CHAPTER 5
ISSUES IN SELECTING OBSERVATIONS AND ATTRIBUTES

From
Exploratory Factor Analysis
Ledyard R Tucker
and
Robert C. MacCallum

CHAPTER 5
ISSUES IN SELECTING OBSERVATIONS AND ATTRIBUTES

5.0. Introduction

The previous chapter provided a detailed presentation of the common factor model in a sample. The collection of sample data requires that a researcher do two things: (a) select a sample of observations from the population; and (b) select a battery of attributes from the domain of interest. The objective of the present chapter is to closely examine a number of critical issues involved in the process of selecting observations and attributes to be measured. It will be shown that there is a number of aspects of this process which can have a substantial impact on the results obtained in a factor analytic study.

5.1. Effects of Random Sampling

Let us first consider the process of selecting a sample from a population, and how factor analytic solutions may be affected by certain aspects of this process. As shown in the previous chapter, when the common factor model is applied to a sample, a degree of sampling error will be present and will influence the obtained factor solution. In examining the nature of this influence, it is useful to consider two distinct cases defined in terms of the type of sampling which is conducted. The first case is random sampling, where a sample is drawn by some random process from the population. In random sampling, each individual has an equal probability of being included in the sample. The second case is selective sampling, where observations are selected for the sample according to their level on one or more attributes; i.e., the probability of a given individual being included in the sample depends on the individual's score on one or more attributes. We will examine these two cases in turn.

In the following subsections we will consider three factors which influence the effect of sampling error under random sampling: (a) sample size; (b) the presence of unique factors; and (c) the standardization of the attributes in the sample.

5.1.1 Effects of Sample Size

Factor analysis is similar to other multivariate statistical methods in that the effects of sampling error are reduced as sample size increases. This phenomenon can be understood most easily by considering some equations developed in the previous chapter. In Eq. (4.11), the model defines population variances and covariances for the surface attributes as a function of the parameters in $\boldsymbol{B}$, $\boldsymbol{\Phi}$, and $\boldsymbol{U^2}$. Lack of fit of this model, as represented in $\boldsymbol{\Delta_\Sigma}$, arises only from model error. In Eq. (4.28), the model defines sample variances and covariances for the surface attributes as a function of the parameters in $\widetilde{\boldsymbol{B}}$, $\widetilde{\boldsymbol{C}}_{\beta\beta}$, and $\boldsymbol{U^2}$. Lack of fit of this model, as

represented in $\boldsymbol{\Delta}_y$, arises from a combination of model error and a primary source of sampling error. Recall that the presence of sampling error here is due to the fact that simplifying assumptions are made about the sample (i.e., that the unique factors are uncorrelated with each other and with the common factors in the sample), and that these assumptions almost surely do not hold exactly. Clearly, however, as sample size becomes very large these assumptions will hold more closely. That is, since these properties of unique factors do hold by definition in the population, they will tend to hold more closely in larger samples than in smaller samples. As a result, the impact of sampling error is reduced in larger samples, and lack of fit of the model to $\boldsymbol{C}_{yy}$, as represented in $\boldsymbol{\Delta}_y$, becomes primarily attributable to model error.

The impact of two other sources of sampling error described in Chapter 4 will also be reduced in large samples. The sources of sampling error were identified as arising from the effects of standardization of factors and of attributes in the sample. However, in large samples such a standardization would be very similar to the corresponding standardization in the population, thus indicating a reduced effect of these types of sampling error also.

It can also be recognized that solutions obtained from analyses of large samples will tend to be more similar to population parameter values. This can be understood by comparing Eqs. (4.1 1) and (4.28). When sample size becomes very large, the sample covariance matrix $\boldsymbol{C}_{yy}$ will become more similar to the population covariance matrix $\boldsymbol{\Sigma}_{yy}$. In addition, the lack of fit term in the sample, $\boldsymbol{\Delta}_y$, arising in large samples primarily from model error, will correspond closely to $\boldsymbol{\Delta}_\Sigma$, representing model error in the population. As a result, obtained sample solutions represented by matrices $\boldsymbol{B}$, $\boldsymbol{R}_{\beta\beta}$, and $\boldsymbol{U}^2$ will tend to be more similar to solutions for population parameters in $\boldsymbol{B}$, $\boldsymbol{R}_{\beta\beta}$, and $\boldsymbol{U}^2$ obtained by corresponding methods. All of these tendencies become stronger as sample size becomes larger. In addition, sampling variability of these sample solutions will be reduced as sample size increases; i.e., solutions will be more stable over repeated sampling in large samples. The mathematical basis for this phenomenon will be developed in later chapters.

In general, the effects of random sampling on obtained factor solutions are greatly reduced in large samples. This point will be emphasized and discussed in both theoretical and practical terms at several points in this and subsequent chapters.

5.1.2. The Presence of Unique Factors

A very interesting phenomenon is that the primary effect of sampling error in the common factor model depends on whether or not unique factors are present. In examining this phenomenon, we will make use of the model as expressed in terms of covariances of modeled attributes (Eqs. (4.4) and (4.14)), since we have shown that the primary impact of sampling error is revealed in that context (Eq. (4.16)). Equivalently, we will examine the issue by employing a

context where no model error is present. Let us first consider the case where no unique factors are present. In this case, Eq. (4.4) for the population can be rewritten as follows:

$$\boldsymbol{\Sigma}_{zz} = \boldsymbol{B}\boldsymbol{\Phi}\,\boldsymbol{B}' \tag{5.1}$$

In addition, Eq. (4.14), which represents the complete expression of the common factor model in the sample, would be greatly simplified. Since no unique factors are present, the last three terms in that equation would vanish, yielding

$$\boldsymbol{C}_{zz} = \boldsymbol{B}\boldsymbol{C}_{\beta\beta}\boldsymbol{B}' \tag{5.2}$$

A comparison of Eqs. (5.1) and (5.2) clearly shows that the common factor weights in the sample will be the same as the population common factor weights in this situation. However, the common factor covariance matrix for the sample will not match the corresponding matrix from the population. Thus, when no unique factors are present, sampling error will not influence the common factor weights but will influence the common factor variances and covariances.

Two aspects of this phenomenon should be pointed out. The first is that sample factor solutions are generally represented in a form where the common factors are standardized in the sample. This issue was discussed in Chapter 4, and the mathematical framework for this standardization was presented in Eqs. (4.17)-(4,20). Substituting from Eq. (4.20) into Eq. (5.2) yields the following representation of $\boldsymbol{C}_{zz}$ in terms of parameters corresponding to common factors standardized in the sample:

$$\boldsymbol{C}_{zz} = \widetilde{\boldsymbol{B}}\,\widetilde{\boldsymbol{C}}_{\beta\beta}\widetilde{\boldsymbol{B}}^{\,\prime} \tag{5.3}$$

In this context where no model error and no unique factors are present, matrix $\widetilde{\boldsymbol{B}}$ would correspond to an obtained sample common factor weight matrix $\boldsymbol{B}$. The relationship between this obtained solution and the population common factor weight matrix $\boldsymbol{B}$ is given by Eq. (4.19). Rewriting this equation to suit the present context yields

$$\boldsymbol{B} = \boldsymbol{B}[\boldsymbol{C}_d]^{\frac{1}{2}}_{\beta\beta} = \widetilde{\boldsymbol{B}} \tag{5.4}$$

This relationship further isolates one effect of sampling on common factor weights discussed in Chapter 4: when the common factors are standardized in the sample, the obtained common factor weights will equal the population common factor weights for standardized factors multiplied by the standard deviations of the common factors in the sample. It must be kept in mind that this simple relationship is defined in the particular context where model error and unique factors <u>are not</u> present. When model error and unique factors <u>are</u> present, the multiplicative effect given by Eq. (5.4) will still occur, but the relationship between $\boldsymbol{B}$ and $\widetilde{\boldsymbol{B}}$ will be influenced by these additional effects.

A second important aspect of the relationship between obtained and population common factor weights in the present context involves the fact that most methods for obtaining a factor solution provide an initial solution characterized by orthogonal common factors. Thus an initial sample common factor weight matrix $\boldsymbol{A}$ would be obtained such that

$$\boldsymbol{C}_{zz} = \boldsymbol{A}\boldsymbol{A}' \tag{5.5}$$

Matrix A will be of order $n \times r$ and will contain entries a_{jk} representing the sample weight for factor *k* on attribute *j*. Matrix $\boldsymbol{B}$ in Eq. (5.4) would then actually correspond to a transformed solution obtained by transforming matrix $\boldsymbol{A}$. That is, following the framework developed in Chapter 3, a sample trait matrix $\boldsymbol{T}$ will exist such that

$$\boldsymbol{A}\boldsymbol{T}^{-1} = \boldsymbol{B} \tag{5.6}$$

Since, as noted above, $\boldsymbol{B}$ is equivalent to $\widetilde{\boldsymbol{B}}$ in the present context, we could also write

$$\boldsymbol{A}\boldsymbol{T}^{-1} = \widetilde{\boldsymbol{B}} \tag{5.7}$$

This indicates that when no unique factors are present the obtained matrix of common factor weights can be transformed into the population weights rescaled by sample common factor standard deviation. In addition, the intercorrelations of the transformed factors in the sample would be given by

$$\widetilde{\boldsymbol{C}}_{\beta\beta} = \boldsymbol{T}\boldsymbol{T}' \tag{5.8}$$

Thus, the relationship defined in Eq. (5.4) would be observable only after an appropriate transformation of an initial orthogonal solution.

It must be kept in mind that the relationships just defined represent the case of no model error and no unique factors. Let us now consider the case where unique factors are present. In effect, the impact of sampling error in this situation has already been described in the context of Eq. (4.14). That equation represents the complete explanation of the common factor model in a sample. That model is then simplified by assuming that unique factors are uncorrelated with each other and with the common factors in the sample, thus yielding Eq. (4.15). When that equation was introduced, it was explained that the presence of sampling error gives rise to non-zero correlations of unique factors with each other and with common factors, meaning that the simplifying assumptions generally do not hold in practice. This in turn gives rise to the matrix $\boldsymbol{\Delta}_z$ in Eq. (4.16), representing a lack of fit arising from this of sampling error. Solutions obtained via factor analytic methods which attempt to optimize fit will thus be influenced by the presence of sampling error. That is, obtained common factor weights in matrix $\boldsymbol{A}$ will be influenced by the presence of sampling error when unique factors are present. As a result, it will not be possible to

transform the obtained solution $\boldsymbol{A}$ to a matrix $\boldsymbol{B}$ satisfying Eq. (5.4), as it is when no unique factors are present.

This influence of sampling error on obtained common factor weights when unique factors are present can be expected to increase as the importance of the unique factors increases. Consider Eqs. (4.14) - (4.16), it can be seen that if the influence of the unique factors is very slight, the contribution of the last three terms in Eq. (4.14) will be minor, thus yielding only a slight influence of sampling error as represented in $\boldsymbol{\Delta}_z$ in Eq. (4.16). As the importance of the unique factors increases, this influence of sampling error will also increase. This effect of sampling error will also increase as sample size is reduced. As explained in section 5.1.1, in a large sample the properties of unique factors in the population (uncorrelated with each other and with the common factors) will hold more closely in the sample, and the influence of sampling error will be reduced. In a small sample, there is greater likelihood of substantial correlations among unique factors and between unique and common factors, thus causing a greater impact of sampling error. Thus, the influence of sampling error on obtained factor solutions is reduced when sample size is large and the influence of unique factors is small.

One final interesting point regarding this issue of the impact of unique factors on sampling error concerns the components analysis model, described in section 3.5. The components model can be viewed as being obtained by deleting unique factors from the common factor model. In the present context, it can thus be seen that sampling error would not affect weights obtained in the components model (except for the multiplicative effect given in Eq. (5.4)), but would affect variances and covariances of the components in a random manner.

5.1.3. The Effect of Standardization of Attributes

It is of interest to consider the effect of sampling error when attributes have been standardized. (The reader must be careful here to distinguish between standardization of attributes and standardization of factors. One issue treated in the previous section was standardization of factors; standardization of attributes is a completely separate issue.) To this point in this chapter, we have dealt with the case where attributes are unstandardized in both the population and sample. Let us now examine the influence of standardization in both of those groups. We will still employ the context of modeled attributes, and we will make use of the case where no unique factors are present, so as to allow us to examine the effect of standardization when no other sources of sampling error is operating. This represents a more detailed and isolated study of this effect than was presented in Chapter 4.

Considering the population first, we have defined a population covariance matrix $\boldsymbol{\Sigma}_{zz}$ for modeled attributes. This can be converted to a correlation matrix (i.e., a covariance matrix for standardized attributes) by making use of a diagonal matrix $[\boldsymbol{\Sigma}_d]_{zz}$ containing variances of the modeled attributes on the diagonal. That is

$$[\Sigma_d]_{zz} = \boldsymbol{Diag}(\Sigma_{zz}) \tag{5.9}$$

We can then define a correlation matrix $\boldsymbol{R}_{zz}$ for the modeled attributes according to

$$\boldsymbol{R}_{zz} = [\Sigma_d]_{zz}^{-\frac{1}{2}} \Sigma_{zz} [\Sigma_d]_{zz}^{-\frac{1}{2}} \tag{5.100}$$

substituting from Eq. (5.1) into Eq. (5.10) yields

$$\boldsymbol{R}_{zz} = [\Sigma_d]_{zz}^{-\frac{1}{2}} \boldsymbol{B\Phi}\ \boldsymbol{B}'[\Sigma_d]_{zz}^{-\frac{1}{2}} \tag{5.11}$$

Let us next define a matrix $\boldsymbol{B}^*$ containing population common factor weights for standardized attributes as follows

$$\boldsymbol{B}^* = [\Sigma_d]_{zz}^{-\frac{1}{2}} \boldsymbol{B} \tag{5.12}$$

We can then substitute from Eq. (5.12) into Eq. (5.11) to obtain

$$\boldsymbol{R}_{zz} = \boldsymbol{B}^* \boldsymbol{\Phi}\ \boldsymbol{B}^* \tag{5.13}$$

This equation expresses the common factor model for standardized attributes in the population, for the case of no unique factors.

