A Comparison of Principal Component Analysis, Common Factor Analysis, and Image Analysis: A Monte Carlo Study

by

Todd Stephen Woodward
B.Sc., University of Victoria, 1989

A Thesis Submitted in Partial Fulfilment of the
Requirements for the Degree of

ACCEPTED

[redacted] S

MASTER OF ARTS

in the Department of Psychology

We accept this thesis as conforming
to the required standard

[redacted]

Dr. Michael A. Hunter, Supervisor (Department of Psychology)

[redacted]

Dr. Lorne K. Rosenblood, Departmental Member (Department of Psychology)

[redacted]

Dr. John O. Anderson, Outside Member (Department of Psychological Foundations in Education)

[redacted]

Dr. Bill McCarthy, External Examiner (Department of Sociology)

University of Victoria

Supervisor: Dr. Michael A. Hunter

Abstract

Principal component analysis, common factor analysis, and image analysis are compared on their ability to reproduce loading patterns underlying simulated data sets. Sample size, number of variables per factor, saturation, and percentage of measurement error were systematically varied. Common factor analysis was found to be more accurate at retrieving the population loading pattern when there was no measurement error, but principal component analysis gave the clearest representation of the population loading pattern when there was 25% measurement error or greater. Image analysis severely underestimated the loadings, except when there was no measurement error. The technique differences diminished as the number of variables per factor increased, as population loading saturation increased, and as the amount of measurement error decreased. Numerical explanations for these differences are presented, and recommendations for future research are proposed.

Examiners:

[redacted]

Dr. Michael A. Hunter, Supervisor (Department of Psychology)

[redacted]

Dr. Lorne K. Rosenblood, Departmental Member (Department of Psychology)

[redacted]

Dr. John O. Anderson, Outside Member (Department of Psychological Foundations in Education)

[redacted]

Dr. Bill McCarthy, External Examiner (Department of Sociology)

Table of Contents

List of Tables

List of Figures

Acknowledgements

The completion of this thesis has taken place under exceptionally difficult circumstances, and it certainly could never have been carried out without the aid and support of many. First, a very big thank-you to Dr. Michael Hunter, who's guidance, patience, and suggestions were invaluable and very much appreciated. Thanks also to my committee members, Dr. Lorne Rosenblood, Dr. John Anderson, and Dr. Bill McCarthy for their helpful suggestions and revisions. A very special thank-you to Dr. Richard Harshman at U.W.O., my advisor for the first year of my M.A., for introducing me to MATLAB and data simulation, and especially for encouraging me to think as creatively and clearly as I am able.

I cannot express the amount of gratitude I feel for my mother Gail, my late father Steve, and my sister Lynne. The strength of these people to stand by me under any circumstances has taken me to this point in life, and I owe every word in this paper to them. I would also like to thank a recent very close companion, Karin Christensen, whose support has been so important over the last half a year that I don't think she or anyone else will ever understand how grateful I am for her friendship.

Last but not least, I would like to thank Donald Hine, Jennifer Maggs, Shawn O'Conner, Dan Slick, and Brent Small for their comments and suggestions, and Chris Bowles for help in the graphics and wardrobe department.

Dedication

I would like to dedicate this thesis to my father, Stephen William Woodward, who passed away as the thesis was nearing completion. Without his confidence, love and support, I'm sure I would never have even gone to university. I wish he could be here to see the project completed, but I'm not positive that he's unaware of my circumstances. As he would have wished, I did not give up in the face of grief and sorrow, and now I realize the strength and optimism that he has instilled within me.

INTRODUCTION

In the behavioral sciences, various techniques have arisen to reduce the dimensionality of a data set from p (the number of observed variables) to m (the number of "summary" variables which can be used to describe the data set, where $m << p$). These techniques can generally be classified into two categories: principal component analysis (PCA) and common factor analysis (CFA). The contention surrounding the comparative virtues of these techniques began almost as soon as they co-existed in the early 1930's[1]. This controversy continues to spark lively debate into this decade, with an entire 1990 issue of the journal *Multivariate Behavioral Research* being devoted to the topic (e.g. Bentler & Kano, 1990; Bookstein, 1990; Gorsuch, 1990; Mulaik, 1990; Rozeboom, 1990; Velicer & Jackson, 1990a; Velicer & Jackson, 1990b). The purpose of this thesis was to present a critical analysis of the empirical literature comparing PCA and CFA, and to introduce an improved methodology for evaluating the two procedures. To begin, an overview of the statistical and conceptual models underlying each technique is presented.

Principal Component Analysis

The purpose of PCA is to represent a set of p interrelated observed variables by a much smaller set of m orthogonal underlying variables called components. The components are linear combinations of the observed variables, with the weights chosen such that the maximum amount of observed variance is consolidated into the fewest components.

---

[1]Thurstone, in his book *The Vectors of Mind* (1935, p. 130), criticized Hotelling's (1933) new method of principal components, primarily for using 1's in the main diagonal of the correlation matrix instead of communalities.

For example, consider a researcher who wishes to explain as much of the variance in a data set as possible with one "summary" variable, or component (i.e., the researcher wishes to reduce the dimensionality of his data set from p to 1). When weighting the manifest variables to construct this summary component, the variables that share variance are given large weights. Creating a component that consists largely of shared variance allows this component to incorporate the maximum possible amount of total variance.

The greater the number of manifest variables that can be given large weights, the more total variance this summary variable will account for. The researcher would like this summary variable, described as the first principal component, to have nearly as much variance distributed to it as exists in the entire data set[2].

Most often, however, more than one component is required to adequately summarize a data set. Therefore, a second set of weights may be obtained to form a second principal component. A restriction placed on this second set of weights is that they must result in a second component that is orthogonal to (uncorrelated with) the first principal component. Similarly, should a third set of weights be derived, they must result in a component that is orthogonal to the second and the first component. The maximum number of sets of weights that can be obtained (and therefore the maximum number of components that can be created) is equal to p, the full dimensionality of a data set.

Because less than p variables are desired to explain the data set, m components are decided upon as sufficient to summarize the relationships among the variables. The

---

[2]The total variance in the data, in this case, is equal to the sum of the variances of all manifest variables.

value of m is decided upon according to some predetermined criterion, and the remaining $p - m$ components are considered to constitute error variance.

A procedure for finding weights that satisfy these criteria was labelled principal component analysis and formulated by Hotelling (1933)[3], and further developed by Girshick (1936). They demonstrated that the variance of a component and the weights to combine the manifest variables to form the component can be found by solving the following eigenproblem:

(1) $$({}_{p}\mathbf{R}_{p} - \lambda_{p}\mathbf{I}_{p})_{p}\mathbf{w}_{1} = \mathbf{0}$$

where $\mathbf{R}$ is a correlation matrix of the manifest variables, λ is the variance of the component (the eigenvalue), $\mathbf{I}$ is the identity matrix, and $\mathbf{w}$ is the weight vector (the eigenvector). See Appendix A for further explanation of how to solve this equation for p components.

This redistribution of variance can also be carried out using the matrix decomposition technique of singular value decomposition (SVD) (see Appendix B). Both SVD and eigen-decomposition can provide the weights needed to create p orthogonal components that account for descending, orthogonal amounts of variance. The following formula shows how the weights can be applied to the variables to compute components:

(2) $$\mathbf{C}_{ik} = \mathbf{w}_{k1}\mathbf{x}_{i1} + \mathbf{w}_{k2}\mathbf{x}_{i2} + \ldots + \mathbf{w}_{kp}\mathbf{x}_{ip}$$

where $\mathbf{C}_{ik}$ is the ith subject's score on the kth component, $\mathbf{w}_{kp}$ is the variable p weight from the kth eigenvector, and $\mathbf{x}_{ip}$ is the ith subject's score on variable p from the data

---

[3]Pearson (1901) conceived of the idea, but did not develop an explicit method for calculating components.

matrix.

Once *m* components are decided upon, they can be multiplied by their weights to observe how closely the reduced rank solution reproduces the original data set:

(3) $$ {}_{n}\mathbf{x}_{redp} = {}_{n}\mathbf{C}_{m}\mathbf{W}_{p}'. $$

where $\mathbf{x}_{red}$ is the reduced rank reproduced data matrix, **C** is the matrix of components, and **W** is the matrix of eigenvectors. The nearer the reduced rank solution approximates the original data matrix, the more total variance the components have accounted for, and the more efficiently the *m* principal components have summarized the data set.

The preceding discussion of weighted manifest variables is useful for understanding the inner workings and goals of PCA. However, although the eigenvectors determine the structure of the factors, they are not usually used to interpret the relationship between the variables and the components. A *component loading matrix* (**A**) is usually interpreted to identify which variables determine the components. The elements of the loading matrix are equal to the correlation between the variable and the component in question.

The loading matrix can also be used to assess how closely the reduced rank solution reproduces the original data set:

(4) $$ {}_{n}\mathbf{x}_{redp} = {}_{n}\mathbf{Z}_{m}\mathbf{A}_{p}'. $$

where $\mathbf{x}_{red}$ is the reduced rank reproduced data matrix, **Z** is the matrix of component scores (see Appendix B), and **A** is the loading matrix. Again, the nearer the reduced rank solution approximates the original data matrix, the more total variance the components have accounted for, and the more efficiently the *m* principal components have

summarized the data set.

The reproduced (reduced dimensionality) correlation matrix can also be calculated by using the loading matrix:

(5) $$ {}_{p}\mathbf{R}_{red\,p} = {}_{p}\mathbf{A}_{m}\mathbf{A}_{p}' $$

where $\mathbf{R}_{red}$ is the reproduced correlation matrix, and $\mathbf{A}$ is the component loading matrix. $\mathbf{R}_{red}$ can be compared to the full correlation matrix to evaluate how closely the solution has approximated the data. This is an identical fit index to that presented in equations (3) and (4).

As the reader may have noticed in equation (1), the correlation matrix must be used for eigen-decomposition; furthermore, it is usually used for SVD (see Appendix B for an explanation of how the eigenvectors can be converted to component loadings). Previous researchers have used the loading matrix to simulate both correlation matrices and data sets, so the equivalence of the two cases should be recognized.

Common Factor Analysis

As is true for PCA, the purpose of CFA is to reduce the dimensionality of a data set and expose its underlying structure (common factors are akin to the components of PCA, and the terms will be used interchangeably when discussing dimension reduction). Factors, however, are thought to be latent causes of the variation in the observed variables, not merely summary variables as components are. The primary difference between PCA and CFA is that CFA specifies a model that distinguishes between the *common* and *unique variance* of the variables, and attempts to explain only the common variance. The common variance of a variable consists of variance that is shared with the

other variables, whereas the unique variance of a variable is comprised of variance not shared with the other variables in the data set. PCA, however, simply redistributes all variance without postulating a model or making this distinction.

