and so on, and (c) uses fewer items, raters, and so on than the generalizability study. Such cases would appear to be uncommon.

REVIEW GUIDE

The Spearman general factor model (6.2) for a homogeneous test is the central topic of this chapter. Its basic formulation

needs close study. [The demonstration that the parameters can be determined from estimable variances and covariances, from (6.3) through (6.6), may be conceptually interesting to some students but is not essential knowledge.] The general model and its special cases, the true-score equivalence model and the parallel items model, through equations (6.3), (6.4), (6.12), and (6.14), give restrictive hypotheses to be fitted and tested for goodness of fit. Attention should concentrate on the use of both goodness of fit indices and the sizes of the discrepancies as illustrated in Tables 6.4(c), 6.5(a), and 6.5(b) to judge the fit of the model, and generally on the illustrative examples in the tables and text.

It is desirable that the student understand the decomposition of the total test score in (6.17) given by the factor model. Given the parameters of the general factor model, the student should be able to compute the reliability coefficient omega by (6.20) and, desirably, understand why it is a reliability coefficient for a homogeneous test.

The Guttman-Cronbach alpha (6.28) estimated by (6.27) and the Spearman-Brown reliability (6.33) are important special cases of omega. The student should be able to compute them and desirably should understand them as reliability coefficients when their special assumptions are satisfied and as lower bounds to the reliability of a homogeneous test. Again, close study of the examples in the tables should be helpful.

For binary data, KR20 and KR21 are of historical interest only. The numerical illustration of the application of the general factor model to a binary data set should be closely studied, and the student should simply accept the warning that we are here using the model as a possibly crude approximation to a more sophisticated model studied later.

The section on generalizability theory is regarded as an optional extra. It is probably enough to recognize (a) that coefficient omega is a coefficient measuring generalizability from a given set of items to a behavior domain, and (b) that further reading is available to the student who wishes to pursue the ANOVA treatment of generalizability to other items and conditions of observation as developed by Cronbach and his associates.

END NOTES

General: There are books on factor analysis at various levels of technicality. Most require matrix algebra. An exception is McDonald (1985). Students with a knowledge of matrix algebra will wish to consult Mulaik (1972).

Special Note: Estimation by unweighted least squares is easy to understand, but there are more sophisticated, more efficient methods of estimation, forming a class of *weighted least squares* estimators. These give differential "value" or "weight" to different discrepancies $s_{jk} - _{jk}$ in forming a function of them for the computer to minimize by a search algorithm. This is a very technical topic. Note the following points: (a) Certain weights for the discrepancies, based on the sample covariances s_{jk} give a computer program option called generalized least squares (GLS). If the item scores give a large-sample distribution of the item covariance matrix corresponding to a normal distribution—this condition may not reach the reader's intuition—GLS gives good estimators, and, if the sample size is sufficiently large—say, greater than 100—provides usable standard errors of estimate (*SE*s) of the parameters λ_j, ψ_j^2, and a chi-square test of the hypothesis (in the present case) of a single-factor model—that is, homogeneous items. (b) Certain weights for the discrepancies, based on the fitted covariances $\hat{\sigma}_{jk}$, give maximum likelihood (ML) estimators of the parameters if the unique parts of the item scores are normally distributed. ML estimators, like GLS, give *SE*s of estimate and a chi-square test of goodness of fit. (c) If the item scores, or at least their unique parts, are not normally distributed—for example, binary scores or Likert scores with few categories—special weights can be chosen to suit the case, to give good properties to the estimates. Generally these methods are limited to quite small numbers of items, and for quite good reasons have mainly been applied to item response models, and not to simple common factor models such as the single-factor model of this chapter.

1. CONFA is a simple self-contained program for confirmatory factor analysis, written by Colin Fraser and made available with this book.
2. Originally given by McDonald (1970). See McDonald (1985, Chap. 7).
3. See later section, Some Principles of Generalizability.
4. Original sources are Guttman (1945) and Cronbach (1951).
5. This proof, given by McDonald (1970), is patterned after a non-factor-analytic proof by Novick and Lewis (1967). See Lord and Novick (1968) for a conveniently accessible account of the Novick and Lewis proof.
6. See Lord and Novick (1968, pp. 112–114).
7. It is actually difficult to model binary data having this rather strange property. It seems that it has not been done, and indeed cannot be done, because it assumes a linear model. So KR20 is at best an approximation for binary items, which implicitly assumes the linear model.
8. This case has a simple item response model—see Chapter 13—in which every item has the same functional relation to the attribute.
9. Hoyt and, seemingly independently, Jackson and Ferguson gave this basic result in 1941. See Lord and Novick (1968).
10. Brennan (1983) is a good resource.
11. For a technical but usable source account, with references to follow up, see Cox, Hinkley, and Barndorff-Nielsen (1996).