Let us now consider a similar framework for the sample. The sample covariance matrix for the modeled attributes can be converted to a correlation matrix by making use of a diagonal matrix $[\boldsymbol{C}_d]_{\mathbf{zz}}$ containing sample variances of the modeled attributes on the diagonal. That is

$$[\boldsymbol{C}_d]_{\mathbf{zz}} = \boldsymbol{Diag}(\boldsymbol{C}_{zz}) \tag{5.14}$$

The desired correlation matrix for the modeled attributes will be designated $\boldsymbol{R}_{zz}$ and can be defined as follows:

$$\boldsymbol{R}_{zz} = [\boldsymbol{C}_d]_{zz}^{-\frac{1}{2}} \boldsymbol{C}_{zz} [\boldsymbol{C}_d]_{zz}^{-\frac{1}{2}} \tag{5.15}$$

Substituting from Eq. (5.5), which represents an obtained orthogonal common factor solution for $\boldsymbol{C}_{zz}$, into Eq. (5.15) gives us

$$\boldsymbol{R}_{zz} = [\boldsymbol{C}_d]_{zz}^{-\frac{1}{2}} \boldsymbol{AA}'[\boldsymbol{C}_d]_{zz}^{-\frac{1}{2}} \tag{5.16}$$

Let us next define a matrix $\boldsymbol{A}^*$ containing obtained weights for orthogonal common factors on standardized modeled attributes as follows:

$$\boldsymbol{A}^* = [\boldsymbol{C}_d]_{zz}^{-\frac{1}{2}} \boldsymbol{A} \tag{5.17}$$

We can then substitute from Eq. (5.17) into Eq. (5.16) to obtain

$$\boldsymbol{R}_{zz} = \boldsymbol{A}^{*}\boldsymbol{A}^{*} \tag{5.18}$$

We now have equations representing the model for standardized attributes in the population (Eq. (5.13)) and in the sample (Eq. (5.18)), Considering the two expressions, the issue of interest here is whether $\boldsymbol{A}^{*}$ can be transformed to obtain $\boldsymbol{B}^{*}$. If that is possible, then standardization has no negative impact. If such a transformation is not possible, however, then standardization of the attributes has introduced another source of error. To resolve this issue, note first that it has already been shown that in the case of unstandardized attributes the obtained matrix $\boldsymbol{A}$ can be transformed into $\widetilde{B}$, as given in Eq. (5.7). By making use of relationships defined in Eqs. (5.4) and (5.12) we can obtain the following:

$$\boldsymbol{B} = [\boldsymbol{\Sigma}_d]_{zz}^{\frac{1}{2}}\boldsymbol{B}^{*}[\boldsymbol{C}_d]_{\beta\beta}^{\frac{1}{2}} \tag{5.19}$$

Substituting from this equation into Eq. (5.7) yields

$$\boldsymbol{A}\boldsymbol{T}^{-1} = [\boldsymbol{\Sigma}_d]_{zz}^{\frac{1}{2}}\boldsymbol{B}^{*}[\boldsymbol{C}_d]_{\beta\beta}^{\frac{1}{2}} \tag{5.20}$$

Employing the relationship given in Eq. (5,17), we can rewrite Eq. (5.20) as follows:

$$[\boldsymbol{C}_d]_{\mathbf{zz}}^{\frac{1}{2}}\boldsymbol{A}^{*}\boldsymbol{T}^{-1} = [\boldsymbol{\Sigma}_d]_{zz}^{\frac{1}{2}}\boldsymbol{B}^{*}[\boldsymbol{C}_d]_{\beta\beta}^{\frac{1}{2}} \tag{5.21}$$

Finally, solving for $\boldsymbol{B}^{*}$ on the right side yields

$$[\boldsymbol{\Sigma}_d]_{zz}^{-\frac{1}{2}}[\boldsymbol{C}_d]_{\mathbf{zz}}^{\frac{1}{2}}\boldsymbol{A}^{*}\boldsymbol{T}^{-1}[\boldsymbol{C}_d]_{\beta\beta}^{-\frac{1}{2}} \tag{5.22}$$

This equation is important because it shows that there is not a simple direct transformation of $\boldsymbol{A}^{*}$ to $\boldsymbol{B}^{*}$. That is, there is no trait matrix $\boldsymbol{T}$ which would provide this transformation using the framework defined in section 3.3. Rather, the rows of $\boldsymbol{A}^{*}$ would have to be rescaled using population and sample standard deviations as shown in Eq. (5.22) before such a transformation would be possible. It can be seen that the adverse impact of standardization will be reduced as sample sizes increases. With very large samples, the sample standard deviations in $[\boldsymbol{C}_d]_{\mathbf{zz}}$ will be very similar to the population standard deviations in $[\boldsymbol{\Sigma}_d]_{zz}$. In this case, the product of the first two terms in Eq. (5.22) will be very nearly an identity matrix, thus allowing the transformation of $\boldsymbol{A}^{*}$ to more closely approximate $\boldsymbol{B}^{*}$.

Regardless of the reduced effect in large samples, the fact remains that the standardization of attributes in a sample introduces an additional source of sampling error which further hinders the degree to which the population common factor weights can be recovered. This source of error can be easily eliminated via the use of covariance matrices in factor analytic studies.

5.1.4. Summary and Demonstration of Effects of Sampling Error Under Random Sampling

We have described three factors which influence the degree of effects of sampling error on obtained factor solutions under the condition of random sampling. The factors are (a) sample size, (b) the role of unique factors, and (c) the standardization of attributes. Sampling error will be reduced when sample size is large, since parameter estimates obtained in large samples are more stable and provide more accurate estimates of true parameter values. Sampling error will have only a multiplicative influence on common factor weights for each factor when unique factors are absent, but will randomly influence common factor variances and covariances. The multiplicative effect arises from the standardization of the factors in the sample. When unique factors are present, estimates of all of these parameters will be affected by sampling error, with the magnitude of the effect being reduced when sample size is large and the influence of the unique factors is small. Finally, standardization of attributes in a sample introduces an additional source of sampling error in that common factor weights obtained for such attributes cannot be transformed into population factor weights, even when no unique factors are present.

We will now offer a demonstration of these phenomena. The demonstration was carried out using a $2 \times 2 \times 2$ design. The factors represented in this design were (a) 2 levels of sample size; (b) two levels of communality, representing cases where unique factors were either absent or present; and (c) two types of matrices analyzed, covariance or correlation. An artificial sample data set was generated for each of the eight conditions in this design. This was accomplished by first constructing two population factor weight matrices containing weights for two factors on ten attributes. These two matrices, shown in Table 5.1, contain weights for standardized, uncorrelated factors on attributes considered to be standardized in the population. The first such matrix represents the case where are no unique factors; this is revealed by the fact that the communalities for the attributes, as given by the sum of the weights in each row, are all equal to unity. The second weight matrix represents the case where substantial unique factors are present; the weights in this matrix were obtained by multiplying all weights in the first matrix by .7, thus yielding a case where the communalities of all attributes were .49. Each of these two weight matrices can be considered to be a population factor weight matrix $\boldsymbol{B}$.

Next a sample of scores on factors was generated for each of two sample sizes, N=25 and N=1000. Scores for the individuals in these samples were generated on two common factors and ten unique factors by sampling from a multidimensional normal population with mean zero and variance unity. Based on these two samples of factor scores, a factor covariance matrix was computed for each sample. Each of the two population factor weight matrices were then applied to each of the two sample factor covariance matrices, as in Eq. (4.12), to yield four sample covariance matrices. Each of these could be considered to be a $\boldsymbol{C}_{zz}$ matrix. Each of these covariance matrices was then standardized to obtain a correlation matrix; each of the resulting

Table 5.1

Population Factor Weight Matrices*

Population Communalities=1.00				Population Communalities=.49			
	Weights on Factors				Weights on Factors		
	1	2	Unique		1	2	Unique
1	1.000	.000	.000	1	.700	.000	.714
2	1.000	.000	.000	2	.700	.000	.714
3	1.000	.000	.000	3	.700	.000	.714
4	.960	.280	.000	4	.672	.196	.714
5	.800	.600	.000	5	.560	.420	.714
6	.600	.800	.000	6	.420	.560	.714
7	.600	.800	.000	7	.420	.560	.714
8	.000	1.000	.000	8	.000	.700	.714
9	.000	1.000	.000	9	.000	.700	.714
10	.000	1.000	.000	10	.000	.700	.714

* In the population the factor scores were standardized, uncorrelated.

four correlation matrices could be considered to be a $\boldsymbol{R}_{zz}$ matrix. This yielded eight sample data matrices, one for each cell defined by the $2 \times 2 \times 2$ design.

Each of these eight matrices was then analyzed by factor analysis procedures to be discussed in Chapter 7. As with previous demonstrations, it is not necessary at this point to understand the methodology employed. All that is necessary to understand at this point is that for each of the eight constructed matrices a factor solution was obtained which, according to a specific criterion, optimally fit the matrix in question. In terms of the developments in Chapter 4, this was achieved by considering each of the constructed $\boldsymbol{C}_{zz}$ and $\boldsymbol{R}_{zz}$ matrices to be equivalent to $\boldsymbol{C}_{yy}$ and $\boldsymbol{R}_{yy}$ matrices, respectively, and fitting the models given by Eqs. (4.28) and (4.37). Consideration of the constructed $\boldsymbol{C}_{zz}$ and $\boldsymbol{R}_{zz}$ matrices as $\boldsymbol{C}_{yy}$ and $\boldsymbol{R}_{yy}$ matrices is appropriate in the present context since no model error is incorporated into this demonstration. Of primary importance in the present context is the relationship between these obtained solutions and the known population parameters. It should be pointed out that in order to evaluate this relationship it was necessary to transform each of the obtained orthogonal common factor weight matrices to approximate the appropriate population common factor weight matrix. This was achieved by a method called canonical congruence transformation. The mathematical framework for this transformation will be presented in Chapter 9. For present purposes, it is sufficient to understand that each obtained orthogonal weight matrix $\boldsymbol{A}$ was transformed to an oblique solution $\boldsymbol{B}$, such that $\boldsymbol{B}$ would be maximally similar to the corresponding population weight matrix, $\boldsymbol{B}$.

Based on the developments in the previous section, we can state the results that we would expect. First, we would expect that the variability of obtained statistics from population parameters will be more evident for the small sample than for the large sample. In the analysis of covariance matrices, we expect that the columns of obtained common factor weights will equal the corresponding population common factor weights, multiplied by the sample standard deviations of the corresponding common factor scores. This relationship, as defined in Eq. (5.4), will be exact when unique factor weights are zero but not when unique factor weights are greater than zero. In this case, additional variability in attribute factor weights is brought on by unique factor variances and covariance. Thus, when communalities are less than unity, we expect that the relationship between population and obtained common factor weights will be randomly perturbed from that defined in Eq. (5.4). Finally, we expect that solutions obtained from the analysis of correlation matrices will be characterized by differential effects on the common factor weights for the attributes, as implied by Eq. (5.22) and arising from the effects of standardization of the attributes in the sample.

Major results are shown in the figures. Figure 5.1 and 5.2 are for the analyses of covariance matrices, Figures 5.3 and 5.4 are for the analyses of correlation matrices. Figures 5.1

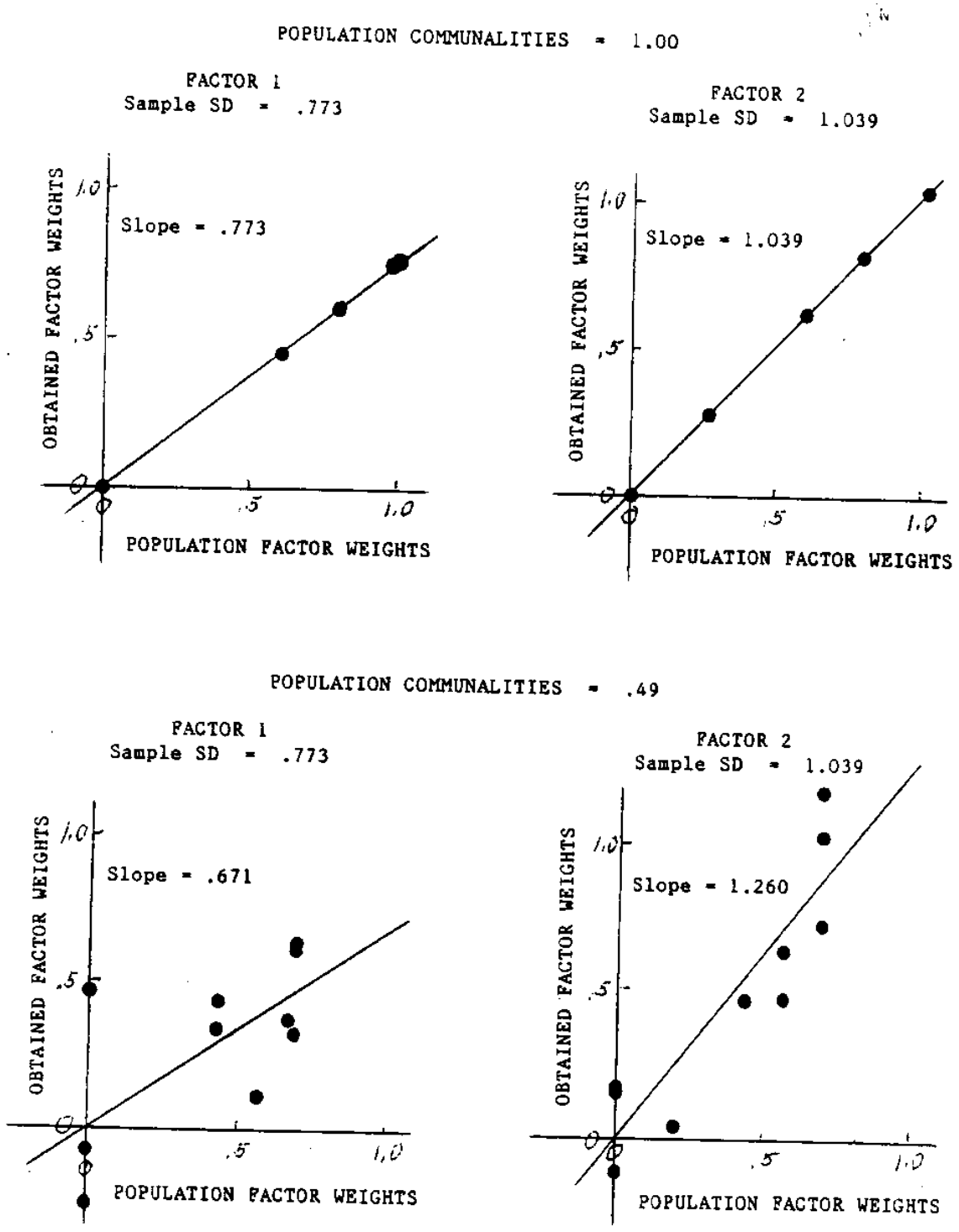

Figure 5.1: Examples of Effects of Random Sampling; Comparison of Obtained Factor Weights with Population Factor Weights; Sample Covariance Matrix Analysed; Sample Size = 25