In CFA, the variables of a data set are expressed as linear functions of *m* hypothetical variables, or common factors. The common factor model is:

(6) $$_{n}\mathbf{x}_{p} = {}_{n}\mathbf{f}_{m}\mathbf{A}_{p}' + {}_{n}\mathbf{e}_{p}$$

where **x** is a data matrix of standardized variables, **A** is the factor loading matrix (the correlation between the variables and the factors for an orthogonal solution), **f** is the factor score matrix (the standardized factors), and **e** is the matrix of uniqueness (variation in the data not related to the factors). The columns of **e** are assumed to be uncorrelated to themselves, and to the columns of **f**.

The similarity of this model to that of regression is striking[4]. However, there is a very important difference, which is that neither **A** nor **f** are known in CFA. In regression, **f** would be known (the independent variables **X**), and only **A** would be estimated (the beta weights **b**). Since there are two unknowns and only one known element of the factor model, there will be indeterminacy in the solutions (the "best fitting" solution will not be unique [Jolliffe, 1986, p. 117]).

The common factor model conceives of variables as linear combinations of factors, whereas the model implicit in PCA defines components as linear combinations

---

[4]The matrix equation for multiple regression is:

(7) $$_{n}\mathbf{y}_{1} = {}_{n}\mathbf{X}_{p}\mathbf{b}_{1} + {}_{n}\mathbf{e}_{1}$$

where **y** is the dependent variable, **X** is the matrix of independent variables, **b** is the constant and beta weights, and **e** is the vector of prediction error.

of variables. In common factor analysis, it is the *manifest variables* that are envisioned as being composed of weighted *latent variables* (the factors). Since the weights are applied to the factors, they are applied to the common parts of the variables.

In factor analysis, as in PCA, the estimation of the model initially focuses on **A**, with **f** being estimated later. Therefore, we can transform the above model (equation [6]) to be based on the correlation matrix:

(8) $$_{p}\mathbf{R}_{p} = {}_{p}\mathbf{A}_{m}\mathbf{A}_{p}' + {}_{p}\mathbf{D}_{p}$$

where **R** is the correlation matrix, **A** is the loading matrix, and **D** is a diagonal matrix of the covariance of the unique parts of the variables. Either this model or the one shown in equation (6) is frequently called the fundamental equation for CFA. The equivalence of these two representations of the fundamental equation is important, as both have been used by previous researchers when creating data sets for Monte Carlo studies.

Notice that factor analysts focus on the off-diagonals of the correlation matrix, ignoring the diagonals or adjusting them to remove the correlation due to the unique parts of the variables. This can be contrasted with principal component analysts, who redistribute all variance in the data set--the variance being represented on the diagonal of the correlation/covariance matrix.

A major problem for factor analysts is that it is impossible to know the true values that determine the **D** matrix in the fundamental equation (8). These uniqueness values are calculated as $1\text{-}h^2$, where h^2 is the *communality*. The communality is defined as the proportion of a variable's variance that is attributable to the common factors in the

population[5]. Since the common factors are never known before the analysis, the communality of a variable cannot be known, and must be estimated. This and other steps used in CFA procedures will be discussed below.

Factor analysts look upon the factors as latent variables measurable only through the inter-correlations of the observed variables. Component analysts, however, conceive of the components simply as summary descriptions of the manifest variables themselves, disregarding the idea of latent causal factors. The CFA model can therefore be described as a reflective model (variables are reflections of underlying factors), and PCA as formative (components are formed as linear combinations of variables). This distinction is illustrated in Figure (1).

Rotation

Rotation of either the components or the factors is usually performed to simplify interpretation of the loadings. This practice is almost always based on the belief that only one or a few factors are involved in most variables. This belief prompts a search for *simple structure*, a concept defined by having most of the loadings for a variable very small, and its variance concentrated on only a few (preferably one) factors (Thurstone, 1947, p. 335).

Varimax rotation is the most commonly used technique for finding simple structure (Kaiser, 1958). This method finds the position of the factor axes where the

---

[5]Technically, this value should be referred to as the "true" communality. Communality can also define an estimate of the true communality that is placed on the diagonal of the correlation matrix (as in principal factor analysis), or the sum of squared factor loadings of each variable on the calculated factors (the latter "communalities" should be referred to as the reproduced or output communalities).

variance of the squared loadings across a factor is maximized. The resulting axis position is called the *varimax* solution.

Theoretical Debate

Under ideal conditions, the solutions from PCA and CFA are very similar (e.g. Gorsuch, 1983, p. 108; Snook and Gorsuch, 1989; Velicer, Peacock & Jackson, 1982; Velicer & Fava, 1987). Recommendations have been made for the use of both techniques with equally strong conviction. However, many of these recommendations have been based largely on theoretical grounds, and have often involved discussion and opinions rather than empirical evidence (Bentler & Kano, 1990; Bookstein, 1990; Comrey, 1973; Gorsuch, 1990; Mulaik, 1990; Rozeboom, 1990; Velicer & Jackson, 1990a; Velicer & Jackson, 1990b).

For example, much of the 60-year-old controversy over the relative benefits of the two techniques has revolved around the debatable method of decomposing a correlation matrix that has ones on the diagonal (Bookstein, 1990; Gorsuch, 1983, p. 20, 1990; Mulaik, 1972, p. 181; Nunnally, 1978, p. 399, 418; Rozeboom, 1990; Thurstone, 1947, p. 484, 1935, p. 130; Velicer & Jackson, 1990a). Factor analysts often attempt to adjust the correlation matrix for unique variance before SVD by subtracting out the diagonal matrix (**D** from equation [8]) of uniquenesses.

As the first critic of calculating components of a full correlation matrix, Thurstone (1935, p. 130) wrote:

> To record unity in the diagonal cells of R_0 implies that the total variance of each trait is to be described by *common* factors...Any solution in which the intercorrelations of n tests are accounted for exactly by n common factors must be an artifact as far as the psychological problem is concerned, because it is definitely known that each test has

some *unique* variance.

Similarly, in 1990 Gorsuch wrote:

> In selecting a model, ask if your variables are truly measured without error; if so, component analysis is a possibility. (p. 34).

Although most researchers acknowledge this argument, it is also generally agreed that if there are a large number of variables, the influence of the diagonals on the final solution becomes negligible (Nunnally, 1978, p. 419; Gorsuch, 1983, p.123). Conversely, when there are a small number of variables, PCA will have higher loadings than CFA (Nunnally, 1978, p. 399; Gorsuch, 1983, p. 124; Velicer & Jackson, 1990a), because it explains the unique variance left on the diagonals. The issue is, however, whether PCA's higher loadings are *inflated*, as Gorsuch (1983, p. 124) claims, or CFA's lower loadings *deflated*, as Velicer and Jackson (1990a) suggest. This matter, among others, has been clarified with the aid of empirical investigations using data sets created with known factor structures.

Empirical Investigations

Empirical investigation can help to uncover circumstances under which dimension reducing methods produce divergent results. The methods are compared on their ability to recover the factor patterns that underlie simulated data sets (e.g. Borgatta, Kercher, & Stull, 1986; McArdle, 1990; Snook & Gorsuch, 1989; Velicer & Fava, 1987; Velicer, Peacock, & Jackson, 1982; Widaman, 1990). If there are differences in the methods' abilities to recover the factor pattern in some situations, this should be considered when recommendations are being made for their use.

Snook and Gorsuch (1989), Borgatta et al. (1986), McArdle (1990) and Widaman

(1990) have carried out such simulations, and have concluded that CFA is considerably more accurate than PCA at retrieving the population loading pattern. However, Velicer and Fava (1987) and Velicer, Peacock, and Jackson (1982) concluded that PCA, image component analysis (considered by the authors to be a type of component analysis), and maximum likelihood factor analysis (a CFA approach) produce equivalent results. Before explaining these contradictory findings, the procedural differences among the studies will be discussed.

There are four general areas where the above studies differed: (a) data creation methodology (b) population loading pattern construction (c) comparison of sample loadings to population loadings, and (d) types of analyses performed. The following review will attempt to magnify these differences, and in some cases suggest improvement. A decision can then be made as to the most appropriate methods to employ when comparing dimension reducing techniques on the analysis of simulated data.

Data Creation

There are two steps previous researchers have taken when simulating matrices with known factor structures. The first is to create a preliminary matrix with variance due only to the factors, called the factor contribution matrix (FCM). The second step is to add a subsequent matrix representing variance not due to the underlying factors to this preliminary matrix. The types of matrices that can be simulated are: (a) a p x p correlation/covariance matrix, or (b) an n x p data matrix (where p is the number of variables and n is the number of subjects). The simulated matrix is referred to as the population matrix, and the pattern matrix used to generate it is labelled the population

pattern matrix. Sampling error may be added in a later step to transform the population matrix to a sample matrix.

Factor Contribution Matrix

The FCM is created either by multiplying the population pattern matrix (**A**) by its transpose to create a factor contribution correlation/covariance matrix (as Borgatta et al. [1986], Velicer et al. [1982], Velicer and Fava [1987], Widaman [1990], and McArdle [1990] did), or by pre-multiplying the transpose of the population pattern matrix by a matrix of factor scores (**f**) to create a factor contribution data matrix (as Snook and Gorsuch [1989] did). Both resulting matrices will be of insufficient rank (if $m < p$), with only m patterns of variation. These methods are actually equivalent, as the covariance of the factor contribution data matrix is equal to the factor contribution covariance matrix. The FCM exists only in the population.

Uniqueness Addition

In the common factor model, *unique factors* or *unique variance* normally refer to the part of the variance of a variable that is not explained by the common factors. A variable's *uniqueness* is defined as that proportion of the variance not attributable to the common factors (Gorsuch, 1983, pp. 26-29). In the present discussion, reference will be made to the above concepts as they exist in the sample as *sample* unique variance and *sample* uniquenesses. The terms can also refer to population variances that exist before measurement, and discussion involving this type of variance will use the terms *population* unique variance and *population* uniqueness.

The addition of population uniqueness to the FCM will result in an increase in the

complexity of the variation patterns. This variation will no longer be due only to the underlying factor structure, thus increasing the rank of the matrix to p. This step can be performed in two ways, depending on the type of FCM that was created.

Borgatta et al. (1986), Velicer et al. (1982), Velicer and Fava (1987), and Widaman (1990) added a p x p diagonal matrix of population uniquenesses to the factor contribution correlation/covariance matrix. The values of this matrix were calculated to result in unities on the diagonal of the simulated population correlation matrix. McArdle (1990) probably followed this procedure of population uniqueness addition, but does not indicate how or whether these steps were performed. Proponents of this method of matrix simulation are basing their methods on the fundamental equation for factor analysis presented in equation (8).

Snook and Gorsuch (1989), taking an alternative approach, added columns of random numbers to the factor contribution data matrix. The variance of the columns of this population unique variance matrix were constrained to result in a data set of standardized variables (variance of 1.0 and mean of 0.0). Standardized variables are used in dimension reduction to allow all variables to have equal influence on the solution. This method of data simulation is based on the fundamental equation for factor analysis presented in equation (6).