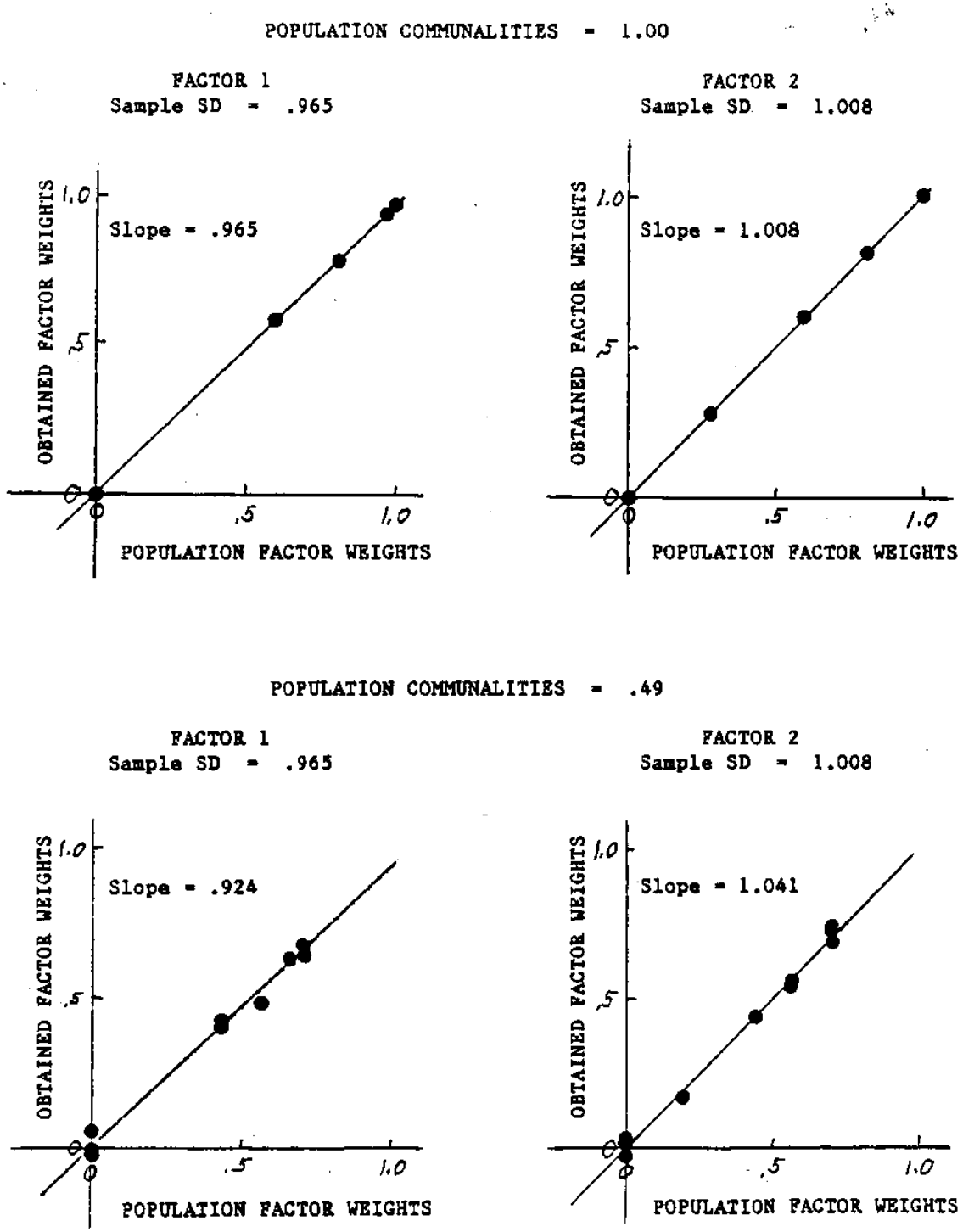

Figure 5.2: Examples of Effects of Random Sampling; Comparison of Obtained Factor Weights with Population Factor Weights; Sample Covariance Matrix Analysed; Sample Size = 1000

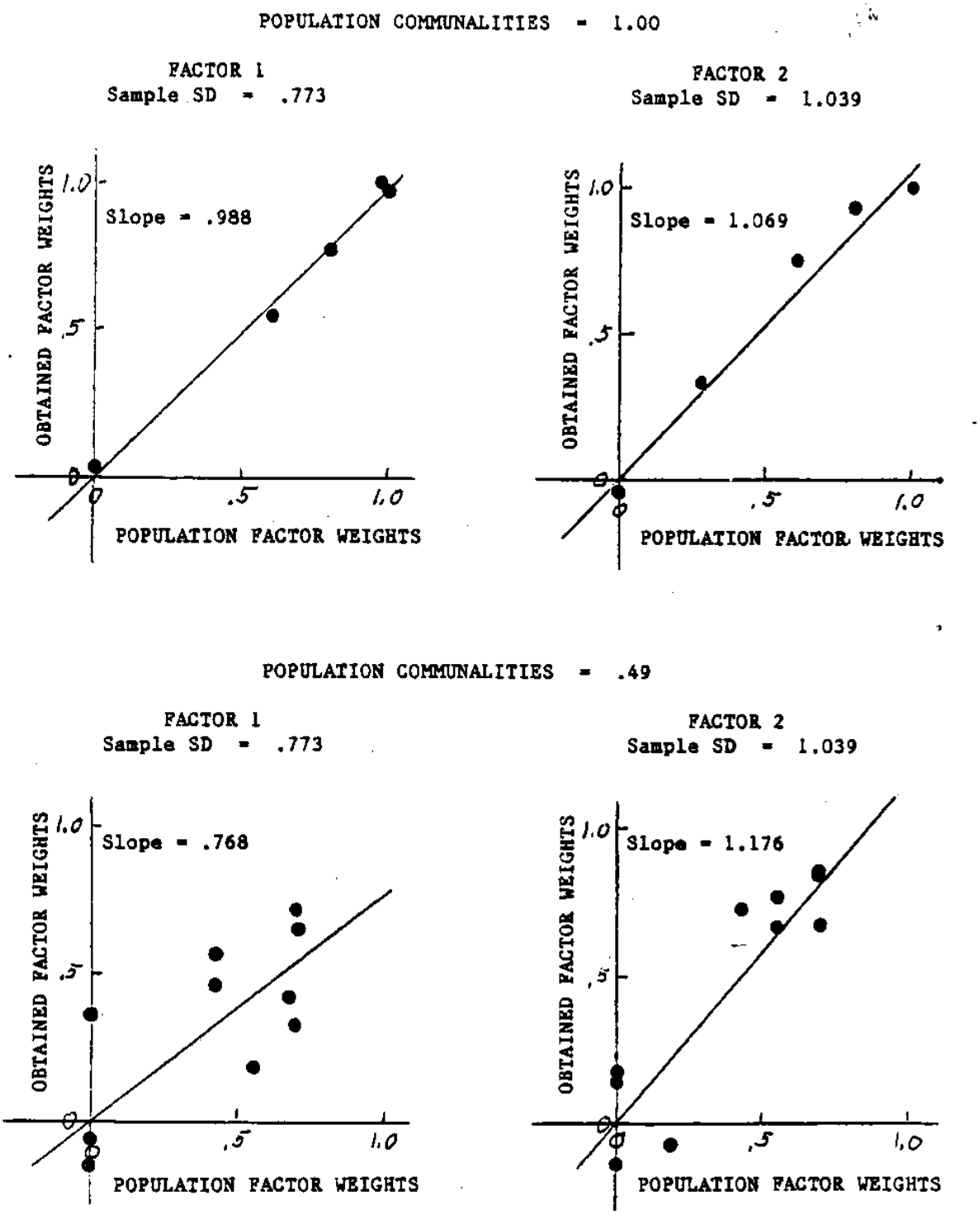

Figure 5.3: Examples of Effects of Random Sampling; Comparison of Obtained Factor Weights with Population Factor Weights; Sample Covariance Matrix Analysed; Sample Size = 25

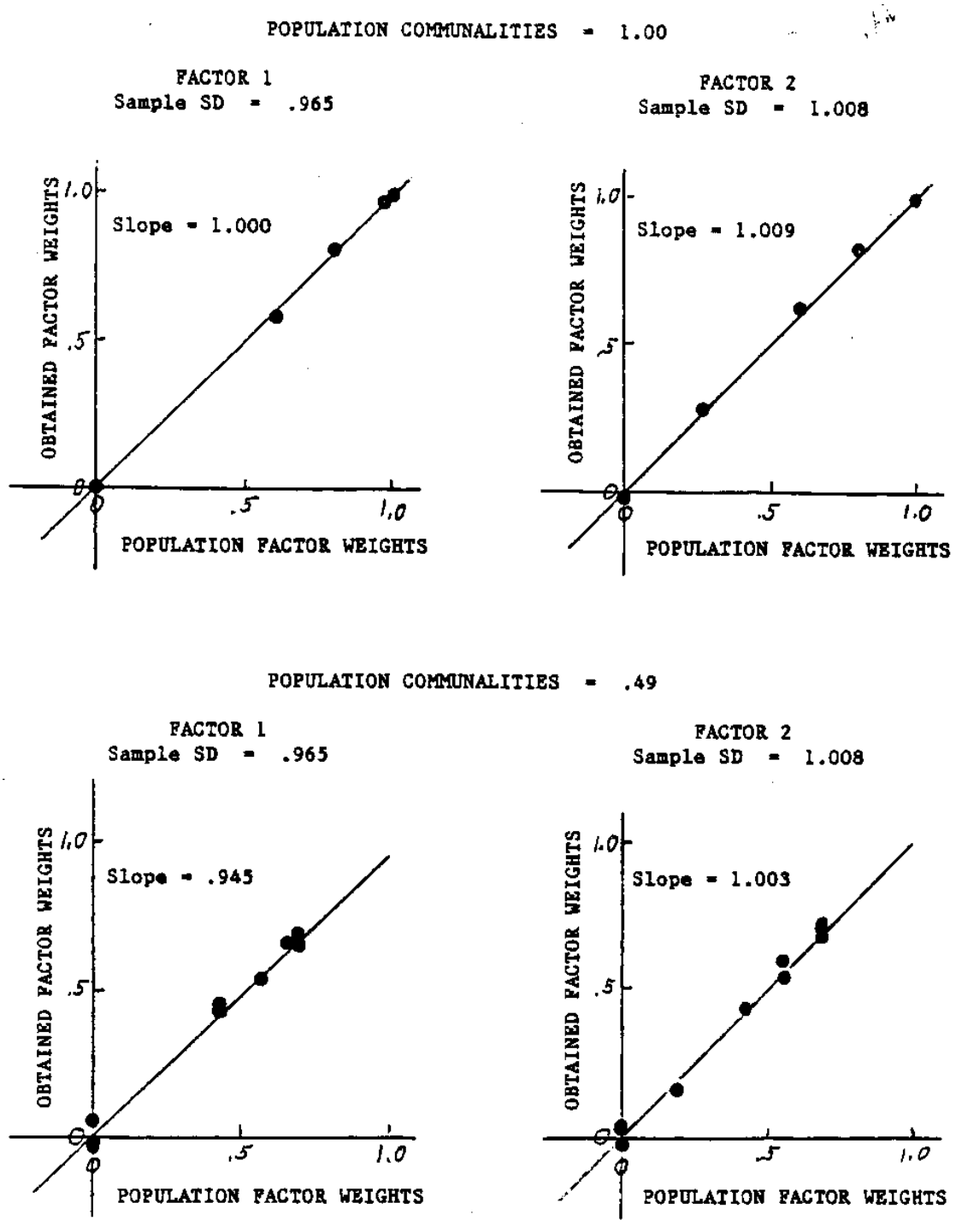

Figure 5.4: Examples of Effects of Random Sampling; Comparison of Obtained Factor Weights with Population Factor Weights; Sample Covariance Matrix Analysed; Sample Size = 1000

and 5.3 are for the sample size of 25 and Figures 5.2 and 5.4 are for the sample size of 1000. The top pair of graphs in each figure show results for the population factor matrix having communalities equal to unity; the lower pair of graphs show results for the population factor matrix having communalities equal to .49. For each figure, the left graph shows results for factor 1 while the right graph shows results for factor 2. Each graph plots obtained factor weights against corresponding population factor weights. There is a plotted point for each attribute. In several instances, however, several of the attributes plot on the same point. A line of best fit to the points, with discrepancies measured vertically, is drawn on each graph.

Consider Figure 5.1 and 5.2 for the analysis of covariance matrices. The upper pair of graphs in each figure are for the matrices having population communalities equal to 1.00. On both graphs on both figures, the points are precisely on the lines of best fit. Note also that the slopes of these lines equal the standard deviations of the factor scores in the sample. Thus, random sampling results in the population factor weights being multiplied by the sample standard deviations of the factor scores. This is a systematic effect for each factor and corresponds to the relationship defined in Eq. (5.4). The size of the effect depends on the sampling variability of the factor score standard deviation, this variability being larger for small samples than for larger samples. This is demonstrated by the fact that the slopes in the upper pair of graphs in Figure 5.2 are closer to 1.00 (the population standard deviation of the factors) than are the corresponding slopes in Figure 5. 1.

Consider the lower pair of graphs in Figures 5.1 and 5.2; these represent the case of population communalities equal to .49, meaning that substantial unique factors are present. It is seen that the points in the lower graphs in Figure 5.1, where sample size is 25, have considerable divergence from the line of best fit. As discussed above, this is brought on by the sampling variances and covariances of the unique factors. Increasing sample size to 1000 (see Figure 5.2) reduces the divergence of the points from the line of best fit. A further effect is that the lines of best fit have slopes which differ from the sample factor score standard deviations. Thus, random sampling has two effects when the unique loadings do not equal zero: (a) there is a general multiplicative effect for each factor, with population factor weights being multiplied by sample factor standard deviations; and (b) there is perturbation of individual attribute weights due to the effects of unique factor variances and covariances.

Consider next Figures 5.3 and 5.4, representing results from the analysis of correlation matrices. For the case of population communalities equal to 1.00, the points no longer are precisely on the lines of best fit. This is brought on by differential effects between attributes arising from standardization of attribute scores in the sample, as implied by Eq. (5.20). These effects are more pronounced in the small sample than in the large sample, as seen by comparing the upper pair of graphs in Figure 5.3 to those in Figure 5.4. Further, the slopes of the lines of

best fit no longer equal the factor score standard deviations in the samples. A further comment about the slopes will be made in the discussion of Table 5.3. Considering the lower graphs in Figures 5.3 and 5.4, representing the case of population communalities equal to .49, considerable divergence of the points from the line of best fit remains for the small sample in Figure 5.3. For the large sample represented in Figure 5.4, the points are close to the line of best fit.

Table 5.2 presents statistics which summarize the goodness of fit of the lines of best fit to the points in the graphs. The statistic presented in this table is called a coefficient of congruence. This is a measure of correspondence between two factors. In the present context, if we consider factor *k* from obtained weight matrix $\boldsymbol{B}$ and factor 1 from population weight matrix $\boldsymbol{B}$, the coefficient of congruence, g_{kl}, between these two factors would be defined as follows:

$$g_{kl} = \frac{\sum_{j=1}^{n} b_{jk}\beta_{jl}}{\underline{\nu}(\sum_{j=1}^{n} b_{jk}^2)(\sum_{j=1}^{n} \beta_{jl}^2)} \qquad (5.23)$$

Geometrically, this value represents the cosine of the angle between the factors when they are plotted in the same factor space. When two factors are exactly congruent, the coefficient of congruence will equal 1.00. As factors become less congruent, the value of the coefficient decreases. Table 5.2 provides coefficients of congruence between corresponding obtained and population common factors for each of the eight analyses conducted.

For the analysis of the covariance matrices and population communalities of 1.00, the coefficients of congruence equal 1.000 indicating a very good fit of the line of best fit to the points. This result is expected from theory and agrees with results represented in Figures 5.1 and 5.2. Continuing to consider the analysis of the covariance matrices, the coefficients of congruence for the communalities of .49 with the small sample are quite low, especially for factor 1. The influence of sample size is indicated by the quite high coefficients of congruence when the sample size is increased to 1000.

In the analyses of the correlation matrices, for the case of population communalities of 1.000 in the small sample, the coefficients of congruence decreased from the corresponding values for the analysis of covariance matrices. There was a similar decrease in the coefficients of congruence for the large sample, but beyond the three decimal places used in the table. For the case of population communality equal to .49, there was a general increase in the coefficients of congruence from analysis of correlation matrices over the values from analysis of covariance matrices.

This somewhat unexpected event was found to occur fairly consistently in a Monte Carlo study involving 100 samples of the same type analyzed here. For factor 1, 81 of the 100

Table 5.2

Coefficients of Congruence: Obtained Factor Weights with Population Factor Weights

Population Communalities	Sample Size	Coefficients of Congruence Factor 1	Factor 2
	Sample Covariance Matrix Analyzed		
1.00	25	1.000	1.000
.49	25	.849	.962
1.00	1000	1.000	1.000
.49	1000	.998	.999
	Sample Correlation Matrix Analyzed		
1.00	25	.999	.995
.49	25	.888	.962
1.00	1000	1.000	1.000
.49	1000	.999	.999

replications yielded a higher coefficient of congruence in the solution obtained from analysis of a correlation matrix over that obtained from analysis of a covariance matrix. For factor 2, 78 of the 100 replications followed the same trend. Both of these divisions were significantly different from chance.

Table 5.3 summarizes results for the slopes of the lines of best fit. In the analysis of the covariance matrices and population communalities of unity, the slopes equal the sample factor score standard deviations. This is the expected result. Comparison of the slopes of the best fitting lines for the analyses of the covariance matrices and the analyses of the correlation matrices indicates that there is a strong tendency for the slopes for the analyses of the correlation matrices to change from the analyses of the covariance matrices toward unit slopes. This could be interpreted that the standardization of the attributes in the sample compensated for the effects of considering standardized factors in the sample. This compensation is not completely accurate.

Table 5.4 presents the correlations between the factors in two different contexts: (a) for the input factor scores in the sample; (b) for the factors obtained from the factor analysis of the sample data. In the analysis of the covariance matrices and population communality of 1.00, the correlations obtained from the analyses equal the factor score correlations in the sample. This equivalence does not hold in any other case. The foregoing equivalencies are the expected results from theory. One other observation is that the correlations for the large samples more nearly equal the zero correlation between the factors in the population than do the correlations for the small samples. This is an expected result, since the accuracy of these estimates of population parameters from sample statistics are dependent on sample size.

These examples were undertaken to illustrate the source of variation in factor weights produced by random sampling of entities in samples. Since, in the factor analytic model, the attribute measures are dependent on the factor scores, sampling effects on the attribute scores are dependent on the sampling variabilities of the factor score statistics. In these examples the population distribution of the factor scores was taken to be multidimensional normal. Thus, sampling effects may be considered in terms of sample means and sample variances and covariances. The effects of sample means are eliminated by the use of deviation measures in the samples (as implied by the computation of covariance and correlation matrices). Thus, the remaining sample effects arise from the sample factor score variance and covariances.

Factors may be divided into the common factors and the unique factors. Each of these classes of factors will have different sampling effects on the factor weights obtained from analyses. Major effects from the common factor variances arise from standardization of the factor scores in the sample, a process inherent in the factoring methods. The restandardization of the factor scores in the sample results in the population factor weights being multiplied by the sample standard deviations of the common factor scores. This effect is illustrated by the slopes of

Table 5.3

Comparisons between Sample Factor SD's and Slopes of Best Fitting Lines:

Obtained Factor Weights on Population Factor Weights

Population Communalities	Sample Size	Factor 1		Factor 2	
		Sample SD	Slope	Sample SD	Slope
Sample Covariance Matrix Analyzed					
1.00	25	.773	.773	1.039	1.039
.49	25	.773	.671	1.039	1.260
1.00	1000	.965	.965	1.008	1.008
.49	1000	.965	.924	1.008	1.014
Sample Correlation Matrix Analyzed					
1.00	25	.773	.988	1.039	1.069
.49	25	.773	.768	1.039	1.176
1.00	1000	.965	1.000	1.008	1.009
.49	1000	.965	.945	1.008	1.003

Table 5.4

Correaltion between Factors (Population Correlation=.000)

Population Communalities	Sample Size	Coefficients of Congruence	
		Factor 1	Factor 2
Sample Covariance Matrix Analyzed			
1.00	25	-.206	-.206
.49	25	-.206	-.389
1.00	1000	-.033	-.033
.49	1000	-.033	-.047
Sample Correlation Matrix Analyzed			
1.00	25	-.206	-.166
.49	25	-.206	-.325
1.00	1000	-.033	-.026
.49	1000	-.033	-.023

the lines of best fit in Figures 5.1 and 5.2 being equal to sample standard deviations when population communalities equal unity. This effect applies to the analysis of matrices of covariance when the attribute scores are expressed in the population unit of measures. Standardization of the attribute scores in the sample by analysis of the sample correlation matrix introduces sampling variability for each attribute. This restandardization of the scores on each attribute may partially adjust for the restandardization of the factor scores. However, this adjustment is incomplete as shown in the examples in Figures 5.3 and 5.4. Since the unique loadings are zero for the case when the population communalities equal unity, there are no effects of the unique factor score variances and covariances.

When the population communalities are less than unity and, thus, the unique factor weights are greater than zero, sampling variation is introduced for individual attribute factor weights by the sampling variation in the unique factors. These variations of the factor weights for the individual attributes are most pronounced in the results for the small sample. See the lower pair of graphs in Figures 5.1 and 5.3. These sampling results are greatly reduced for the large sample.

Finally, as expected, all sampling variations are substantially reduced from the small to the large sample.

5.2. Effects of Selective Sampling

Let us now turn our attention to effects of sampling error on factor analysis under selective sampling. The previous section focused on the effects of sampling error under random sampling, where each individual in the population has a probability of being included in the sample. By contrast, selective sampling involves the case where each individual does not have an equal probability of being selected; rather, the selection of a given individual for a sample is dependent at least in part on that individual's level or score on one or more attributes. These attributes will be referred to as selection variables. If a given individual's scores on the selection variables satisfy the selection criteria, then that individual may be included in the sample; if the scores do not satisfy the criteria, then the individual may not be included in the sample.

Two basic facts about selective sampling must be recognized: (a) selective sampling occurs routinely in practice; and (b) selective sampling may affects results obtained in factor analysis. To consider the first point, samples of individuals are influenced not only by the actions of the experimenter but also by actions of many others and of institutions, as well as by actions of the prospective members of the sample. The experimenter may choose among sources of subjects for experiments, thus imposing a type of selection. In addition, other individuals and institutions may affect the pool from which an experimenter draws samples. For instance, the student body at a highly prestigious college has undergone a stringent system of selection. This system of

selection not only raises the level of talent of the student body, but also tends to reduce the variability of talent and, thus, of the correlations between measures of attributes. Some colleges of lower prestige suffer restriction on both ends of the range of talent. Top applicants are drained away to the more prestigious colleges while the college of lower prestige rejects less qualified applicants. These effects also reduce the range of talent and reduce the variability and correlations among attributes. Thus, samples drawn from student bodies at colleges surely are characterized by various forms of selection. To consider another example, the body of enlistees in the armed forces tend not only to be of lower talent, but also of lower variability. Any sample drawn from such sources will be affected by the selection that has occurred. Finally, one must keep in mind the potential impact of self-selection. That is, in many situations, prospective members of a sample must take some action (e.g., sign up for an experiment, apply to a college) which influences whether or not they may be included in a sample. Considering all of these potential sources of selection, it is clear that most research sample will have been subjected to some type of selective sampling.

To consider the second point made above (i.e., that selective sampling will affect results of factors analysis), we need only recognize some basic influences of selection. As illustrated in the previous paragraph, selected samples will tend to be characterized by low variability on the selection variables, and in turn on other attributes related to the selection variables. This restriction of the range of scores on these attributes will affect not only the variances of those attributes, but also the correlations (and, thus, covariances) of those attributes. Since factor analysis solutions are defined so as to optimally account for the variation and covariation of the attributes in question, it is clearly possible that any selection process which influences the variances and covariances of those attributes will in turn influence results of factor analyses.

To demonstrate these potential affects more clearly, let us consider two simple illustrations. Some effects of selective sampling by an institution are illustrated in a study by French, Tucker, Newman, and Bobbitt (1952). This project involved a study of the system for selection of Reserved Officer candidates for the United States Coast Guard. Some results are shown in Table 5.5. The large pool of applicants was subjected to three stages of selection, where more information was used at each subsequent state. A quantitative test and a verbal test were used at each stage, and some relevant statistics for these tests are shown in Table 5.5. Note that the mean scores increased while the standard deviations decreased in general from stage to stage. In addition, there was a drastic effect on the correlation between the two tests. This effect on the correlation is in part a result of the two tests being tested directly in the selection process. However, some reduction of the correlation would still have occurred if parallel forms of the tests had been used in making the selections. Note that the final entering class included only

Table 5.5
Sequence of Statistics During Selection for a Quantitative Test and a Verbal Test*

	Applicants	After Initial Selection	After Second Selection	Entering Class
N	2253	968	374	128
Quantitative				
M	38.6	42.6	46.2	48.7
D	9.8	9.3	9.0	9.0
Verbal				
M	54.5	64.8	70.1	75.4
SD	14.5	12.5	12.2	10.9
Correlation	.50	.38	.32	.12

* These data are for the class of 1949 at the United States Coast Guard Academy, see French, et al (1952).

French, John W., Tucker, Ledyard R, Newman, Sidney H., and Bobbitt, Joseph M. A factor analysis of aptitude and achievement entrance tests and course grades at the United States Coast Guard Academy. *Journal of Educational Psychology*, 1952, 43, 65-80.

about $5\frac{1}{2}\%$ of the applicants. Any study using this class of cadets must take into account the effects of the considerable selection which occurred.

A different form of selective sampling is illustrated by some results drawn from the study of intelligence by Thurstone and Thurstone (1941). Results are shown in Table 5.6. These results involve two samples which differ in heterogeneity as to semesters in high school; students in sample 1 were from a single half-grade, while students in sample 2 were from a span of one and one-half grades. To illustrate the effects of this difference in heterogeneity, six tests were selected from the two batteries included in the studies by Thurstone and Thurstone (1941). These six tests were similar in the two batteries, though some editing did occur from the tests given to sample 1 to the test given to sample 2. The two correlation matrices for the six tests were taken from the larger matrices published by Thurstone and Thurstone (1941) and are shown in Table 5.6. Since sample 2 is more heterogeneous than sample 1, we would expect that the correlations for corresponding pairs of tests would be higher for sample 2. With one exception, this pattern holds. This effect would in turn lead us to expect difference in the factor solutions obtained from the two samples. In particular, we would expect two things to occur: (a) that the loadings for the more heterogeneous sample would be larger, since they must account for larger correlations among the tests; and (b) that the factor intercorrelations would be larger, since the correlations among the surface attributes are larger. To determine whether these predictions would hold, factor analyses were performed for the two correlation matrices yielding the factor weights given in Table 5.6. Factors V and W correspond to the Thurstones' interpretation of a verbal factor and a word fluency factor. There is seen to be a tendency toward the weights being larger for the more heterogeneous sample, though the strength of this tendency is mediocre. Reasons for the weakness of this tendency will be explored later in this section. The expectation of a larger interfactor correlation in sample 2 is borne out quite distinctly.

A point of considerable importance is that individuals conducting factor analytic studies should be aware of the effects which may be generated by the nature of the samples they use. It must be recognized that selection of some type occurs almost universally in obtaining samples for such studies, and it is clear that such selection can impact on the nature of the factor analysis results obtained.

5.2.1. Mathematical Framework for Theory of Linear Selective Sampling

Given this state of affairs, we will next develop a conceptual and mathematical framework for systematically examining the effects of selective sampling in factor analysis. The development of this framework involves some concepts additional to the common factor model as it has been to this point. A primary addition is a domain of variables involved directly in selection among individuals in a population. These variables, as noted above, will be called

Table 5.6

Comparison of Statistics for Two Samples Differing in Heterogeneity as to School Grade of Individuals*

Sample 1: 710 Students in Grade VIIIB

	Correlation Matrix						Factor Weights		
Test	1	2	3	4	5	6		V	M
1 Completion	1.00	.68	.72	.35	.37	.31	1	.73	.14
2 Sentences	.68	1.00	.77	.23	.29	.22	2	.87	-.07
3 Vocabulary	.72	.77	1.00	.24	.30	.30	3	.87	-.02
4 First Letters	.35	.23	.24	1.00	.47	.53	4	-.04	.71
5 Four-Letter Words	.37	.29	.30	.47	1.00	.43	5	.09	.57
6 Suffixes	.31	.22	.30	.53	.43	1.00	6	-.01	.66

Correlation Between Factors=.50

Sample 2: 437 Students in Grade VIIA, VIIIB, VIIIA

	Correlation Matrix						Factor Weights		
Test	1	2	3	4	5	6		V	M
1 Completion	1.00	.77	.78	.43	.35	.43	1	.84	.01
2 Sentences	.77	1.00	.83	.42	.36	.41	2	.92	-.05
3 Vocabulary	.78	.83	1.00	.47	.42	.48	3	.86	.06
4 First Letters	.43	.42	.47	1.00	.65	.56	4	.03	.76
5 Four-Letter Words	.35	.36	.42	.65	1.00	.51	5	-.04	.77
6 Suffixes	.43	.41	.48	.56	.51	1.00	6	.13	.59

Correlation Between Factors=.62

* These data were collected in 1939 by L. L. Thurstone and Thelma G. Thurston from the studies in the Chicago Schools. At that time advancement in school was by half grade. The correlation matrices were taken from: Thurston, L. L. & Thurston , Thelma G. Factor Studies of Intelligence. Psychometric Monograph No. 2. Chicago, University of Chicago Press, 1941.

selection variables. Selection variables may be of many types. Material used in selection among applicants for admission to a college, or for a job may be characterized as selection variables. Less concrete selection variables are the variety of matters considered by individuals in selecting whether or not to apply. These are variables of selfselection. To consider some other examples, some claim may be made that selective mating though years may have increased the variance on some attributes and led to higher correlations among these attributes. An experimenter may choose among several sources of subjects for experiments. All of these possibilities will be included here under the concept of selection variables.