As was the case for FCM creation, the methods of variance/covariance addition are equivalent, as the covariance of **e** is equivalent to **D** (see equations [6] & [8]). This was not a precise relationship in practice, however, because the matrix of uniqueness vectors that Snook and Gorsuch (1989) added was not constrained to have orthogonal

columns, causing random changes to the off-diagonal elements of the simulated correlation matrix. These off-diagonal changes were presumably small, as there were a large number of normally distributed random values per columns (150 subjects were simulated), minimizing accidental correlations.

Since equations (6) and (8) represent compensatory models, both methods of variance addition employed by previous researchers led to a situation where the percentage of population unique variance added was directly related to the size of the population loadings. This is sound methodology for population creation, but the sample uniqueness should be greater than the population uniqueness due to measurement error, and should not depend on the size of the population loadings. The sample uniqueness simulated by previous researchers were consistently equal to the population uniqueness (disregarding sampling error). This point can be clarified through a discussion of the theoretical splitting of the output (sample) unique variance proposed by factor analysis researchers.

Specific and Error Variance

Factor analysts acknowledge that the sample unique variance can be divided into two parts: the *specific variance* and the *error variance* (Comrey, 1973, p. 22; Gorsuch, 1983, p. 109; McDonald, 1985, p. 109; Mulaik, 1972, p. 97; Nunnally, 1978, p. 347). Specific variance is thought to represent true, reliable variance in the variable that is not related to the other variables in the data set. Error variance, however, consists of what we will term measurement error (error due to unreliable measuring instruments, or any random event causing variance in the scores). Because there is normally no knowledge

of the proportion of common variance in the population, or the amount of measurement error in a variable, all specific and error variance is modelled as (sample) unique variance.

By definition, the size of the specific variance depends completely on the size of the population loadings. Therefore, specific variance is simply another term for population unique variance. Furthermore, the variance that was added by previous researchers to the FCM (labelled uniqueness) could be accurately labelled specific variance.

As mentioned above, error variance depends mainly on the quality of the measuring instruments, and does not depend in any way on the size of the population loadings. This source of variation was not included in what previous researchers have termed unique variance.

The size of specific variance, then, should depend completely on the size of the *population* loadings, as it has in previous simulations. Because the amount of error added by the above researchers depended completely on the size of the loadings in the population, they were simulating samples with no error variance, and were therefore neglecting to include the variation due to measurement error. The amount of measurement error added must be independent of the size of the population loadings. Once this is done, the performances of dimension reducing techniques may display increasing divergence, as they may handle measurement error differently.

Example

To further clarify the meaning of the terms specific variance, error variance,

common variance, measurement error, latent factor, and "true" score, an example is given.

For the following discussion, a population will be considered to consist of subjects that, because they have not been chosen to be a member of a sample for research purposes, have not yet been measured. They do, however, possess some amount of the attribute that any given test measures. This can be known as their "true" score on that attribute. Once a sample is chosen from this population, subjects must then be assigned some amount of measurement error. Thus, when the variance (or correlation) of attributes in the population is mentioned, these attributes will be considered to be free of measurement error, consisting only of "true" scores.

When one is simulating a population of scores, it is possible for the researcher to determine the "true" score of each subject on a given attribute. The value of this true score will be produced partially by the underlying factors that are to be exposed by the inter-correlations among the chosen battery of tests, and partially by an unknown subset of all the other causal factors in the universe[6].

The following is an example of the sources of variation that may determine a subject's population score (or "true" index on an attribute). A particular subject, who is a member of our population of interest, has an amount of mental rotation ability that we will index as x. 75% of the variance in x may be due to a latent spatial factor, and the other 25% may be due to the influence of a latent creativity factor and a latent action

---

[6] This assumes that there is no aspect of behaviour that cannot be ultimately explained by a subset of the indeterminable number of factors that exist.

planning factor (or any other combination of factors).

Now imagine that we were also interested in this subject's index on map reading. We will call the index for the amount of map reading ability *y*, and decide that 75% of *y* is due to this latent spatial factor, while 25% is due to latent scanning and memory factors. If the mental rotation test and a test of map reading are included in a test battery, 75% of the subject's "true" score on both tests will be shared variance, presumably due to the spatial factor.

Two other abilities of interest may be word production and word comprehension. Indices can be found for all subjects in the population on all these abilities. Because of our selection of ability tests (mental rotation, map reading, word production and word comprehension), the spatial factor will be exposed through the correlation of the indices for mental rotation and map reading. The influence of the verbal, scanning, memory, creativity, and action planning factors will not affect this correlation, as they are assumed to be uncorrelated to spatial ability (and to each other). Similarly, a verbal factor may be exposed through the correlation of word production and word comprehension, this correlation being unaffected by the rest of the factors.

Of course there are usually more than two variables that are inter-correlated, in which case multiple correlations would expose the latent factor. However, it is when the variable's correlation with itself is observed that the distinction between common and specific variance emerges.

There would be a correlation of 1.0 between the vector of population indices for mental rotation ability and itself. Seventy-five percent of this covariance would be shared

with the other variables in the battery of tests due to the influence of the spatial factor (the common variance), and 25% of the covariance would be attributable to the influence of other factors (the specific variance). The percentage of variance due to the factors not exposed through the intercorrelations of the variables (creativity and action planning in this case) can, of course, only be exposed through the correlation of this variable with itself.

Population factor loadings are the correlation between the subjects "true" scores on an attribute, and a latent factor. Assuming simple structure, any variable is related to only one of the latent measured factors. The rest of its variance (the specific variance) is explained by the multitude of other unrepresented factors (i.e. creativity, scanning, memory, and action planning in our example). These "true" values could never be known in a real research setting, but can be known, and are indeed determined, by a researcher who is creating a population data set.

Each time a subject is contacted to be tested, and the test battery is actually administered, variance due to measurement error (the error variance), must be added to the variables. A subject who is measured twice will receive a different score each time, this being due to differing amounts of error variance per testing time. However, the amount of common variance and the amount of specific variance will remain constant over testing occasions, because these sources of variation are fixed in the population.

In summary, the size of the loadings that result from any given dimension reducing technique should be dependent upon two independent sources (disregarding sampling error): the size of the loadings in the population, and the proportion of

measurement error in the sample. However, because the methods of dimension reduction may differ in the way they apportion these sources, the obtained, sample-based loadings may differ for each technique. These sources of variation can and should be simulated when attempting a Monte Carlo study of this nature, but were ignored by the previous researchers.

Sampling Error

An additional source of distortion of the population correlations is sampling error. The combination of subjects drawn from a population will affect the sample correlations, and also the calculated factors. Velicer et al. (1982) and Velicer and Fava (1987) computed simulated "samples" by entering the matrix into a computer program designed to simulate sampling error for a given sample size. The resulting matrix was labelled the sample correlation matrix.

Snook and Gorsuch (1989) created six samples, each differing from the others only in the random numbers that defined the unique variance. The correlation among these uniqueness vectors was apparently not controlled; therefore, the "sample" correlation matrices had random differences in the off-diagonals. The remaining researchers (Borgatta et al., 1986; McArdle, 1990; Widaman, 1990) did not include variance due to sampling error in their simulations.

The above studies simulated sampling error in ways that may not be realistic. For Velicer et al. (1982) and Velicer and Fava (1987), it was not possible to take true samples from a population because they were working only with correlation matrices. Considering Snook and Gorsuch (1989) were working with subjects, it is surprising that

they did not simply create a population of subjects and sample from this. If possible, actually taking a sample from a population of subjects is the most realistic approach to simulating sampling error.

Construction of Population Loading Pattern

Three of the six aforementioned authors used only simple structure matrices with homogeneous loading sizes of .40, .60 and .80 as population loading matrices (Snook & Gorsuch, 1989; Velicer & Fava, 1987; Widaman, 1990). They give no reason for using this pattern beyond simplicity and replication of others' methods.

Velicer et al. (1982) investigated five types of population loading patterns. They ranged in complexity from "ideal" (simple structure with homogeneous loadings) to "complex" (each variable loading on three of six factors, with a moderate loading on one and small loadings on the other two), keeping the number of factors constant at six. This use of complex patterns led to an overdependence on summary statistics and a complicated interpretation. In later studies, Velicer and associates used only population pattern matrices with simple structure and homogeneous loadings of .40, .60, and .80 (e.g. Fava & Velicer, 1992; Velicer & Fava, 1987).

McArdle (1990) experimented with oblique factors, single factor data sets, and loading sizes that varied from .10 to .90. Although there were perhaps too many visual comparisons to get a clear impression of the divergent behaviour of PCA and CFA, the superior precision of CFA in this study was not lost in the complexity of the factor structure. Borgatta et al. (1986) appear to have arbitrarily chosen their complex seven variable, three factor population pattern matrix, but also found CFA to be the superior

method. In all studies, the number of factors underlying the data ranged from one (McArdle, 1990) to twelve (Widaman, 1990).

Comparison to the Population Loading Pattern

Matrices compared

Velicer et al. (1982) compared the loading matrices that resulted from the dimension reducing methods to (a) the matrix that was used to create the population correlation matrix, and (b) the pattern matrix that resulted from analysis of the population correlation matrix. Velicer and Fava (1987) compared the sample loadings only to the loadings that resulted from analysis of the population correlation matrix. Borgatta et al. (1986), Snook and Gorsuch (1989), McArdle (1990), and Widaman (1990) compared the sample correlation matrices only to the loading matrices that were used to create the data.

The object of these simulations was to compare the methods on their ability to match the loading pattern that caused the score variation in the population. Velicer et al. (1982) and Velicer and Fava's (1987) method of comparing the pattern matrix resulting from analysis of the sample correlation matrix to that resulting from an analysis of the population correlation matrix seems to add an additional and unnecessary step. Moreover, it clouds the purpose of the simulations, and the theoretical advantage of analyzing the population correlation matrix is not clear.

Method of Comparison

These researchers also differed on ways of comparing the sample matrix to the population matrix. Velicer and Fava (1987), and Velicer et al. (1982) created a statistic for comparing loading matrices. This was called g, and was calculated as:

(9) $$\mathbf{g} = \text{trace}({}_{m}\mathbf{e}_{p}'{}_{p}\mathbf{e}_{m})/\mathbf{pm}$$

where **e** is a matrix of the difference scores of the two matrices being compared, **p** is the number of variables, and **m** is the number of factors. They verbalize this as the "average (squared) difference between comparable loadings" (Velicer & Fava, 1987, p. 201). This statistic was then manually compared over the methods in Velicer et al. (1982), and submitted to an ANOVA in Velicer and Fava (1987). The standard deviations of *g* were also recorded and observed by these authors.

The major problem with the *g* statistic is that it does not allow researchers to observe the direction of the inaccuracy of a given method. Accuracy is defined as a squared difference, ignoring the direction of the error. The direction of error for the sample loadings must be observed, because this is the only way to determine whether the technique is underestimating or overestimating the loadings.

Additionally, this statistic does not distinguish between zero and non-zero loadings. If there is a relatively small number of non-zero loadings, large differences between them may be lost when averaging over the many small difference scores that result from the zero loadings represent. Moreover, Pennell (1968) points out that different standard errors are expected from zero and non-zero loadings, as is the case with all correlations. These loading types should be observed separately.