Another important notion in the present framework is that of a subpopulation. For present purposes, selection will be represented as operating at the level of populations. That is, when selection is applied to some population, those individuals who satisfy the selection criteria comprise a subpopulation. The effects of selection in factor analysis can be examined by determining how the factors in the general population are affected by selection of a subpopulation. In practice one would actually be observing a sample drawn from the subpopulation, and the results of factor analysis of that sample would be subject to the additional effects of random sampling discussed in the previous section. Thus, to examine the effects of selection, it is necessary to employ a framework defined at the level of populations and subpopulations.

A schemata for a theory of selective sampling is shown in Figure 5.5. This schemata is a modification of the schemata shown in Figure 1.2, which represented the basic common factor model and which was discussed in Chapter 1. The common factor model involves internal attributes (or factors) pictured on the left, and surface attributes pictured on the right. A domain of selection variables is pictured in the lower middle. Solid lines with arrowheads are drawn from the internal attributes to the surface attributes to indicate possible causal relations. These effects correspond to the common and unique factor weights. It is important to understand the nature of the relations of the selection variables to the other attributes included in this framework. No causal relations are assumed for the selection variables. Rather, the selection variables are assumed to have associational relations with the internal attributes. These non-zero associational relations are represented in the schemata by dashed lines with double arrowheads. Note also that selection variables may be associated with each other. In the schemata three selection variables are pictured which have associational relations with the two common factors. In addition, one association is indicated between one of the selection variables and one of the unique factors. Of course, other associations are possible between selection variables and all unique factors; however, only one is shown. Relations of unique factors with selection variables are special cases and result in complex effects to be described later. Primary attention will be given to effects of relations of selection variables with common factors. It is these relations which underlie the

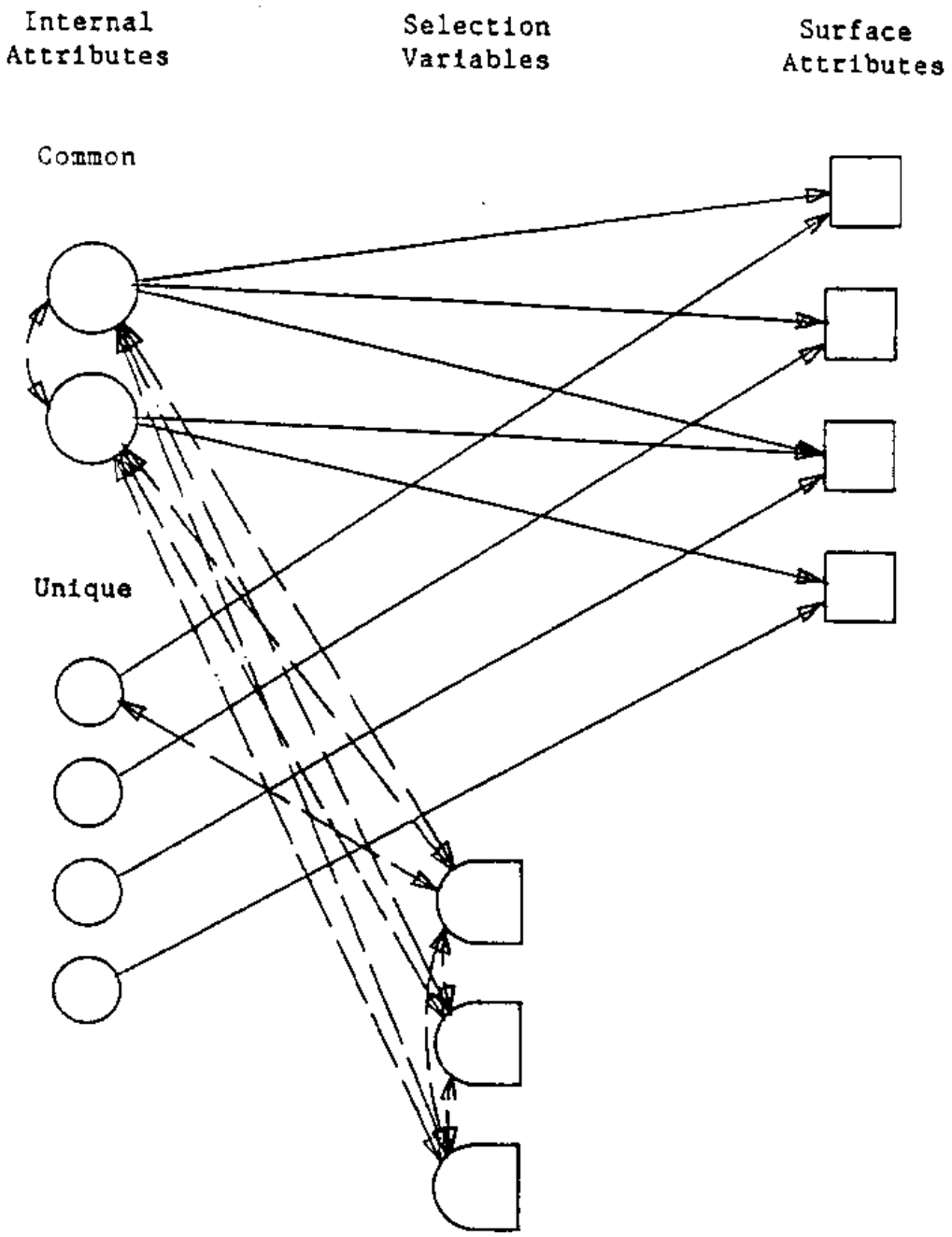

Figure 5.5: Schemata for Theory of Selective Sampling

effects of selection on the observed data. That is, according to the schemata, a selection process based on the selection variables will yield a subpopulation in which the underlying factors have been influenced by selection. Since the factors directly influence the surface attributes, the statistical properties of the surface attributes will also have been influenced by selection. Thus, the relations of the surface attributes to the selection variables are taken to be mediated by relations of the factors to the selection variables.

We now begin the development of a mathematical framework to represent these concepts and relations. The objective is to determine how the parameters of the common factor model, primarily the common factor weights and intercorrelations, are affected by selective sampling; i.e., how the parameters in a subpopulation are different from those in the general population, and how the parameters in different subpopulations are different from each other. We will continue to employ notation used in this and previous chapters for the common factor model, and we will extend this notation and define new notation as necessary to incorporate the selection variables and subpopulations into this framework. A vector of scores on the selection variables is designated $\underline{\boldsymbol{v}}$. In the case considered here, the factor scores are taken to be linearly related to the selection variables. There is a close similarity of the situation to linear, homoscedastic, multiple regression of the factor scores on the selection variables. However, no causal relation is implied; only a relation of association is taken to exist. Even so, the factor scores may be divided into two components: one component of composite measures from the selection variables and a second component of discrepancies. Stated differently, the factor scores may be divided into a component which can be accounted for by their linear association with the selection variables, and a component which cannot be accounted for by that association, The vector of scores on the discrepancy component will be designated $\underline{\ddot{\boldsymbol{v}}}$. These discrepancy scores are taken to be unrelated to the selection variables.

Given these concepts, the relation of a factor score vector, $\underline{\boldsymbol{x}}$, to the selection variables vector, $\underline{\boldsymbol{v}}$, is given by:

$$\underline{\boldsymbol{x}} = \underline{\boldsymbol{v}}\boldsymbol{W}' + \underline{\ddot{\boldsymbol{v}}} \tag{5.23}$$

where $\boldsymbol{W}$ is a weight matrix representing weights applied to the selection variables to produce that part of the factor scores which is linearly associated with the selection variables. In further developing this approach, it will be necessary to employ notation to differentiate among subpopulations. The reader should keep in mind that, as noted above, different selection processes or criteria would yield different subpopulations. The letter *p* will be used as an index for subpopulation. Given this, we define $\boldsymbol{\Sigma}_{xxp}$ as the subpopulation covariance matrix among the factor scores. The assumption stated in the previous paragraph that the discrepancies are unrelated to the selection variables implies that:

$$\boldsymbol{\Sigma}_{v\ddot{v}p} = \mathbf{0} \tag{5.24}$$

for all *p*. A further assumption of homoscedaticity implies that the covariance matrix for the discrepancies is constant across subpopulations. That is,

$$\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}p} = \boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} \tag{5.25}$$

for all *p*.

Let us briefly consider relations between mean vectors on the factors and selection variables. The mean factor score vector, $\boldsymbol{\mu}_{xp}$, for subpopulation *p*, is given, according to Eq. (5.23), by:

$$\boldsymbol{\mu}_{xp} = \boldsymbol{\mu}_{vp}\boldsymbol{W} + \boldsymbol{\mu}_{\ddot{v}p} \tag{5.26}$$

where $\boldsymbol{\mu}_{vp}$ is the mean score vector for the selection variables and $\boldsymbol{\mu}_{\ddot{v}p}$ is the mean discrepancy score vector. No assumption is made in the present context as to the constancy of the mean vectors. Though some assumptions may be made in particular contexts about the mean vectors, emphasis here is on the covariance and correlation relations. In the linear system the covariance and correlation relations are independent of the mean vectors.

Fundamental covariance relations are considered next. The relation of the factor score covariance matrix $\boldsymbol{\Sigma}_{xxp}$ in subpopulation *p* to the selection variable covariance matrix $\boldsymbol{\Sigma}_{vvp}$ in subpopulation *p* and the covariance matrix $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$ is developed from Eq. (5.23) with modifications indicated by Eqs. (5.24) and (5.25). This yields:

$$\boldsymbol{\Sigma}_{xxp} = \boldsymbol{W}\boldsymbol{\Sigma}_{vvp}\boldsymbol{W}' + \boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} \tag{5.27}$$

Another fundamental covariance relation is represented by the common factor model, which is modified in the present context to apply to each subpopulation *p*. The model would then be written as follows:

$$\boldsymbol{\Sigma}_{zzp} = \boldsymbol{\Omega}\boldsymbol{\Sigma}_{xxp}\boldsymbol{\Omega}' \tag{5.28}$$

It is especially important to note that the factor weight matrix, $\boldsymbol{\Omega}$, is not changed in the modification; i.e., there is no subscript *p* on $\boldsymbol{\Omega}$ to indicate a different factor weight matrix for each subpopulation. This is a very strong assumption: the basic attribute score equation is invariant over selective sampling. However, this assumption follows quite naturally from common factor theory, where it is assumed that the basic model defining the linear relations between the modeled attributes and the internal attributes (see Eq. (3.3)) applies to every individual in a population; this would not change when individuals are selected into subpopulations. It is interesting to note here that any particular individual may be a member of several subpopulations; from that view, it would be difficult to argue for different weight

matrices in different subpopulations. However, in some contexts objections may be raised that factor patterns may change from one subpopulation to another, such as for males vs. females. However, some aspects of the current framework may account for such differences. Such difference will be important in factor analysis in multiple populations, which will be discussed later in this section, as well as in Chapter 19.

It will be useful to define a basic, or reference subpopulation. This will be a subpopulation for which the variances and covariances of the selection variables are zero. Such a subpopulation would be the result of "absolute" selection; i.e., in order for an individual to be selected, the individual would have to have exactly a particular set of scores on the selection variables. Let this subpopulation be designated as subpopulation *o* with *p* = *o*. We could then write

$$\boldsymbol{\Sigma}_{vvo} = \mathbf{0} \tag{5.29}$$

Combining this definition with the relationship given by Eq. (5.27), we obtain

$$\boldsymbol{\Sigma}_{xxo}=\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} \tag{5.30}$$

This indicates that in the case of absolute selection, the covariance matrix for the factors in the subpopulation will equal the covariance matrix for the discrepancies. Combining this result with Eq. (5.28), we obtain

$$\boldsymbol{\Sigma}_{zzo} = \boldsymbol{\Omega}\, \boldsymbol{\Sigma}_{xxo} \boldsymbol{\Omega}' = \boldsymbol{\Omega} \boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} \boldsymbol{\Omega}' \tag{5.31}$$

This equation shows the representation of the common factor model in the subpopulation in the case of absolute selection. According to this result, the factor weights are unchanged, and the factor covariances will equal the covariances of the discrepancies.

To this point we have not differentiated among common, specific, and error factors in the development of this mathematical framework. However, since different assumptions may be made and different relations may hold for the different classes of factors, it is useful to do so. The factor weight matrix $\boldsymbol{\Omega}$ and covariance matrix $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$ may be expressed in terms of sections for the common, specific, and error factors. For the factor weight matrix, the following partitioning was presented in Eq. (3.6):

$$\boldsymbol{\Omega} = [\boldsymbol{B}, \boldsymbol{\Xi}, \boldsymbol{E}] \tag{5.32}$$

To distinguish different classes of factors for $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$, it is necessary to employ a second level of subscripts. These are β for common factors, ξ for the specific factors, and ϵ for the error factors. The matrix $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$ can be represented as follows:

$$\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\ddot{v}_\beta \ddot{v}_\beta} & \boldsymbol{\Sigma}_{\ddot{v}_\beta \ddot{v}_\xi} & \boldsymbol{\Sigma}_{\ddot{v}_\beta \ddot{v}_\epsilon} \\ \boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\beta} & \boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\xi} & \boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\epsilon} \\ \boldsymbol{\Sigma}_{\ddot{v}_\epsilon \ddot{v}_\beta} & \boldsymbol{\Sigma}_{\ddot{v}_\epsilon \ddot{v}_\xi} & \boldsymbol{\Sigma}_{\ddot{v}_\epsilon \ddot{v}_\epsilon} \end{bmatrix} \tag{5.33}$$

Let us examine the submatrices within defined in this equation. Based on Eq. (5.27), matrix $\boldsymbol{\Sigma}_{\ddot{v}_\beta \ddot{v}_\beta}$ is the covariance matrix among the common factors in the basic subpopulation. Let us consider the common factors to be standardized in that subpopulation, and let us define matrix $\boldsymbol{\Phi}_o$ as the correlation matrix among the common factors in the basic subpopulation. That is,

$$\boldsymbol{\Phi}_o = \boldsymbol{\Sigma}_{\ddot{v}_\beta \ddot{v}_\beta} \tag{5.34}$$

Applying the conventional definitions of factor theory to the basic subpopulation, covariances of the specific factors with the common factors are assumed to equal zero, as are the covariances of the error factors with the common factors and the specific factors. Thus, we can write

$$\boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\beta} = \boldsymbol{\Sigma}_{\ddot{v}_\epsilon \ddot{v}_\beta} = \boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\epsilon} = \mathbf{0} \tag{5.35}$$

Covariances among the specific factors and among the error factors are also assumed to equal zero. Considering these to be standardized in the reference subpopulation allows us to write

$$\boldsymbol{\Sigma}_{\ddot{v}_\xi \ddot{v}_\xi} = \boldsymbol{\Sigma}_{\ddot{v}_\epsilon \ddot{v}_\epsilon} = \boldsymbol{I} \tag{5.36}$$

Applying the foregoing definition and assumptions to Eq. (5.33), the matrix $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$ can be represented as follows:

$$\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}} = \begin{bmatrix} \boldsymbol{\Phi}_o & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I} \end{bmatrix} \tag{5.37}$$

Recall that the representation of the common factor model in the basic subpopulation was given in Eq. (5.31). A more detailed representation of this model can now be obtained by substitution from Eqs. (5.32) and (5.37) into (5.31). This yields the following:

$$\boldsymbol{\Sigma}_{zzo} = \boldsymbol{B}\boldsymbol{\Phi}_o\boldsymbol{B}' + \boldsymbol{\Xi}^2 + \boldsymbol{E}^2 \tag{5.38}$$

When the specific factors and the error factors have zero associations with the selection variables, the specific variances and error variances given in $\boldsymbol{\Xi}^2$ and $\boldsymbol{E}^2$, respectively, may be combined into unique variances. (As previously noted, when specific factors have nonzero relations with the selection variables, considerable complexities arise, which will be discussed later in this section.) The combining of specific and error variances to form unique variances was defined in Eq. (3.27), which is repeated here for convenience:

$$\boldsymbol{U^2 = \Xi^2 + E^2} \tag{5.39}$$

Substituting from this equation into Eq. (5.38) yields the following:

$$\boldsymbol{\Sigma_{zzo} = B\Phi_o B' + U^2} \tag{5.40}$$

This equation provides a more detailed (than Eq. (5.31)) expression of the common factor model in the basic subpopulation. It indicates that the common factor weights and unique variances in that subpopulation will be the same as in the general population, but the factor intercorrelations, given in $\boldsymbol{\Phi_o}$, may be different.

Given this representation for the basic subpopulation, let us now consider further examination of the covariance relations defined in Eqs. (5.27) and (5.28) for any subpopulation *p*. It is useful to define partitioned forms of some of the matrices in those equations. In particular, the weight matrix, $\boldsymbol{W}$, expressing the relations of the factors to the selection variables, can be defined as having sections corresponding to the several types of factors. This would be represented as follows:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W_\beta} \\ \boldsymbol{W_\xi} \\ \boldsymbol{W_\epsilon} \end{bmatrix} \tag{5.41}$$

It will be assumed that the weights for the error factors are zero; i.e., there is no association between the selection variables and the error of measurement factors. Logically, the errors of measurement in the surface attributes should be independent of the selection variables. Thus, we can write

$$\boldsymbol{W_\epsilon = 0} \tag{5.42}$$

Considering the relation of the specific factors to the selection variables, when these associations are zero then section $\boldsymbol{W_\xi}$ will equal zero. For present purposes, this will be taken to be the case. However, this submatrix will be treated later in this section as being, possibly, nonzero, to represent the case of associations of specific factors to the selection variables.

The preceding developments provide a basis for consideration of the effects of selective sampling on the factor covariance matrix. This necessitates the sectioning of matrix $\boldsymbol{\Sigma_{xxp}}$, the factor covariance matrix for subpopulation *p*, into sections representing the several types of factors as follows:

$$\boldsymbol{\Sigma_{xxp}} = \begin{bmatrix} \boldsymbol{\Sigma_{x_\beta x_\beta p}} & \boldsymbol{\Sigma_{x_\beta x_\xi p}} & \boldsymbol{\Sigma_{x_\beta x_\epsilon p}} \\ \boldsymbol{\Sigma_{x_\xi x_\beta p}} & \boldsymbol{\Sigma_{x_\xi x_\xi p}} & \boldsymbol{\Sigma_{x_\xi x_\epsilon p}} \\ \boldsymbol{\Sigma_{x_\epsilon x_\beta p}} & \boldsymbol{\Sigma_{x_\epsilon x_\xi p}} & \boldsymbol{\Sigma_{x_\epsilon x_\epsilon p}} \end{bmatrix} \tag{5.43}$$

It will be useful to examine relations between sections of $\boldsymbol{\Sigma}_{xxp}$ and matrices $\boldsymbol{W}$, $\boldsymbol{\Sigma}_{vvp}$, and $\boldsymbol{\Sigma}_{\ddot{v}\ddot{v}}$. This can be achieved via Eqs. (5.27), (5.33), and (5.43), and results in the following equations:

$$\boldsymbol{\Sigma}_{x_\beta x_\beta p} = \boldsymbol{W}_\beta \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\beta' + \boldsymbol{\Phi}_o \tag{5.44}$$

$$\boldsymbol{\Sigma}_{x_\xi x_\beta p} = \boldsymbol{W}_\xi \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\beta' \tag{5.45}$$

$$\boldsymbol{\Sigma}_{x_\epsilon x_\beta p} = \boldsymbol{W}_\epsilon \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\beta' \tag{5.46}$$

$$\boldsymbol{\Sigma}_{x_\xi x_\xi p} = \boldsymbol{W}_\xi \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\xi' + \boldsymbol{I} \tag{5.47}$$

$$\boldsymbol{\Sigma}_{x_\xi x_\epsilon p} = \boldsymbol{W}_\xi \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\epsilon' \tag{5.48}$$

$$\boldsymbol{\Sigma}_{x_\epsilon x_\epsilon p} = \boldsymbol{W}_\epsilon \boldsymbol{\Sigma}_{vvp} \boldsymbol{W}_\epsilon' + \boldsymbol{I} \tag{5.49}$$

With the assumption of Eq. (5.42) that the weights for the error of measurement factors equal zero, the covariance of the error factors with the other factors, given by Eqs. (5.46) and (5.48) remain zero. That is,

$$\boldsymbol{\Sigma}_{x_\epsilon x_\beta p} = \boldsymbol{\Sigma}_{x_\epsilon x_\xi p} = \boldsymbol{0} \tag{5.50}$$

The covariance matrix among the error of measurement factors remains an identity matrix:

$$\boldsymbol{\Sigma}_{x_\epsilon x_\epsilon p} = \boldsymbol{I} \tag{5.51}$$

More complex is the status of the covariances involving the specific factors; as shown in Eq. (5.45) and (5.47), these covariances will depend upon the specific factors weight matrix, $\boldsymbol{W}_\xi$. As noted above, we will assume for present purposes that $\boldsymbol{W}_\xi$ equals zero, meaning that there is no relation between the selection variables and the specific factors.

A very important and interesting topic for mathematical consideration is that of standardization of factor scores and of attribute scores in subpopulations. Usually, factor analytic procedures define the factor scores to be standardized. When factor analysis is applied to a correlation matrix, the attribute scores have also been standardized. It will be shown that these standardizations affect comparisons between results obtained from studies in different subpopulations. That is, factor analysis results obtained from factor analyzing data drawn from different subpopulations may be different simply as a result of the standardizations of factors and attributes within those subpopulations.

Let us consider first the standardization of common factor scores in a subpopulation. Let us define a diagonal matrix $[\boldsymbol{\Sigma}_d]_{x_\beta x_\beta p}$ which contains the variances of the common factors in

subpopulation *p*. That is,

$$[\boldsymbol{\Sigma}_d]_{x_\beta x_\beta p} = \boldsymbol{Diag}(\boldsymbol{\Sigma}_{x_\beta x_\beta p}) \tag{5.52}$$

Then the correlation matrix among the common factors in subpopulation *p* can be defined as follows:

$$\boldsymbol{\Phi_p} = [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta p} \boldsymbol{\Sigma}_{x_\beta x_\beta p} [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta p} \tag{5.53}$$

A compensating rescaling of the common factor weight matrix involves definition of a matrix $\widetilde{B_p}$ as follows:

$$\widetilde{B_p} = \boldsymbol{B}[\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta p} \tag{5.54}$$

From the definitions given in Eqs. (5.53) and (5.54) we can write

$$\boldsymbol{B\Sigma}_{x_\beta x_\beta p}\boldsymbol{B'} = \widetilde{\boldsymbol{B}}_{\boldsymbol{p}}\boldsymbol{\Phi}_p\widetilde{\boldsymbol{B}}_{\boldsymbol{p}}{}' \tag{5.55}$$

When the specific factor weights in $\boldsymbol{W_\xi}$ are zero, the factor equation for covariance matrix $\boldsymbol{\Sigma}_{zzp}$, obtained from Eqs. (5.28), (5.32), (5.39), and (5.43), is

$$\boldsymbol{\Sigma}_{zzp} = \boldsymbol{B\Sigma}_{x_\beta x_\beta p}\boldsymbol{B'} + \boldsymbol{U^2} \tag{5.56}$$

The usual equation for factor analysis of a covariance matrix can then be obtained by combining equations (5.55) and (5.56), yielding the following:

$$\boldsymbol{\Sigma}_{zzp} = \widetilde{\boldsymbol{B}}_{\boldsymbol{p}}\boldsymbol{\Phi}_p\widetilde{\boldsymbol{B}}_{\boldsymbol{p}}{}' + \boldsymbol{U^2} \tag{5.57}$$

Note that matrices $\widetilde{B_p}$ and $\boldsymbol{\Phi_p}$ represent the common factor weights and intercorrelations in subpopulation *p*, and that they are particularly affected by the standardization of the common factor scores within the subpopulation, as given in Eqs. (5.53) and (5.54).

Now that we have developed a representation of the common factor model within a given subpopulation *p*, an important topic concerns the relations between the factors obtained in the subpopulation and factors obtained from a different subpopulation. Researchers often wish to study differences in factor structures across populations. It will now be shown that certain types of such differences can be represented using the present framework of subpopulations. Let *q* designate a second subpopulation. Eq. (5.54) may be written for subpopulation q as follows:

$$\widetilde{B_q} = \boldsymbol{B}[\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta q} \tag{5.58}$$

We wish to determine the relation between $\widetilde{B_q}$ and $\widetilde{\boldsymbol{B}}_{\boldsymbol{p}}$; i.e., between the factor weights obtained from the two subpopulations. This relation can be found from Eqs. (5.54) and (5.58) to be

$$\widetilde{B}_q = \widetilde{\boldsymbol{B}}_p [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta p} [\boldsymbol{\Sigma}_d]^{\frac{1}{2}}_{x_\beta x_\beta q} \tag{5.59}$$

Since both $[\boldsymbol{\Sigma}_d]_{x_\beta x_\beta p}$ and $[\boldsymbol{\Sigma}_d]_{x_\beta x_\beta q}$ are diagonal matrices, their product is also a diagonal matrix which may be designated $\boldsymbol{\Psi}_{pq}$:

$$\boldsymbol{\Psi}_{pq} = [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{x_\beta x_\beta p} [\boldsymbol{\Sigma}_d]^{\frac{1}{2}}_{x_\beta x_\beta q} \tag{5.60}$$

Substituting from this equation into Eq. (5.59) yields

$$\widetilde{B}_q = \widetilde{\boldsymbol{B}}_p \boldsymbol{\Psi}_{pq} \tag{5.61}$$

This is an important equation because it indicates that the effect of standardization of common factors within subpopulations is to make subpopulation factor weight matrices proportional by columns. This has implications for a number of subsequent topics such as comparing results from different factor analytic studies and factor analysis in multiple populations.

Let us next consider the second type of standardization which affects results obtained from subpopulations: standardization of attribute scores. Such standardization involves converting the covariance matrix among attribute scores into a correlation matrix. Let us define a matrix $[\boldsymbol{\Sigma}_d]_{zzp}$ containing variances of the modeled attributes:

$$[\boldsymbol{\Sigma}_d]_{zzp} = \boldsymbol{Diag}(\boldsymbol{\Sigma}_{zzp}) \tag{5.62}$$

The population correlation matrix among the attribute scores is then given by

$$\boldsymbol{R}_{zzp} = [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{zzp} \boldsymbol{\Sigma}_{zzp} [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{zzp} \tag{5.63}$$

This transformation must be applied to the common factor weight matrix to yield a common factor weight matrix for analysis of the correlation matrix. This can be seen by applying the transformation shown in Eq. (5.63) to both sides of Eq. (5.57). This leads us to define the common factor weight matrix for the analysis of the correlation matrix as follows:

$$\widetilde{\widetilde{\boldsymbol{B}}}_p = [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{zzp} \widetilde{\boldsymbol{B}}_p \tag{5.64}$$

The unique variances must also be transformed, thus yielding

$$\widetilde{\boldsymbol{U}^2_p} = [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{zzp} \boldsymbol{U}^2 [\boldsymbol{\Sigma}_d]^{-\frac{1}{2}}_{zzp} \tag{5.65}$$

Substitution from Eqs. (5.64) and (5.65) into Eq. (5.57) yields the usual factor equation for analysis of a correlation Matrix:

$$\boldsymbol{R}_{zzp} = \widetilde{\widetilde{\boldsymbol{B}}}_p \boldsymbol{\Phi}_p \widetilde{\widetilde{\boldsymbol{B}}}_p{}' + \widetilde{\widetilde{\boldsymbol{U}}}^2 \tag{5.66}$$

Note that the correlation matrix, $\boldsymbol{\Phi}_p$, for the common factors in subpopulation p is not transformed when the attributes are standardized. However, matrices $\widetilde{\widetilde{\boldsymbol{B}}}_p$ and $\widetilde{\widetilde{\boldsymbol{U}}}_p^2$, which represent the usual results from factoring a correlation matrix, have been affected by this standardization.

To complete the discussion of this issue of standardization of attributes, let us consider how the relation between solutions from two different subpopulations is affected. Let us define a factor weight matrix $\widetilde{\widetilde{\boldsymbol{B}}}_q$ for subpopulation q. Equation (5.64) allows us to write

$$\widetilde{\widetilde{\boldsymbol{B}}}_q = [\boldsymbol{\Sigma}_d]_{zz\mathbf{q}}^{-\frac{1}{2}} \widetilde{B}_q \tag{5.67}$$

Substituting into this equation from Eq. (5.61) yields

$$\widetilde{\widetilde{\boldsymbol{B}}}_q = [\boldsymbol{\Sigma}_d]_{zz\mathbf{q}}^{-\frac{1}{2}} \widetilde{\boldsymbol{B}}_p \boldsymbol{\Psi}_{pq} \tag{5.68}$$

Finally, substituting from Eq. (5.64) into this equation yields

$$\widetilde{\widetilde{\boldsymbol{B}}}_q = [\boldsymbol{\Sigma}_d]_{zz\mathbf{q}}^{-\frac{1}{2}} [\boldsymbol{\Sigma}_d]_{zzp}^{\frac{1}{2}} \widetilde{\widetilde{\boldsymbol{B}}}_p \boldsymbol{\Psi}_{pq} \tag{5.69}$$

This equation shows that when attributes are standardized within subpopulations, the relation between common factor weight matrices obtained from the subpopulations will be affected by that differential standardization. In particular, there will be a proportionality of rows in the two matrices. Combining this result with the effect of standardization of common factors, also shown in Eq. (5.69), it can be seen that standardization of factors and attributes within subpopulations causes differences in factor weights obtained when factor analysis is applied to data from those subpopulations. Specifically, there is a proportional rescaling of both rows and columns of the common factor weight matrix. This result has considerable implications for the matching of factor matrices from different studies. When these studies involve analysis of correlation matrices, the factor matrices would have to be rescaled by rows so that there is a common unit of measure between the two studies for each attribute.