Snook and Gorsuch (1989) compared the sample loadings to the population loadings by taking the sum of the difference between the sample loadings and the population loadings and dividing by the number of loadings, making a mean difference score. This was done separately for non-zero and zero loadings, therefore avoiding the

aforementioned problems. This statistic was then analyzed in a three-factor ANOVA to test for significant variation. The standard deviations were not reported.

Widaman (1990) and McArdle (1990) took the mean of all non-zero computed loadings, and compared that to the non-zero population loading. They do not give detailed reports as to the behaviour of the zero loadings, but Widaman (1990) mentions that there was an absence of bias in all conditions for the nondefining loadings. Borgatta et al. (1986) simply displayed the three matrices (generating pattern, PCA, and CFA) in a table, and noted that the CFA pattern was more similar to the generating pattern. None of these researchers reported the standard deviations of their statistic.

Although Velicer et al. (1982) reported the standard deviation of the *g* statistics, they could not report the crucial information regarding the direction of technique divergence. Other researchers reported the direction of the dissimilarities, but did not report the standard deviations. Both types of information are especially important if different sample sizes are to be tested.

The use of ANOVAs to detect significant changes in loading sizes, although useful for studying interactions of effects, may unnecessarily complicate matters. It is difficult to determine what should be considered a subject, making degrees of freedom difficult to determine, and raising questions about assumptions (e.g., random sampling and/or random assignment to conditions), and the validity of the alpha levels (Glass, Peckham, & Sanders, 1972). The researchers that visually compared loading sizes certainly did not forfeit clarity of results, and conveyed their findings conclusively with the use of tables.

Rotational Indeterminacy

An additional complication with comparison of the output loading pattern to that of the population is that when rotation is performed, the order and sign of the factors can change arbitrarily. Varimax rotation was carried out by all researchers who specified their rotation technique (McArdle, 1990, did not specify) with the exception of Velicer et al. (1982) and Velicer & Fava (1987), who used procrustes rotation.

Snook and Gorsuch (1989), McArdle (1990), and Widaman (1990) do not mention rotational indeterminacy, and simply state that they compared the population loading and the comparable loading in the sample pattern. This may indicate that they visually chose the comparable factors, and/or manually reversed the signs. Because they only performed two analyses, Borgatta et al. (1986) did not encounter this complication and simply reported the two loading matrices.

Velicer et al. (1982) and Velicer and Fava (1987) overcame this problem by employing an orthogonal procrustes rotation, which "rotates the patterns into positions of maximum agreement before comparisons are performed" (p. 201, Velicer & Fava, 1987). If the target pattern has simple structure, this rotation will act much like a varimax. However, if the target (population) pattern is not simple structure, procrustes rotation will be more accurate than varimax in finding the population loading pattern.

When researchers do not have access to the population matrix, they usually perform a varimax or another more common rotation. The use of procrustes rotation, therefore, may not be a realistic test of how these factor analytic techniques would perform under typical circumstances when the nature of the population pattern in

unknown.

Analyses Performed

Principal Component Analysis

All of the aforementioned researchers included PCA as one of their analytic methods. This technique has been explained in the introduction to this paper, and is the procedure of choice when selecting a component method.

Common Factor Analysis

All authors also included some type of CFA technique that was compared to PCA. However, the type of CFA that was chosen differed among the researchers. The Principal Factor Analysis (PFA) method has been used by Borgatta et al. (1986), Snook and Gorsuch (1989), and Widaman (1990). Maximum likelihood factor analysis (MLFA) was used by McArdle (1990), Velicer et al. (1982), and Velicer & Fava (1987). Finally, the unweighted least squares procedure (ULS) was also used by McArdle (1990).

According to this sample of studies, the type of CFA technique used was not related to obtained results. For example, McArdle (1990) concluded that CFA was superior to PCA, while Velicer et al. (1982) and Velicer & Fava (1987) concluded that there was no difference between the methods, although all three compared MLFA to PCA. Browne (1968) compared five methods of factor analysis (including MLFA and CFA) on their ability to recover loading patterns from simulated correlation matrices, and found little differences between them. Moreover, McArdle (1990) apparently used MLFA and ULS interchangeably, reporting no differences in loading size.

Image Analysis

Image analysis (IA) originated with Guttman (1953), and was tested by Velicer et al. (1982) and Velicer and Fava (1987). These researchers considered IA a type of component analysis, since the factors are extracted from a matrix derived directly from the raw scores, and no estimation of uniqueness is necessary. However, IA is considered by others to be a type of CFA, since its model explicitly includes error (e.g. Gorsuch, 1990). For these reasons, IA is often viewed as a compromise between PCA and CFA (e.g. Gorsuch, 1990; Kaiser, 1970; Nunnally, 1978, p. 416).

Velicer et al. (1982) and Velicer and Fava (1987) utilize a variant of Guttman's IA called *image component analysis* (ICA). This involves performing SVD on the matrix $\mathbf{R}^*$:

$$(10) \qquad {}_p\mathbf{R}^*_p = {}_p\mathbf{S}_p^{-1}\,{}_p\mathbf{R}_p\mathbf{S}_p^{-1},$$

where **S** is a diagonal matrix containing the square root of the reciprocals of the diagonal elements of $\mathbf{R}^{-1}$ (which is equal to the standard deviation of the anti-images, $\sqrt{\mathbf{1}-\mathbf{R}^2}$), and **R** is the correlation matrix. $\mathbf{R}^*$, therefore, is simply the correlation matrix rescaled to have the metric of the anti-images (this allows ICA to be a scale-invariant technique). The final loading pattern is produced by pre-multiplying the loadings calculated from the SVD of $\mathbf{R}^*$ by **S**. This step serves to re-adjust the scale of the loading matrix from that of the anti-images to that of the correlation matrix.

Methodological Biases

Although theoretical discussion has not yet produced a clear preference for CFA or PCA, empirical evidence has been more conclusive. All empirical research in the

area, excluding that involving Velicer, has found CFA to be more accurate at retrieving the population loading pattern than PCA. However, as previously noted, when variables without measurement error are used to compute the population correlation or data matrix (as has been the case with all previous simulations), studies will be biased towards CFA.

When unique variance is added to the FCM, and the resulting matrix is subjected to PCA, all variance is treated as common and distributed to components. Each factor must explain units of *perceived* common variance in addition to the true common variance (recall that all true common variance is represented in the FCM). These additional units of variance are split over the non-zero loadings for that factor, biasing the sample loadings upward.

CFA, however, attempts to remove this additional variance before performing SVD by adjusting the diagonal entries of the correlation matrix. Because CFA eliminates the unique variance before SVD, it does not need to explain variance above that attributable to the population loadings, and can accurately retrieve the population loadings when there is no measurement error.

The exceptions to this postulation are the Velicer studies, but this inconsistency is explainable. Most importantly, recall that the statistic *g* averages zero and non-zero loading differences together. Widaman (1990) and Snook & Gorsuch (1989) report no differences between PCA and CFA on the zero loadings, but important differences on the non-zero loadings. Therefore, because the majority of the population loadings in Velicer's studies were zero, differences between the techniques on non-zero loadings may have been missed.

There are further possible explanations for these findings. Velicer et al. (1982) may not have found a difference between the methods because they used a large number of variables (36), allowing the extra variance to be spread thinly over a greater number of non-zero loadings. Velicer & Fava (1987) were the only researchers to use a small number of variables (12, 18 and 24) and not detect a difference between PCA and CFA, although they report "small differences favouring MLFA over PCA over ICA" (p. 203). This oddity may have occurred because Velicer and Fava (1987) compared the sample loadings only to the loadings that resulted from analysis of the population correlation matrix, instead of to the population loadings used to create the population correlation matrix. It may also have been the case that PCA was slightly overestimating the loadings while MLFA and ICA were slightly underestimating the loadings (or vice-versa), but the researchers were not alerted to this because of the use of the statistic g to evaluate performance (recall that g averages squared differences). The present study will attempt to test the above conclusions when there are varying proportions of measurement error added to the data.

Present Research

For the present simulation, a computer program was written that creates sample data sets as follows: a population FCM was created by multiplying the population pattern matrix by the population factor scores. Specific variance was added to the FCM. Multiple samples of the desired size were taken from this population. Each subject, once sampled, was then assigned a score to represent error variance.

The following results were expected:

1. CFA will recover population loadings more accurately than PCA when there is no measurement error and a small number of variables.

2. PCA, CFA, and IA will produce essentially equivalent results when there is no measurement error and a large number of variables.

3. Loading sizes will decrease as measurement error is added.

METHOD

Apparatus

The computer program for this simulation was written using 386-MATLAB for 80386-based MS-DOS personal computers. 386-MATLAB is an interactive software package that integrates numerical analysis and matrix computation. All program source code has been included in Appendix D. Some simulated data sets have been subjected to identical analyses using SPSSPC for Windows or BMDP, the results of which agreed with those of the present study.

Procedure

Parameters

Following the methods of previous researchers (Snook & Gorsuch, 1989; Widaman, 1990; Velicer & Fava, 1987), the population loading patterns were constructed to have simple structure, with uniform loading sizes of .40, .60, or .80. To facilitate comparison to Snook & Gorsuch (1989), who used a similar data construction technique, the number of factors was kept constant at three, with all variables loading equally onto a single factor.

The simple structure, homogeneous loading pattern seems reasonable in

preliminary simulations, although there is no major theoretical problem with using more complicated patterns. Using simple structure reduces the probability that variations in loading pattern will be due to an inappropriate rotation technique, since varimax searches for simple structure. It also simplifies the specific variance addition step, because all columns in the FCM will have the same variance when homogeneous loadings are used. In addition, it allows the researcher to report one average non-zero loading size per analysis, simplifying interpretation. More complicated and/or realistic population loading patterns should be attempted after this simplest case is understood.

Researchers have recommended 36 as a large enough number of variables to nullify differences between PCA and CFA (Gorsuch, 1983, p. 123; Nunnally, 1978, p.419; Velicer et al., 1982). Conversely, simulated data sets with nine variables have invoked discrepancies between PCA and CFA (Snook & Gorsuch, 1989). Nine variables also meets the minimum identifiability constraint of 3 recommended by Anderson and Rubin (1956) ($p/m = 3$). Therefore, variable sets ranged from 9 to 36, with 18 included as an intermediate comparison, replicating Snook and Gorsuch (1989). Factor identifications under these circumstances were 3, 6, and 12 variables. Examples of population loading matrices are presented in Table 1.

The minimum acceptable sample size is generally recognized to be between 100 and 200 subjects, although this may depend on the number of variables (Gorsuch, 1982, p. 332; Comrey, 1973, p. 200; Guadagnoli and Velicer, 1988). However, it is fair to suppose that many researchers have attempted to perform dimension reduction with less than an acceptable number of subjects. For this reason, we will test the performance of

the techniques at poor (50), marginal (100), and acceptable (200) sample sizes. Due to the constraints of random-access memory size and speed of computation, population size was kept constant at 1000, and 20 samples were taken per condition. Percentage of measurement error ranged from 0 to 75 percent (0, 25, 50, and 75%).