5.2.2. Demonstration of Effects of Selective Sampling

We will now describe and report results of a simulated demonstration of selective sampling designed to illustrate various features of the theory of linear selective sampling. This example was designed to parallel in a number of aspects the results given in Table 5.6 from the studies by Thurstone and Thurstone (1941). This demonstration includes no effects of random sampling; that is, it shows the effects of selection in the context of subpopulations as described above. Further, no model error was incorporated into the demonstration; i.e., there is no lack of

fit of the factor model to the simulated data. Consequently, theoretical relations are fitted precisely. This demonstration reveals possible unobservable relations and intermediate results not available in the results given in Table 5.6 for the Thurstone and Thurstone (1941) data.

Parameters for the demonstration are given in Table 5.7. The basic common factor matrix, $\boldsymbol{B}$, representing the basic subpopulation, and uniquesses, $\boldsymbol{U^2}$, are given at the upper left. Diagonal entries in $\boldsymbol{U^2}$ are listed in a column. There were six attributes and two common factors. Three of the attributes had high weights on the first factor and zero weights on the factor. The reverse was true for the last three attributes, having zero weights on the first factor and high weights on the second factor. The basic covariance matrix among common factors, $\boldsymbol{\Phi}_o$, is given at the upper right. The foregoing matrices were used to compute the basic covariance matrix among attributes, $\boldsymbol{\Sigma}_{zzo}$, by Eq. (5.42). This matrix is given in the center of Table 5.7. One selection variable was used in the demonstration. This variable might be thought of as general mental development. To parallel the Thurstone and Thurstone (1941) studies, the standard deviation on this selection variable was taken to equal 1.00 (for one semester for the subjects in sample 1 of the Thurstone's studies) for subpopulation 1, and equal to 3.00 (for three semesters for the subjects in sample 2 of the Thurstone's studies) for subpopulation 2. These standard deviations and associated variance are given next to the bottom of Table 5.7. The covariance matrix, $\boldsymbol{\Sigma}_{vvp}$, is 1 $\times$ 1 and contains for each subpopulation the variance of the selection variable for the subpopulation. The weight matrix, $\boldsymbol{W}_\beta$, for the common factors is given at the bottom of Table 5.7.

Covariance relations in the two subpopulations are given in Table 5.8. The top half of this table is for subpopulation 1 and the bottom half is for subpopulation 2. For both subpopulations, the covariance matrix for the common factors is at the upper left of the halves for the subpopulations. The standard deviations are given below each covariance matrix. These covariance matrices are $\boldsymbol{\Sigma}_{x_\beta x_\beta 1}$ and $\boldsymbol{\Sigma}_{x_\beta x_\beta 2}$. They were computed by Eq. (5.44). Note that the variances and covariances have increased from the values in the basic covariance matrix in Table 5.7. These increases are greater for subpopulation 2 than for subpopulation 1, this being the result of the variance on the selection variable being greater for subpopulation 2 than for subpopulation 1.

The correlation matrices among the common factors are given to the right of the covariance matrices. These are matrices $\boldsymbol{\Phi}_1$, and $\boldsymbol{\Phi}_2$; they were computed by Eq. (5.53). Note that the correlations of .50 and .64 between the common factors are greater than the corresponding correlation of .47 for the basic population given in Table 5.7. These two values correspond closely to the correlations for the Thurstone's results given in Table 5.6.

Covariance matrices among the attributes are given in Table 5.8 for both subpopulations.

Table 5.7
Parameters for Demonstration of Effects of Selective Sampling in Factor Analysis

Basic Factor Weight Matrix

	Factor		Uniqueness	
Attribute	1	2	Loading	Variance
1	.90	.00	.44	.19
2	.70	.00	.71	.51
3	.50	.00	.87	.75
4	.00	.60	.80	.64
5	.00	.70	.71	.51
6	.00	.80	.60	.36

Basic Covariance Matrix Among Common Factors

Factor	1	2
1	1.00	.47
2	.47	1.00

Basic Covariance Matrix Among Attributes

Attribute	1	2	3	4	5	6
1	1.00					
2	.63	1.00				
3	.45	.35	1.00			
4	.25	.20	.14	1.00		
5	.30	.23	.16	.42	1.00	
6	.34	.26	.19	.48	.56	1.00

Subpopulation SD and Variance on Selection Variable

Subpopulation	SD	Variance
1	1.00	1.00
2	3.00	9.00

Weight Matrix of Common Factors on Selection Variable

Common Factor	Weight
1	.20
2	.30

These are matrices $\boldsymbol{\Sigma}_{zz1}$, and $\boldsymbol{\Sigma}_{zz2}$. They were obtained by Eq. (5.56) using the basic factor matrix $\boldsymbol{B}$ and uniqueness $\boldsymbol{U^2}$ given in Table 5.7 and the covariance matrices among the factors in Table 5.8, which were described previously. The variances and covariances are larger for subpopulation 2 than for subpopulation 1. The factor matrices for the covariance matrices were obtained by Eq. (5.54). These are matrices $\widetilde{\boldsymbol{B}}_{\boldsymbol{1}}$ and $\widetilde{B}_{\boldsymbol{2}}$. These factor matrices along with the correlation matrices among the common factors and the uniqueness reproduce the covariance matrices according to Eq. (5.57). Note that the factor loadings are larger for subpopulation 2 than for subpopulation 1. The relation between the factor weights is given by Eq. (5.61). This relation will be discussed further in connection with the graphs in subsequent figures.

Table 5.9 presents the correlation relations in the subpopulations. The correlation matrices are $\boldsymbol{R}_{zz1}$ and $\boldsymbol{R}_{zz2}$ which were computed by Eq. (5.63). As in the Thurstone data given in Table 5.6, the correlations are greater for subpopulation 2 than for subpopulation 1. The common factor matrices obtained by factor analyzing these correlation matrices are given on the right of each correlation matrix. These are matrices $\widetilde{\widetilde{\boldsymbol{B}}}_{\boldsymbol{1}}$ and $\widetilde{\widetilde{\boldsymbol{B}}}_{\boldsymbol{2}}$. These factor matrices are related to the factor matrices in Table 5.8 for covariance matrices by Eq. (5.64). The factor weights are greater for subpopulation 2 than for subpopulation 1. The factor weights for the correlation matrices will be compared graphically.

Figures 5.6 and 5.7 present graphical comparisons of factor weights in two samples or two subpopulations. A comparison of the factor weights for the Thurstone studies is given in Figure 5.6. The graph on the left is for factor V (verbal ability) and the graph on the right is for factor W (word fluency). In each of these graphs there is a point for each attribute with coordinates equal to the factor loadings in the two samples, A line of mutual relation is drawn on each graph. The slopes of these lines of relation are greater than unity indicating that the factor weights are larger for sample 2 than for sample 1. This corresponds to the greater heterogeneity of sample 2 than of sample 1 with respect to the grades in school. The dispersion of the points from the lines of relation may be due to both sampling variability and to editing of the tests between the two experiments.

It is interesting to compare these results to those shown in Figure 5.7 which presents factor weights for the two subpopulations in the simulation study. The upper pair of graphs are for the factor weights obtained from the covariance matrices while the lower pair of graphs are for the factor weights obtained from the correlation matrices. These factor weights themselves are shown in Table 5.8 and 5.9, respectively. As in Figure 5.6, there is a point on each graph for each attribute with coordinates equal to the factor weights in the two subpopulations. The graphs on the left are for factor 1 and the graphs on the right are for factor 2. A line of relation is drawn on each graph with slopes as indicated.

Table 5.8
Covariance Relations in Subpopulations

SUBPOPULATION 1

Covariance and Correlation Matrices Among Common Factors

Covariance Matrix			Correlation Matrix		
Factor	1	2	Factor	1	2
1	1.04	.53	1	1.00	.50
2	.53	1.09	2	.50	1.00
SD	1.02	1.04			

Covariance Matrix Among Attributes							Factor Matrix	
Attribute	1	2	3	4	5	6	1	2
1	1.03						1 .92	.00
2	.66	1.02					2 .71	.00
3	.47	.36	1.01				3 .51	.00
4	.29	.22	.16	1.03			4 .00	.63
5	.33	.26	.19	.46	1.04		5 .00	.73
6	.38	.30	.21	.52	.61	1.06	6 .00	.84
SD	1.02	1.01	1.00	1.02	1.02	1.03		

SUBPOPULATION 2

Covariance and Correlation Matrices Among Common Factors

CovarianceMatrix			Correlation Matrix		
Factor	1	2	Factor	1	2
1	1.36	1.01	1	1.00	.64
2	1.01	1.81	2	.64	1.00
SD	1.17	1.35			

Covariance Matrix Among Attributes							Factor Matrix	
Attribute	1	2	3	4	5	6	1	2
1	1.29						1 1.05	.00
2	.86	1.18					2 .82	.00
3	.61	.48	1.09				3 .58	.00
4	.55	.42	.30	1.29			4 .00	.81
5	.64	.49	.35	.76	1.40		5 .00	.94
6	.73	.57	.40	.87	1.01	1.52	6 .00	1.08
SD	1.14	1.08	1.00	1.14	1.18	1.23		

Table 5.9
Correlation Relations in Subpopulations

SUBPOPULATION 1

	Correlation Matrix Among Attributes							Factor Matrix	
Attribute	1	2	3	4	5	6		1	2
1	1.00						1	.90	.00
2	.64	1.00					2	.71	.00
3	.46	.36	1.00				3	.51	.00
4	.28	.22	.16	1.00			4	.00	.62
5	.32	.25	.18	.44	1.00		5	.00	.72
6	.37	.29	.21	.50	.58	1.00	6	.00	.81

Correlation Between Factors=.50

SUBPOPULATION 2

	Covariance Matrix Among Attributes							Factor Matrix	
Attribute	1	2	3	4	5	6		1	2
1	1.00						1	.92	.00
2	.70	1.00					2	.75	.00
3	.52	.42	1.00				3	.56	.00
4	.42	.34	.26	1.00			4	.00	.71
5	.47	.39	.29	.57	1.00		5	.00	.80
6	.52	.42	.31	.62	.70	1.00	6	.00	.87

Correlation Between Factors=.64

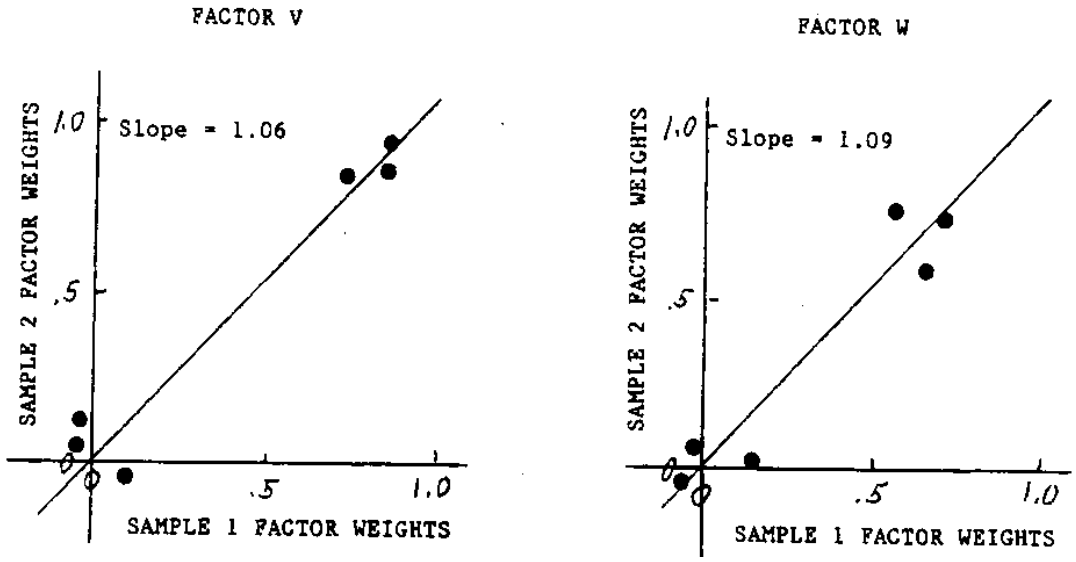

Figure 5.6: Comparison of Factor Weights for Two Samples Differing in Heterogenity as to School Grade of Individuals