Analyses

Principal Component Analysis

All simulated data sets were subjected to PCA, virtually the only component method used by social scientists.

Common Factor Analysis

Since the type of CFA used seems unrelated to previously obtained results, PFA was chosen as the CFA method to facilitate comparison to Snook and Gorsuch (1989). It differs from PCA in that estimates of communality, instead of 1.0s, are placed on the diagonals of the observed correlation matrix before extraction of the factors. These estimates of communality begin with squared multiple correlations resulting from the regression of each variable on all other variables. The output communalities that result from factor extraction from this adjusted correlation matrix (through SVD) are then placed on the diagonals, and the process is repeated until there is no significant change in the size of the communalities.

Gorsuch (1983, p. 107; 1990) suggests limiting the number of iterations to two or three because it is not clear what the communalities are converging towards. However, most researchers allow the communalities to be reestimated until negligible change occurs in the estimates, as this is the default for most statistics packages (Tabachnick & Fidell,

1989, p. 660-661; Norusis, 1988, p. B-52). The present program performed 7 iterations before stopping in an attempt to simulate a realistic but conservative situation[7].

Image Analysis

Image Analysis (Guttman, 1953) clearly defines what is to be considered the common and unique parts of a variable. The common part of a test score is defined as that part of its variance which is predictable by linear multiple correlation from all the other variables (called the *partial image scores*). The unique part is defined as that part of the variance of a variable that is not predictable from all the other variables (called the *partial anti-image scores*)[8].

Conceptually, IA proceeds by performing SVD on the variance-covariance matrix of the image scores of the variables. This image covariance matrix can also be calculated directly from the correlation matrix:

$$(11) \qquad {}_pG_p = {}_pR_p + {}_pS^2{}_pR_p^{-1}{}_pS^2{}_p - 2_pS^2{}_p$$

where **G** is the variance-covariance matrix of the images, and $\mathbf{S^2}$ is a diagonal matrix of the reciprocals of the diagonal elements in $\mathbf{R^{-1}}$ (or the variance of the anti-images, $\mathbf{1-R^2}$).

Inspection of available statistical software packages reveals that different packages use very different algorithms for IA. The Velicer et al. (1982) and Velicer & Fava (1987) variant of IA (image component analysis) is not used by SPSS, BMDP, or SAS. Thus, in the interest of investigating procedures most likely to be used in practice, the

---

[7]The presence or absence of Haywood cases (an estimation of communality that results in a diagonal element greater than 1.0) was not observed in this simulation.

[8]Guttman referred to the sample images as partial image scores but to the population image scores as *total image scores*.

current study based IA on the algorithm used in BMDP and SPSS (ref. Kaiser, 1963).

This approach consists of performing SVD on the matrix $\mathbf{G}^*$, which results from the following formula:

(12) $$ {}_pG_p^* = {}_pS_p^{-1}\,{}_pG_pS_p^{-1} $$

where $\mathbf{S}$ is the square root of $\mathbf{S}^2$ from equation (11), and $\mathbf{G}$ is the image variance-covariance matrix from equation (11). To arrive at the final loadings, the loadings calculated from the SVD of $\mathbf{G}^*$ must be rescaled by pre-multiplying by $\mathbf{S}$.

Performance Evaluation

All loading matrices were rotated to varimax criterion. The MATLAB program was designed to switch reversed signs, and to order the factors as they are ordered in the population. The output loading pattern was visually compared to the generating pattern by averaging all non-zero loading and comparing to the corresponding population non-zero loading, and similarly for zero loadings. The mean standard deviation of the loadings over multiple samples was also recorded.

Data Generation

The population FCM was created using the formula

(13) $$ {}_Nx_{fcp} = {}_Nf_mA_p' $$

where $\mathbf{x}_{fc}$ is the resulting population factor contribution data matrix, $\mathbf{A}$ is the population loading matrix, and $\mathbf{f}$ is the standardized population factor score matrix. Equation (13) must be modified to account for the addition of specific variance:

(14) $$ {}_Nx_p = {}_Nf_mA_p' + {}_Ne_p $$

where $\mathbf{x}$ is the matrix of population scores on the variables, and $\mathbf{e}$ is the orthogonal

matrix of variation in the data due to unrepresented factors.

When a sample was taken from a population and measured, two sources of variation were added: (a) sampling error, and (b) measurement error. Sampling error was simulated by taking random samples of subjects (**X**) from the population matrix **x** for analysis. Measurement error was simulated for each sample by adding a matrix of random numbers to **X**, the variance of which was be constrained to result in a defined percentage of error variance in the final scores. This relationship can be expressed by the following formula:

(15) $$_{n}\mathbf{DATA}_{p} = {}_{n}\mathbf{X}_{p} + {}_{n}\mathbf{ME}_{p}$$

where **DATA** is the final sample data matrix of standardized scores, **X** is the sample data matrix before the addition of measurement error, and **ME** is the matrix of measurement error. In accordance with the assumptions of factor analysis, the columns of the matrix **ME** are uncorrelated with each other, and are uncorrelated with the columns of **X**.

All sources of variation on the final data set (disregarding sampling error) can be expressed as follows:

(16) $$_{n}\mathbf{DATA}_{p} = {}_{n}x_{fcp} + {}_{n}e_{p} + {}_{n}\mathbf{ME}_{p}.$$

The final scores are constrained to be standardized, with a variance of 1.0 and a mean of 0.

RESULTS

The full range of results have been presented in Appendix C. Tables 2-10 will be referred to in the text for more specific findings. The most salient result of this simulation was that PCA consistently produced high loadings relative to CFA and IA,

whereas IA produced low loadings. These loading size differences diminished as the number of variables increased, but increased as the overall percentage of error (specific + error variance) increased (see Tables 2-10). The largest discrepancy in the loading sizes between the techniques occurred when there were only nine variables, small population loadings, and high measurement error (Table 4, 75% measurement error column). Correspondingly, the smallest discrepancy occurred when there were 36 variables, large population loadings, and no measurement error (see Table 8, 0% measurement error column).

Another noticeable and perhaps more important finding was that when there was 25% or greater measurement error, PCA gave the clearest representation of the population factor structure, particularly when there were only 9 variables (see Tables 2-10, especially Tables 2, 3, and 4). Although this technique sometimes exaggerated the size of the population loadings, this overestimate may simplify interpretation, since the zero loadings did not differ systematically between techniques (see Table 11, and Appendix C). Image analysis, however, so severely underestimated the loadings that at high error levels two or three of the factors were often indistinguishable to the computer (see Table 12). This indicates that the simple structure pattern was lost due to uniformly low loadings. Alternatively, when there was no measurement error, CFA perfectly recovered the loadings on average (see Tables 2-10). This remarkable accuracy of CFA was independent of sample size or number of variables.

The results of Snook and Gorsuch (1989), Borgatta et al. (1986), and McArdle (1990) were replicated in the finding of superior recovery of the factors for CFA when

there was no measurement error, and 9 variables (see Tables 2, 3, and 4, the 0% measurement error column), supporting expected result #1. The results of Velicer et al. (1982), Velicer & Fava (1987), and Widaman (1990) (small number of factors condition only) were also replicated, with small differences between the methods being found for a large number of variables and no measurement error (see Tables 8, 9, 10, the 0% measurement error column), supporting expected result #2. Expected result #3 was also supported, as is obvious from all results tables.

The standard deviation changes over sample size were as was expected, with decreasing variation as sample size increased (see Appendix C, and Table 13). The standard deviations of the mean zero loadings corresponded closely to that of the mean non-zero loadings for identical population loading size, sample size, technique, and percentage of measurement error. There were no systematic differences between the techniques in standard deviation sizes after the different loading sizes were accounted for.

When there was a very low variable to subject ratio (36 variables and 50 subjects) and high measurement error, there was too much variation of the loadings to have faith in any technique under these circumstances. For example, the mean standard deviation of the mean loadings over 20 samples for PCA was .3021 when there were 36 variables, 50 subjects, 75% measurement error, and a population loading size of .8 (see Table 13 and Appendix D). This is because if only 50 subjects must convey information about how 36 variables are related to factors, the combination of subjects sampled will have a profound effect of the clarity of the structure achieved, particularly when the structure has already been distorted by measurement error.

As was the case with previous researchers, the mean of the zero loadings for all conditions was accurate. The number of times two or more factors could not be distinguished by the computer program generally increased as measurement error increased, and as population loading size and sample size decreased (see Appendix C).

DISCUSSION

Snook and Gorsuch (1989), Borgatta et al. (1986), McArdle (1990), and Widaman (1990) found superior recovery of factors when CFA was performed, whereas Velicer et al. (1982) and Velicer and Fava (1987) found no difference in the ability of PCA, CFA, and IA to recover the factor structure of simulated data sets. The present study proposed that the above researchers neglected to include variation due to measurement error in their simulated data sets, and comes to different conclusions when this source of variation is simulated.

The conclusions reached were as follows: if the researcher is confident that the variables have been measured without measurement error (or with very little, i.e. below 25%), CFA is the recommended choice. In this simulation, this technique was extremely accurate under these circumstances, remaining consistent over sample size and number of variables.

If, however, the researcher expects there to be an appreciable amount of measurement error, CFA will produce low loadings. The researcher will not be able to determine the extent to which this is due to either poor factor structure or poorly measured instruments. Conversely, because PCA interprets all variance as common variance, it raises the sample loadings that represent non-zero population loadings when

finding the factor pattern, clarifying the factor structure. When the population loadings were low, PCA clarified the population loading pattern by raising the sample non-zero loadings, even when there was substantial measurement error. When the population loadings were high, PCA increased loadings that had been deflated by measurement error, giving an accurate representation of the population loading pattern. The exact relationship between the size of the sample PCA loadings, the loadings in the population, and the number of variables is presented below.

The following discussion focuses on a simulation where there are 9 variables, 3 factors, no measurement error, no sampling error (i.e. the sample is the population), and .80 population loadings in simple structure. This simulation begins by computing the population FCM, the variance of which is due only to the three factors. Since the standardized factor scores are rescaled by loadings of .8, the total variance of this matrix is $.8^2*9$, or 5.76. An eigen-decomposition of this matrix (preserving the scaling) would produce three positive eigenvalues (one for each of the three factor patterns), and six zero eigenvalues.

After the population uniqueness is added, and the variance of the variables is raised to 1.0, only 64% of the variables' total variance of 9.0 will be due to the factors (5.76/9.0 = .64). Therefore, 36% (3.24 units of the total 9.0 units of variance) of the total variance of the scores is unique, but will be interpreted as common by the method. This unique variance is divided up over the factors, so .36 (3.24/9 = .36) units of additional variance for each of the *p* factors must be explained. Now the previously zero eigenvalues become slightly greater than zero, and the three original eigenvalues become

larger.

For the three large eigenvalue factors, while they originally explained 1.92 units of variance each, they now must explain 2.28 (1.92 + .36) units. This extra variance, being interpreted as common variance, increases only the size of the non-zero loadings (three in our example), since they are the only loadings to represent common variance on that factor. Therefore, since 2.28/3 = .76, .76 is the percentage of variance shared by this variable and the factor, and is the square of the final mean loading found when PCA is performed on the population (.8718). As can be seen in the results (Table 2, PCA, 0% measurement error), sampling error reduces the loading size, but this loading will be found if the full population is analyzed.

This reasoning can be extended to an example where there is 50% measurement error. This will cause the common variance in the scores to be half of when there was no measurement error (32%, or 2.88 units). The loadings resulting from this common variance are now reduced to .566[9], due to measurement error. The variance not due to the factors is increased to 68%, so each factor must explain an additional .68 units of variance. The original 3 factors, when they explained all the common variance, explained .96 units of variance (2.88/3) each. They now must explain 1.64 units (.96 + .68), and spread this over the non-zero loadings. The loadings are raised to the square root of .5467 (1.64/3), which is .7394. This estimate was again reduced by sampling error in our results (see Table 2, PCA, 50% measurement error).

---

[9].566 is the square root of .32, which results from ((2.88/3 [factors])/3 [variables loading on the factor]).

By raising only the non-zero loadings in the sample, PCA clarified the structure of the factors which may have otherwise been lost through attenuation due to measurement error. The greater percentage of measurement error, the more PCA was an improvement over the other techniques that separate that variance out. This effect was reduced as the number of variables increased, since the unique variance could then be spread over more non-zero loadings per factor.

Unlike PCA, CFA attempts to separate out the variance due to the factors and that due to uniqueness. Therefore, using the above example with no measurement error, this technique attempts to explain only the 1.92 units of variance per factor that was present in the population. No extra variance must be explained by each factor. Spread over the three non-zero loadings per factor, the original loadings of .8 are retrieved (see Table 2, PFA, 0% measurement error). The stability of this finding over sample size and variable number also gives evidence that PFA very accurately distinguishes between common and unique variance independent of these parameters.

As measurement error is added, the percentage of total variance in the scores that CFA interprets as common decreases. For example, if there is 50% measurement error, the total common variance is reduced by half, as are the units of variance to be explained by each factor. Relating this back to our PCA example with 50% measurement error, only the 2.88 units of common variance will be split over three factors, resulting in .96 units of variance per factor. As can be seen in Table 2, PFA, 50% measurement error, the loadings of .566 can be found nearly perfectly retrieved. This explains the decrease in the size of the loadings as the amount of measurement error increases.

Image analysis also attempts to separate out common and unique variance, but provided less accurate estimates of the population loadings than both the other techniques. There is little reason to recommend its use based on these empirical results. The loadings it produced were often too low to maintain clarity of factor structure.

The behaviour of this technique may be attributable to its use of multiple regression to estimate the amount of common variance. Guttman (1956) proved that R^2 is a *lower bound* for the communality. Perhaps the true communality, or how much a particular variable has in common with a set of hypothetical factors, is underestimated by R^2. This appears to be the case in these simulations. Analysis of one of the simulated data sets showed that the R^2 for a variable being predicted by all the other variables was lower than its true population communality[10].

An additional unusual result of the image analysis simulation was that the size of the loadings *decreased* as sample size increased, moving away from the population loadings. This effect is probably caused by the decomposition of the covariance matrix of the partial images. When the subject to variable ratio decreases, the goodness of fit to the sample increases, and the proportion of predictable variance in the dependent variable is overestimated relative to the population. Since the partial images have artificially increased variance in relation to the total images, the covariances in the partial image covariance matrix will also be inflated. As sample size increases, the size of the variance of the partial images will decrease, as will the values of the partial image

---

[10] $R^2 = .50$ for a given variable in a population data set with 9 variables made from a simple structure loading matrix of .8's (communality for that variable is .64).

covariance matrix, and the final loadings.

The point is clarified when these results are compared to those from a different type of image analysis known as image component analysis, which was used by Velicer et al. (1982) and Velicer & Fava (1987) in their simulation. This method decomposes the rescaled correlation matrix instead of the partial image covariance matrix. Under these circumstances, the loading sizes increase as sample size increases, just as in PCA. Since the only difference between the two methods is the use of the partial image covariance matrix, herein must lie the reason for the opposing size changes with sample growth.

The image analysis findings lead to a strange conclusion. Although smaller samples seem to cause the population loadings to be better estimated, they are actually biased estimates of loadings that are too low in the first place. This is similar to the inflated loadings of PCA, because in both cases only the non-zero loadings inflate. Under these circumstances the population loadings can be found only when two biases cancel each other out. It is likely that in more realistic situations, the behaviour of these biases would be much less predictable.

In summary, the present study has found that the comparative performance of PCA, CFA, and IA depend largely on the amount of measurement error present in the chosen battery of tests. Because average non-zero loading size decreased as measurement error increased for all dimension reducing methods, and because PCA loadings were least affected by this relationship, PCA is recommended for most research situations in the social sciences. If, however, the researcher feels confident that the variables have no or very little measurement error, CFA is recommended.

Relation to Theory

Considering the preference of social scientists to err on the conservative, it is perhaps unusual to recommend a technique that under certain circumstances exaggerates the pattern of variation underlying the data. However, since many practitioners use dimension reduction as an exploratory technique, the behaviour of PCA can only enhance their understanding of the data they have collected. Mulaik (1990) writes:

> Exploratory common factor analysis is at best one among many methods one might use initially in the course of trying to formulate hypotheses about causal structures underlying the variables of a domain (p. 55).

Psychological problems are often, out of necessity, exploratory. In these cases, the technique that can overcome the distorting effect of measurement error and reflect population patterns is the superior technique.

Most practitioners who wish to reduce the dimensionality of their data set would like to discover something about the latent factors that are causing the variables to correlate. The model implicit in PCA, which attempts to form weighted aggregates of variables, does not take this theoretical approach (as does CFA). However, if PCA more clearly represents the underlying relationships between variables and factors, it may be preferable to use this technique. Mulaik (1990) further writes:

> Much of the motivation to use the component analysis model is that a component analysis model is often a very good approximation of a common factor model...But using component analysis in this way need not mean one has abandoned thinking of the data in terms of a common factor model (p. 54).

Components fit naturally into the CFA model with the modification that there is no uniqueness component. However, factors cannot fit the model implicit in PCA, because factors are not formed directly from variables. Referring to Figure 1, the

direction of the arrows in the causal model for PCA can be changed to fit whichever model one wishes to propose, making components more "theoretically flexible" than CFA.

Since components can always be modelled regardless of what one's theoretical preferences are, proponents of CFA who claim superiority for theoretical reasons must re-think their argument. Again, the solution to theoretical debate lies in empirical investigation when studying the differences between the techniques. Given that there is no opportunity to make clear hypotheses about the expected factor structure, but one wishes to explore the factor structure of one's data, this empirical investigation suggests the superiority of PCA under the normal circumstances of analyzing variables with measurement error.

Directions for Future Research

Although this simulation study may have added the dimension of measurement error to data simulation for this research area, to fully understand the comparative behaviour of PCA, CFA, and IA much more realistic data sets must be created. The present population loading pattern is highly artificial, and the probability of this variable-factor relationship actually occurring is low. It should be further noted that this simple structure population pattern follows the logic of Thurstone and the American school of factor analytic thought that developed within the context of intelligence testing. This pattern in not universally accepted.

For example, Spearman and the British factor analysts believed that there is one general factor of intelligence, and multiple specific factors that are secondary to this main

one (Spearman, 1904). If Spearman's factor-variable relationship existed in the population, using varimax rotation a) would distort the fairly accurate representation of the population factor pattern that the unrotated factor (or component) solution may give, and b) may provide quite different simulation results than those found using Thurstone's hypothesized pattern and varimax rotation.

Attempting more complicated loading patterns (such as the one proposed by Spearman) would be the first logical step towards fully understanding the implications of these results. For example, recalling that PCA spreads extra variance evenly over non-zero loadings and not over zero loadings, would it raise population loadings of .10 by the same amount as population loadings of .80 on the same factor? If more or equal increases are given to low population loadings, this may severely exaggerate the relative perceived importance of a zero population loading and one of .10 or lower.

Another improvement would be to add varying amounts of measurement error to different variables. The amount of measurement error usually varies across tests within a battery, and this may have an effect the way the methods recover the population loadings. One may also attempt using more realistically sized populations - a population size of 1000 may severely underestimate the true amount of sampling error that arises when a sample is taken from a large population.

Furthermore, one could attempt to vary the number of subjects per variable, the number of subjects per factor, or the number of variables per factor. For example, keeping sample size, population loading size, and measurement error constant, a data set with 9 variables and 3 factors (3:1 variable to factor ratio) may be analyzed similarly to

a data set with 12 factors and 36 variables. In this case, recommendations for methods may be affected by the more general *variable-to-factor* ratio instead of only the number of variables.

Thus, in retrospect, the previously mentioned postulation of Thurstone and Gorsuch that PCA is a valid technique only when one's variables are measured without error takes a sharp reversal. The findings from the present empirical study suggest that PCA is an inferior technique when the variables are measured with no measurement error, with CFA being more accurate. PCA, however, gave a clearer representation of the population factor structure than CFA when the variables were measured *with* error. Thurstone and Gorsuch have attempted to make recommendations based on how well the data characteristics correspond to the way the model illustrates the data. However, whether or not the technique handles the data as the model seems to predict is another issue - the resolution of which can be greatly aided by empirical investigation.

REFERENCES

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5,* 111-150.

Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*, 67-74.

Bookstein, F. L. (1990). Least squares and latent variables. *Multivariate Behavioral Research, 25,* 75-80.

Borgatta, E. F., Kercher, K., & Stull, D. E. (1986). A cautionary note on the use of principal components analysis. *Sociological Methods and Research, 15*, 160-168.

Browne, M. W. (1968). A comparison of factor analytic techniques. Psychometrika, 3 267-334.

Comrey, A. L. (1973). *A First Course in Factor Analysis*. New York: Academic Press.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27*, 387-415.

Girshick, M. A. (1936). Principal components. *American Statistical Association, 31*, 519-528.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288.

Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: some well and little known facts. *Multivariate Behavioral Research, 25*, 33-39.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika, 18*, 277-296.

Guttman, L. (1956). Best possible systematic estimates of communalities. *Psychometrika, 21*, 273-285.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417-441, 498-520.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Kaiser, H. F. (1963). Image Analysis. In C. W. Harris (Ed.), *Problems in Measuring Change* (pp. 156-166). Madison, Wis: University of Wisconsen Press.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-415.

McArdle, J. J. (1990). Principles versus principals of structural factor analyses. *Multivariate Behavioral Research, 25*, 81-87.

McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hillsdale: Lawrence Erlbaum Associates.

Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York: McGraw-Hill Book Company.

Mulaik, S. A. (1990). Blurring the distinction between component analysis and common factor analysis. *Multivariate Behavioral Research, 25*, 53-59.

Norusis, M. J. (1988). *SPSS/PC+ Advanced Statistics V2.0*. Chicago: SPSS Inc.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 6*, 559-572.