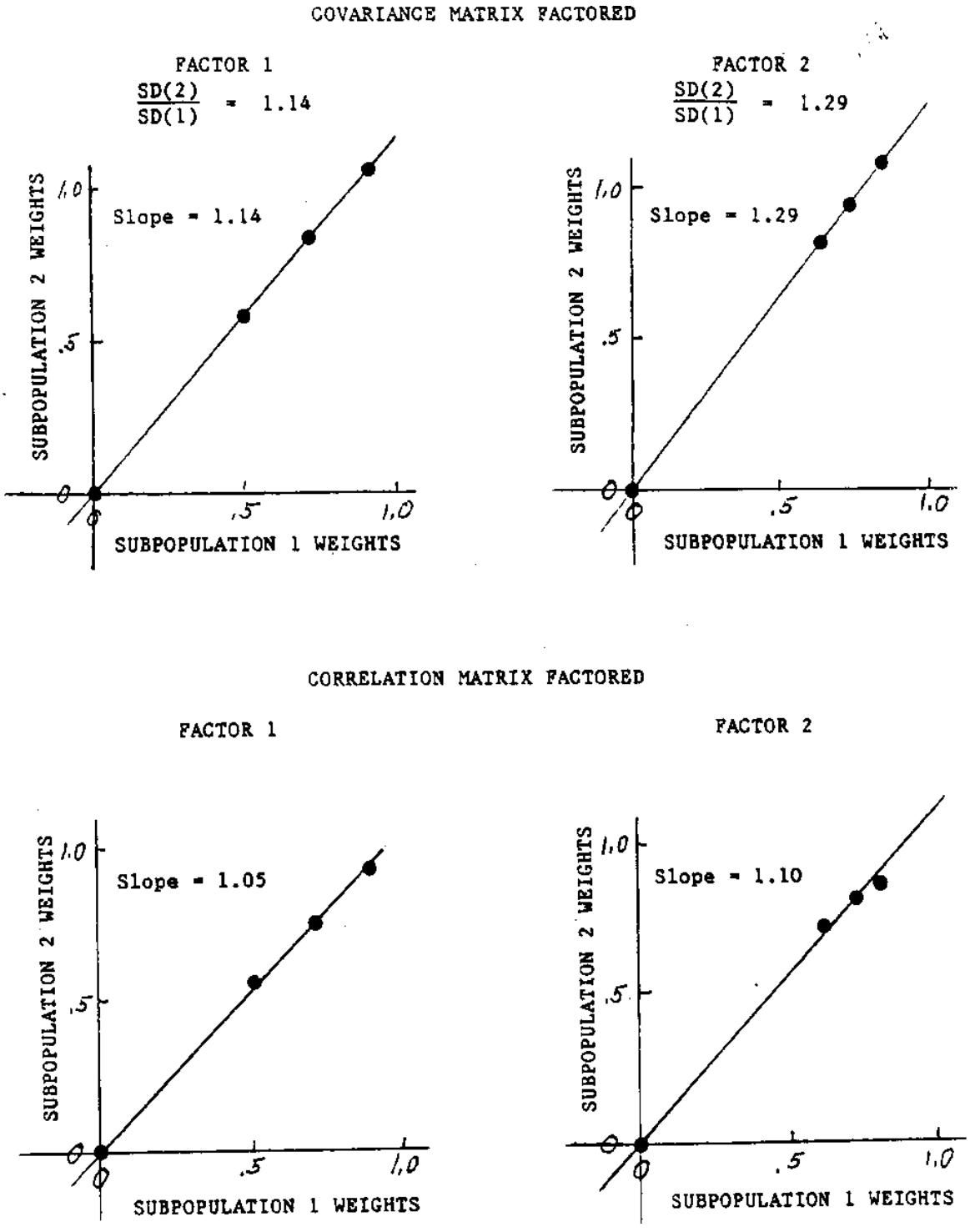

Figure 5.7: Comparison of Factor Weights for Two Subpopulaions in the Selective Sampling Demonstration. SD(1) is the Standard Deviation of Factor Scores in Subpopulation 1; SD(2) is the Standard Deviation of Factor Scores in Subpopulation 2.

For the covariances matrices graphs the ratios of the common factor scores are indicated. Note that the slopes of the lines of relations for these graphs equal the factor score ratios of standard deviations. Note, also, that the points fall precisely on the lines of relation. These results are consistent with and predicted by Eq. (5.61): the covariance matrices factor weights in two rations are strictly proportional by columns.

For the correlation matrices graphs in Figure 5.7, the slopes of the lines of best fit are greater than unity but less than the slopes for the covariance matrices graphs. A point of interest is that these lines of best fit for the correlation matrices have slopes quite close to the slopes of the lines for the Thurstone studies shown in Figure 5.6. Standardizing the attribute scores in subpopulations does not make the factor weights equal between the two subpopulations. There is only a partial adjustment toward equality from the relation for factor weights for the covariance matrices. Further, note that the points in the lower graphs in Figure 5.7 vary a small amount from the lines of best fit. The source of this variation can be understood by considering Eq. (4.69). The product of matrices $[\boldsymbol{\Sigma}_d]_{zz\mathbf{q}}^{-1}$ and $[\boldsymbol{\Sigma}_d]_{zzp}$ is generally not an identity matrix, nor even a constant. Thus, in matching factors from different subpopulations, an assumption of equality for factor weights from correlation matrices is not justified. Nor is it justifiable to assume that a proportionality exists between such weights and that the proportionality represents the ratio of standard deviations on a selection variable.

To conclude this discussion of the effects of selective sampling, let us consider the case in which selection variables are related to specific factors. All previous discussion has assumed that these relations were zero; a variety of complexities occurs when the specific factors have nonzero relations with the selection variables. In these cases the weight matrix $\boldsymbol{W}_\xi$ is not zero, which results in the specific factors not having zero covariances with the common factors. Further, the covariance matrix among the specific factors no longer is an identity matrix. These results are shown in Eqs. (5.45) and (5.47). A number of situations exist which lead to some simplifications; however, none of the special cases will be treated here.

Individuals using factor analysis should recognize that a fairly common type of situation occurs which can result in selection variables being related to specific factors. Frequently, data are available on tests used in selection. These tests may be admissions examinations such as the SAT or the ACT. When results of such tests are incorporated into the battery of attributes being analyzed, there are relations of the selection variables with the specific factors of these tests. In fact, there are relations of the error of measurement factors with the selection variables. Use of parallel tests to the selection tests eliminates the relations of the error of measurement factors to the selection variables, but does not eliminate the relations of the specific factors with the selection variables. If there is only one selection test in the battery, the loadings of this test will

be altered. If there are two or more selection tests in the battery, not only will the loadings of these tests be altered, but also one or more common factors will be added with some negative loadings. Experimenters should be aware of these possibilities.

5.3. Effects of Selection of Attributes

In the preceding sections we have examined how obtained factor solutions can be influenced by various aspects of the selection of a sample of individuals from the population. It is also important to recognize that obtained factor solutions can be influenced by another selection process: the selection of attributes from the domain of interest. A consideration of the nature of common factors themselves will provide a basis for understanding the critical role that selection of attributes plays in affecting the obtained factor solution. A common factor is defined as a factor which is common to more than one attribute in the battery. Unique factors, on the other hand, represent those influences which operate on only a single attribute. (Let us keep in mind that unique factors are composed of both specific and error portions, which correspond to systematic and random influences on single variables, respectively.) Given these definitions, it should be clear that the obtained factor solution is determined by the attributes selected for inclusion in the battery to be factor analyzed. In particular, the obtained common factors will be determined to a great extent by the selection of the attributes.

One way to gain a more explicit understanding of this phenomenon is to consider the effects of altering a given battery of attributes by deleting or adding attributes. Suppose that we have a battery of attributes and have obtained a factor solution that is characterized by a given number of common factors. Let us consider the impact of deleting attributes from the battery. If we were to delete all attributes representing a given common factor and conduct the factor analysis on the reduced battery, the given common factor would disappear completely from the obtained solution. An interesting variation of this phenomenon occurs if we consider the result of deleting all but one of the attributes representing a given factor. In that case, the factor would not be obtained as a common factor, but its influence would still be present in the form of a specific factor affecting the one remaining attribute. In such a case, the remaining attribute would have a much higher unique variance in the reduced battery than it did in the original battery.

It is interesting to consider the converse of this process; i.e., adding attributes to a given battery. Clearly, such a process can give rise to one or more additional common factors in the expanded battery, though this would not have to occur. To simplify the discussion, suppose two new attributes are added to an existing battery. These two new attributes may have in common a factor which was not present at all in the original battery. Thus, one additional common factor would be obtained in the expanded battery. Alternatively, the two new attributes may be influenced by the same common factors already represented in the battery, in which case no

additional common factors would be obtained. Another possible case is that one or both of the new attributes may be influenced by a factor which was a specific factor in the original battery; thus, the expanded battery would yield an additional common factor which involves both new and original attributes.

While a variety of other cases could be pointed out, the critical issue is to recognize that the modification of an existing battery of attributes by adding or deleting attributes can have a substantial impact on obtained factor solutions. Let us consider a simple demonstration of this phenomenon. This demonstration will involve examining the effects of modifying the battery of attributes employed in the illustration presented in Section 1.3. Recall that that battery contained nine attributes, with three attributes representing each of three common factors. The common factors, with corresponding attributes listed in parentheses, were numerical calculation; (addition, multiplication, three-higher), spatial relations (figures, cards, flags), and perceptual speed (identical numbers, faces, mirror reading). The three-factor solution for this battery was presented in Table 1.2. Suppose this battery were modified by deleting the last two tests; i.e., the faces test and the mirror reading test. Based on the discussion above, we can predict the impact of this modification on the obtained factor solution. The perceptual speed factor would no longer be present as a common factor, since it would influence only one remaining attribute. Thus, we would expect to obtain only two common factors. In addition, we would expect that the remaining test representing the perceptual speed factor (i.e., identical numbers) would have a much higher unique variance in the solution obtained from the reduced battery.

A factor solution was obtained for the reduced battery (by applying factor analysis methodology to the 7×7 correlation matrix obtained by deleting the last two attributes from the matrix shown in Table 1.1). A two-factor solution was found to fit these data very well, and the weights and intercorrelations for the two factors, along with the communalities of the attributes, are shown in Table 5.10. The characteristics of this solution correspond to the predicted results. The two obtained factors correspond to numerical calculations and spatial relations, and no perceptual speed factor is obtained. The correlation between the two factors (.16) is very similar to the correlation between those two factors in the three-factor solution (.15). The communality of the identical numbers test is substantially reduced in the two-factor solution (.28) from what it was in the three-factor solution (.44), reflecting the fact that some variance which was accounted for by the three-factor solution has now become unique variance in that attribute. The only reason that the communality has not dropped to nearly zero is that the identical numbers test does have a moderate loading on the numerical calculations test. Other demonstrations using the original nine-test battery could be carried out in order to examine other types of attribute selection on obtained factor solutions. We leave this as an exercise for the reader.

Another issue inherent in the process of attribute selection involves the number of

Table 5.10

Factor Solution for Seven Mental Tests

	Factor Weights		Communalities
Attribute	1	2	
1. Addition	.60	.05	.37
2. Multiplication	.75	-.06	.55
3. Three-Higher	.52	.32	.42
4. Figures	-.03	.71	.50
5. Cards	.00	.79	.63
6. Flags	.03	.71	.51
7. Ident. Numbers	.53	.02	.28

Factor Intercorrelations

Factor	1	2
1	1.00	.16
2	.16	1.00

attributes to be included in the battery which is to be analyzed. For scientifically convincing results, an obtained common factor structure should be overdetermined above a minimum necessary to satisfy mathematical conditions. By overdetermination, we mean that the structure represented in the obtained solution should be uniquely determined and highly constrained, rather than arbitrary. This overdetermination will be shown to depend on a relationship between the number of common factors in the domain being investigated with the number of attributes in the observed battery. A coefficient of overdetermination will be introduced, and an inequality will be developed between the number of attributes to be included in a battery and the number of common factors in the domain.

This development follows and extends Thurstone's (1935, 1947) work on the number of independent common factors. Inequalities developed by Thurstone have been reproduced in texts on factor analysis (e.g., Mulaik, 1972). Further investigation of this problem was reported by Ledermann (1937).

Let *r* represent the number of common factors expected from the analysis, and let *n* represent the number of attributes in the battery. It is useful to consider the number of correlations among the attributes; this would be the number on one side of the diagonal of the *n* $\times$ *n* correlation matrix, which would be *n*(*n*-1)/2. This value represents the number of intercorrelations, which are the values employed to obtain a factor solution. It is also useful to consider the number of independent factor weights being estimated in a factor analysis. The number of weights in the factor weight matrix would be *nr*. However, an important fact here is that it is always possible to transform a given factor weight matrix so that a number of entries in the weight matrix will be zero. It will be shown in Chapter 9 that this number is *r*(*r*-1)/2. Thus, the number of independent factor weights in an obtained solution can be defined as the total number of weights, minus the number that can be set to zero by transformation of the factors. This would yield *nr*-*r*(*r*-1)/2. The basic inequality is that the number of correlations (i.e., the values being used to obtain parameter estimates) should be equal to or greater than the number of independent factor weights (i.e., the parameters being estimated). This would be written as follows:

$$n(n-1)/2 \geq nr - r(r-1)/2 \tag{5.70}$$

In order for the results to be overdetermined, the number of correlations should be somewhat greater than the number of independent factor weights. A coefficient of overdetermination will be designated as *g* and may be defined such that the number of correlations is at least *g* times the number of independent factor weights. This leads to a revision of Eq. (5.70) to yield

$$n(n-1)/2 \geq \mathbf{g}[nr - r(r-1)/2] \tag{5.71}$$

Solution of this inequality for *n* is facilitated by defining n_c which satisfies the equality of Eq. (5.71) and letting

$$n \geq n_c \tag{5.72}$$

The definition of n_c allows us to write

$$n_c(n_c\text{-}1)/2 = \mathbf{g}[n_c r\text{-}r(r\text{-}1)/2] \tag{5.73}$$

Eq. (5.73) can be manipulated to obtain

$$n_c^2 - (1 + 2\mathbf{g}r)n_c + \mathbf{g}r(r-1) = 0 \tag{5.74}$$

Eq. (5.74) may be solved for n_c via the application of the quadratic formula. This yields the following:

$$n_c = (\mathbf{g}r + 1/2) + \underline{\nu}(\mathbf{g}r + 1/2)^2 - \mathbf{g}\boldsymbol{r}(r-1) \tag{5.75}$$

Eq. (5.75) coupled with Eq. (5.72) provides the desired relation of battery size *n* to the estimated number of factors *r*. The minimum value of *n* would be the least integer equal to or greater than n_c. These formulae allow one to calculate the minimum value of *n* for any given number of factors *r* and coefficient of overdetermination *g*. Values of minimum *n* for selected values of *g* in the range of 1 to 4 by steps of 0.5 and the numbers of factors from 1 to 20 are shown in Table 5.11. The column for coefficient of overdetermination 1.0 corresponds to values presented by Thurstone (1947). Thurstone's (1947) suggestion that "In order for a factor analysis to be stable and scientifically significant and convincing, the number of tests must be two or three times greater than those which are shown in the table" corresponds to use of a coefficient of overdetermination in the range of 1.5 to 2.0. In early studies in a domain the number of common factors may be subject to crude estimation and a larger coefficient of overdetermination should be used. In later studies in a domain the number of common factors may be better known and a smaller coefficient of overdetermination would be justified. In any case, using a number of attributes that is too small will result in poorly defined factors and unstable results.

Table 5.11
Minimum Number of Attributes for Given Number of Factors

	Coefficient of Overdermination						
Number of Factors	1.0	1.5	2.0	2.5	3.0	3.5	4.0
1	3	4	5	6	7	8	9
2	5	7	9	11	13	15	17
3	6	9	12	15	18	21	24
4	8	12	16	20	24	28	32
5	9	14	19	24	29	34	39
6	10	17	23	29	35	41	47
7	12	19	26	33	40	47	54
8	13	21	30	38	46	54	62
9	14	24	33	42	51	60	69
10	15	26	36	47	57	67	77
11	17	29	40	51	62	73	84
12	18	31	43	55	68	80	92
13	19	33	47	60	73	86	99
14	20	36	50	64	78	93	107
15	21	38	54	69	84	99	114
16	23	40	57	73	89	105	122
17	24	43	60	78	95	112	129
18	25	45	64	82	100	118	136
19	26	48	67	87	106	125	144
20	27	50	71	91	111	131	151