Pennell, R. (1968). The influence of communality and N on the sampling distributions of factor loadings. *Psychometrika, 33*, 423-439.

Rozeboom, W. W. (1990). Whatever happened to broad perspective? *Multivariate Behavioral Research, 25*, 61-65.

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: a Monte Carlo study. *Psychological Bulletin, 106*, 148-154.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201-285.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using Multivariate Statistics, Second Edition.* New York: Harper & Row.

Thurstone, L. L. (1935). *The Vectors of Mind.* Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple-Factor Analysis.* Chicago: University of Chicago Press.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 22,* 193-209.

Velicer, W. F., & Jackson, D. N. (1990a). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1-28.

Velicer, W. F., & Jackson, D. N. (1990b). Component analysis versus common factor analysis: some further observations. *Multivariate Behavioral Research, 25,* 97-114.

Velicer, W. F., Peacock, A. C, & Jackson, D. C. (1982). A comparison of component and factor patterns: a monte carlo approach. *Multivariate Behavioral Research, 17,* 371-388.

Widaman, K. F. (1990). Bias in pattern loadings represented by common factor analysis and component analysis. *Multivariate Behavioral Research, 25,* 89-95.

Table 1

Examples of Population Loading Matrices

Population Loading = .8

0.8	0	0
0.8	0	0
0.8	0	0
0	0.8	0
0	0.8	0
0	0.8	0
0	0	0.8
0	0	0.8
0	0	0.8

Population Loading = .6

0.6	0	0
0.6	0	0
0.6	0	0
0	0.6	0
0	0.6	0
0	0.6	0
0	0	0.6
0	0	0.6
0	0	0.6

Population Loading = .4

0.4	0	0
0.4	0	0
0.4	0	0
0	0.4	0
0	0.4	0
0	0.4	0
0	0	0.4
0	0	0.4
0	0	0.4

Note. Matrices for 9 variables and 3 factors.

Table 2

Mean Non-zero Loading Sizes for 9 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.8579	0.8052	0.7324	0.6571
100	0.8680	0.8020	0.7397	0.6611
200	0.8700	0.8070	0.7372	0.6621

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7947	0.6985	0.5541	0.4010
100	0.7976	0.6890	0.5612	0.3989
200	0.7999	0.6939	0.5668	0.3979

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7040	0.5371	0.3641	0.1862
100	0.6941	0.5304	0.3627	0.1806
200	0.6839	0.5301	0.3567	0.1849

Note. Population loading size was .80.

Table 3

Mean Non-zero Loading Sizes for 9 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7396	0.7012	0.6424	0.5822
100	0.7500	0.7039	0.6633	0.6186
200	0.7507	0.7130	0.6711	0.6237

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.6069	0.5255	0.4284	0.3050
100	0.6062	0.5247	0.4256	0.2985
200	0.6018	0.5221	0.4268	0.3009

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.4274	0.3383	0.2150	0.1167
100	0.4269	0.3260	0.2210	0.1020
200	0.4120	0.3180	0.2065	0.1057

Note. Population loading size was .60.

Table 4

Mean Non-zero Loading Sizes for 9 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.5633	0.5357	0.4799	0.4693
100	0.5956	0.5762	0.5601	0.5284
200	0.6473	0.6201	0.5930	0.5778

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.4353	0.3597	0.2874	0.2132
100	0.4222	0.3345	0.2867	0.2056
200	0.4022	0.3546	0.2835	0.2047

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.2343	0.1946	0.1149	0.0638
100	0.2119	0.1472	0.1000	0.0498
200	0.1768	0.1429	0.0955	0.0437

Note. Population loading size was .40.

Table 5

Mean Non-zero Loading Sizes for 18 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.8276	0.7485	0.6524	0.5429
100	0.8325	0.7507	0.6541	0.5449
200	0.8340	0.7512	0.6582	0.5465

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7914	0.6905	0.5628	0.3944
100	0.7933	0.6877	0.5632	0.3992
200	0.8008	0.6905	0.5660	0.3987

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7715	0.6258	0.4621	0.2675
100	0.7626	0.6226	0.4665	0.2682
200	0.7554	0.6193	0.4644	0.2682

Note. Population loading size was .80.

Table 6

Mean Non-zero Loading Sizes for 18 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.6648	0.6145	0.5506	0.4721
100	0.6782	0.6212	0.5560	0.4853
200	0.6810	0.6241	0.5595	0.4874

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.5969	0.5160	0.4237	0.2923
100	0.5965	0.5070	0.4201	0.2995
200	0.5986	0.5187	0.4262	0.2995

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.5403	0.4314	0.3086	0.1692
100	0.5322	0.4189	0.2955	0.1659
200	0.5244	0.4134	0.2980	0.1621

Note. Population loading size was .60.

Table 7

Mean Non-zero Loading Sizes for 18 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.4331	0.4236	0.3940	0.3368
100	0.5065	0.4778	0.4388	0.3993
200	0.5402	0.5046	0.4674	0.4314

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.3666	0.3409	0.2784	0.1918
100	0.3948	0.3306	0.2856	0.1929
200	0.3965	0.3469	0.2854	0.2008

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.3307	0.2507	0.1896	0.0908
100	0.3103	0.2478	0.1641	0.0849
200	0.2905	0.2196	0.1561	0.0836

Note. Population loading size was .40.

Table 8

Mean Non-zero Loading Sizes for 36 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.8066	0.7184	0.5627	0.3645
100	0.8119	0.7197	0.6128	0.4772
200	0.8170	0.7223	0.6118	0.4780

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7986	0.6837	0.5296	0.2930
100	0.7966	0.6893	0.5627	0.3994
200	0.7962	0.6905	0.5643	0.3994

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.7997	0.6675	0.4644	0.2871
100	0.7809	0.6610	0.5140	0.3230
200	0.7851	0.6569	0.5138	0.3258

Note. Population loading size was .80.

Table 9

Mean Non-zero Loading Sizes for 36 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.6315	0.5497	0.4135	0.2479
100	0.6400	0.5727	0.4933	0.4003
200	0.6384	0.5753	0.4961	0.4037

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.5911	0.5053	0.3710	0.1871
100	0.5971	0.5130	0.4222	0.2979
200	0.5997	0.5238	0.4254	0.3008

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.6047	0.4990	0.3576	0.2283
100	0.5801	0.4778	0.3616	0.2176
200	0.5632	0.4708	0.3608	0.2170

Note. Population loading size was .60.

Table 10

Mean Non-zero Loading Sizes for 36 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.4340	0.3600	0.2722	0.1991
100	0.4645	0.4166	0.3835	0.3268
200	0.4712	0.4294	0.3872	0.3365

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.3818	0.3131	0.2181	0.1138
100	0.3873	0.3463	0.2785	0.1985
200	0.4010	0.3436	0.2802	0.1997

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.3792	0.2819	0.2060	0.1226
100	0.3813	0.2943	0.2088	0.1157
200	0.3563	0.2853	0.2008	0.1117

Note. Population loading size was .40.

Table 11

Mean Zero Loading Sizes for 9 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.0019	-0.0019	-0.0092	-0.0035
100	0.0057	0.0053	-0.0004	-0.0012
200	0.0009	0.0042	-0.0023	-0.0025

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.0155	0.0039	-0.0053	-0.0014
100	0.0132	0.0014	-0.0011	0.0029
200	-0.0015	0.0023	0.0021	0.0023

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	-0.0036	-0.0011	0.0067	0.0064
100	-0.0088	-0.0066	-0.0024	0.0038
200	-0.0045	0.0038	0.0005	0.0001

Note. Population loading size was .80.

Table 12

Number of Samples Where Factors Were Indistinguishable For 9 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	7	6	7	8
100	2	2	4	1
200	0	0	0	0

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	7	9	11	6
100	2	1	1	5
200	0	0	2	1

Image Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	17	12	17	12
100	15	12	16	17
200	14	14	14	18

Note. Population loading size was .40.

Table 13

Mean Loading Standard Deviations of Non-Zero Loadings For 36 Variables

Principal Component Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.0512	0.0660	0.1778	0.3021
100	0.0321	0.0287	0.0215	0.0175
200	0.0203	0.0186	0.0154	0.0106

Principal Factors Analysis

	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.0536	0.0722	0.1861	0.2636
100	0.0359	0.0313	0.0249	0.0177
200	0.0239	0.0193	0.0169	0.0117

Image Analysis

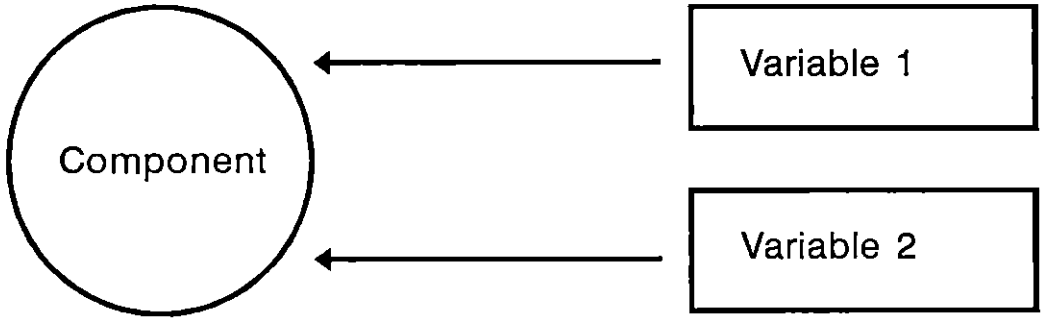
	Percentage of Measurement Error			
Sample Size	0	25	50	75
50	0.0537	0.0883	0.1910	0.2865
100	0.0363	0.0300	0.0239	0.0154
200	0.0216	0.0208	0.0154	0.0105

Note. Population loading size was .80.
Number of samples was 20.

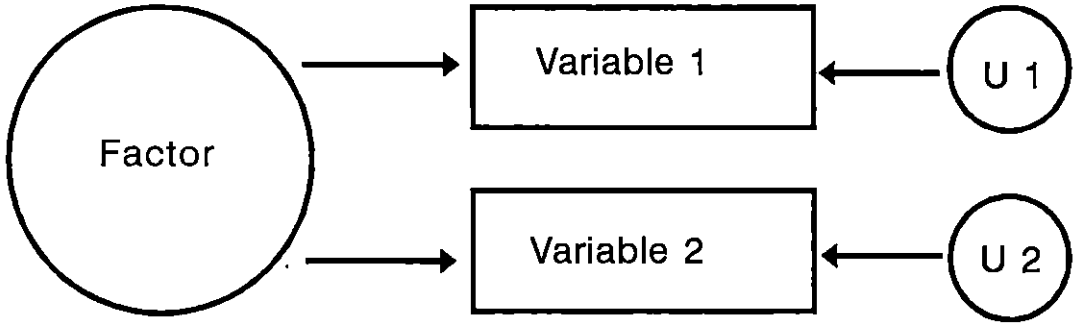
Figure 1

Causal Models

Principal Component Analysis



Common Factor Analysis



Appendix A

The solution of the eigenproblem begins by solving for the eigenvalues. To do this, one must solve the following determinental equation *p* times:

$$\left| {}_{p}\mathbf{R}_{p} - \lambda\, {}_{p}\mathbf{I}_{p} \right| = \mathbf{0}$$

where λ (the eigenvalue) is unknown. After *p* eigenvalues can are found, they can be substituted individually into the original equation:

$$({}_{p}\mathbf{R}_{p} - \lambda\, {}_{p}\mathbf{I}_{p})_{p}\mathbf{w}_{1} = \mathbf{0},$$

which must be solved to find an eigenvector for each eigenvalue.

Appendix B

The MATLAB computer programs written for this study use singular value decomposition (SVD) to extract principal components. Note that when SVD is carried out on a correlation matrix, the singular values are equal to eigenvalues. The matrix formula for SVD of the correlation matrix is:

$$_{p}\mathbf{R}_{p} = {}_{p}\mathbf{V}_{p}\mathbf{L}_{p}\mathbf{V}_{p}'$$

where **L** is a diagonal matrix containing the singular values (eigenvalues), and **V** contains the singular vectors (eigenvectors). To arrive at the component loading matrix , the square root must be taken of the **L** matrix of singular values,

$$_{p}\mathbf{R}_{p} = {}_{p}\mathbf{V}_{p}\sqrt{{}_{p}\mathbf{L}_{p}}\sqrt{{}_{p}\mathbf{L}_{p}}\mathbf{V}_{p}'$$

and multiplied each by the **V** matrix, arriving at:

$$_{p}\mathbf{R}_{p} = {}_{p}\mathbf{A}_{p}\mathbf{A}_{p}'.$$

The component scores can be found by the following equation:

$$_{n}\mathbf{Z}_{p} = {}_{n}\mathbf{X}_{p}\mathbf{A}_{p}'$$

where **Z** is the a matrix of the component scores, **X** is a matrix of the raw data in standardized form, and **A** is the component coefficient matrix. Alternatively, the component scores and the component coefficients can be found simultaneously by performing SVD on an adjusted data matrix:

$$_{n}\mathbf{X}_{p} = {}_{n}\mathbf{U}_{p}\mathbf{L}_{p}\mathbf{V}_{p}'$$

where **X** is the data matrix in *z* score form, with each number divided by the square root of (n-1), **U** is a matrix of the component scores, **L** is a matrix of the singular values (equal to the square root of the eigenvalues), and **V** is a matrix of the singular vectors

(eigenvectors). The component coefficients are found by multiplying **L*V'**.

Appendix C

Complete Output

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.8579	0.8052	0.7324	0.6571	50.0000
0.8680	0.8020	0.7397	0.6611	100.0000
0.8700	0.8070	0.7372	0.6621	200.0000
STD				
0.0316	0.0312	0.0273	0.0212	50.0000
0.0241	0.0200	0.0162	0.0134	100.0000
0.0138	0.0130	0.0117	0.0084	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0019	-0.0019	-0.0092	-0.0035	50.0000
0.0057	0.0053	-0.0004	-0.0012	100.0000
0.0009	0.0042	-0.0023	-0.0025	200.0000
STD				
0.0992	0.0822	0.0805	0.0661	50.0000
0.0600	0.0571	0.0488	0.0428	100.0000
0.0429	0.0414	0.0368	0.0301	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7947	0.6985	0.5541	0.4010	50.0000
0.7976	0.6890	0.5612	0.3989	100.0000
0.7999	0.6939	0.5668	0.3979	200.0000
STD				
0.0647	0.0557	0.0548	0.0356	50.0000
0.0492	0.0477	0.0339	0.0251	100.0000
0.0314	0.0277	0.0217	0.0163	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0155	0.0039	-0.0053	-0.0014	50.0000
0.0132	0.0014	-0.0011	0.0029	100.0000
-0.0015	0.0023	0.0021	0.0023	200.0000
STD				
0.0921	0.0800	0.0649	0.0443	50.0000
0.0596	0.0551	0.0393	0.0318	100.0000
0.0445	0.0400	0.0307	0.0228	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7040	0.5371	0.3641	0.1862	50.0000
0.6941	0.5304	0.3627	0.1806	100.0000
0.6839	0.5301	0.3567	0.1849	200.0000
STD				
0.0578	0.0477	0.0289	0.0166	50.0000
0.0417	0.0312	0.0195	0.0107	100.0000
0.0274	0.0205	0.0148	0.0066	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0036	-0.0011	0.0067	0.0064	50.0000
-0.0088	-0.0066	-0.0024	0.0038	100.0000
-0.0045	0.0038	0.0005	0.0001	200.0000
STD				
0.0965	0.0691	0.0530	0.0312	50.0000
0.0627	0.0511	0.0363	0.0194	100.0000
0.0403	0.0300	0.0255	0.0154	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	4	50
0	0	0	0	100
0	0	0	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7396	0.7012	0.6424	0.5822	50.0000
0.7500	0.7039	0.6633	0.6186	100.0000
0.7507	0.7130	0.6711	0.6237	200.0000
STD				
0.0811	0.0789	0.0878	0.1063	50.0000
0.0483	0.0489	0.0422	0.0398	100.0000
0.0318	0.0311	0.0272	0.0224	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0002	0.0039	0.0003	0.0054	50.0000
-0.0043	0.0011	-0.0034	0.0003	100.0000
-0.0001	-0.0011	0.0009	0.0001	200.0000
STD				
0.1550	0.1457	0.1458	0.1418	50.0000
0.1018	0.1074	0.0894	0.0812	100.0000
0.0712	0.0660	0.0588	0.0566	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.6069	0.5255	0.4284	0.3050	50.0000
0.6062	0.5247	0.4256	0.2985	100.0000
0.6018	0.5221	0.4268	0.3009	200.0000
STD				
0.1319	0.1132	0.0989	0.0717	50.0000
0.0993	0.0864	0.0715	0.0490	100.0000
0.0648	0.0554	0.0461	0.0320	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0089	0.0077	-0.0013	0.0026	50.0000
0.0019	0.0040	-0.0036	-0.0003	100.0000
0.0106	-0.0021	-0.0015	0.0053	200.0000
STD				
0.1314	0.1138	0.0923	0.0660	50.0000
0.0875	0.0785	0.0598	0.0424	100.0000
0.0626	0.0517	0.0426	0.0313	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.4274	0.3383	0.2150	0.1167	50.0000
0.4269	0.3260	0.2210	0.1020	100.0000
0.4120	0.3180	0.2065	0.1057	200.0000
STD				
0.1022	0.0642	0.0601	0.0328	50.0000
0.0649	0.0379	0.0330	0.0178	100.0000
0.0414	0.0304	0.0189	0.0097	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0088	-0.0141	-0.0081	-0.0054	50.0000
0.0016	0.0087	-0.0002	0.0050	100.0000
0.0088	0.0011	-0.0055	-0.0017	200.0000
STD				
0.1213	0.0765	0.0691	0.0392	50.0000
0.0802	0.0544	0.0360	0.0231	100.0000
0.0505	0.0371	0.0313	0.0130	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
5	11	10	17	50
0	2	6	9	100
0	0	0	2	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.5633	0.5357	0.4799	0.4693	50.0000
0.5956	0.5762	0.5601	0.5284	100.0000
0.6473	0.6201	0.5930	0.5778	200.0000
STD				
0.2303	0.2231	0.2364	0.2161	50.0000
0.1666	0.1648	0.1379	0.1381	100.0000
0.0884	0.0768	0.0768	0.0757	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0261	-0.0030	0.0174	-0.0017	50.0000
0.0092	-0.0076	0.0181	-0.0010	100.0000
0.0040	-0.0021	0.0046	-0.0013	200.0000
STD				
0.2643	0.2580	0.2618	0.2309	50.0000
0.2106	0.1921	0.1800	0.1851	100.0000
0.1239	0.1342	0.1372	0.1109	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
7	6	7	8	50
2	2	4	1	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.4353	0.3597	0.2874	0.2132	50.0000
0.4222	0.3345	0.2867	0.2056	100.0000
0.4022	0.3546	0.2835	0.2047	200.0000
STD				
0.2430	0.2133	0.1829	0.1105	50.0000
0.1734	0.1529	0.1226	0.0801	100.0000
0.1247	0.1136	0.0806	0.0571	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0050	0.0125	-0.0155	-0.0059	50.0000
0.0046	-0.0104	0.0042	0.0035	100.0000
0.0013	-0.0013	0.0024	0.0008	200.0000
STD				
0.2005	0.1697	0.1466	0.0987	50.0000
0.1222	0.1216	0.0992	0.0705	100.0000
0.0919	0.0766	0.0680	0.0448	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
7	9	11	6	50
2	1	1	5	100
0	0	2	1	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 9
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.2343	0.1946	0.1149	0.0638	50.0000
0.2119	0.1472	0.1000	0.0498	100.0000
0.1768	0.1429	0.0955	0.0437	200.0000
STD				
0.1782	0.1472	0.0797	0.0478	50.0000
0.1369	0.0989	0.0642	0.0245	100.0000
0.0762	0.0608	0.0474	0.0245	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0551	0.0111	0.0140	0.0072	50.0000
0.0319	0.0112	-0.0161	-0.0083	100.0000
0.0142	0.0038	0.0038	0.0084	200.0000
STD				
0.1640	0.1239	0.0783	0.0408	50.0000
0.1154	0.0880	0.0579	0.0309	100.0000
0.0828	0.0552	0.0429	0.0257	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
17	12	17	12	50
15	12	16	17	100
14	14	14	18	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.8276	0.7485	0.6524	0.5429	50.0000
0.8325	0.7507	0.6541	0.5449	100.0000
0.8340	0.7512	0.6582	0.5465	200.0000
STD				
0.0415	0.0362	0.0316	0.0257	50.0000
0.0264	0.0237	0.0211	0.0173	100.0000
0.0189	0.0160	0.0130	0.0112	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0039	-0.0090	0.0071	0.0043	50.0000
0.0017	-0.0053	0.0010	-0.0007	100.0000
0.0057	-0.0027	-0.0003	0.0009	200.0000
STD				
0.1003	0.0848	0.0725	0.0575	50.0000
0.0621	0.0565	0.0510	0.0425	100.0000
0.0455	0.0408	0.0337	0.0264	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7914	0.6905	0.5628	0.3944	50.0000
0.7933	0.6877	0.5632	0.3992	100.0000
0.8008	0.6905	0.5660	0.3987	200.0000
STD				
0.0580	0.0482	0.0424	0.0282	50.0000
0.0394	0.0332	0.0243	0.0183	100.0000
0.0250	0.0225	0.0179	0.0125	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0045	0.0083	-0.0018	0.0008	50.0000
-0.0007	0.0014	-0.0018	-0.0013	100.0000
-0.0037	0.0052	0.0006	0.0007	200.0000
STD				
0.0953	0.0807	0.0730	0.0483	50.0000
0.0637	0.0574	0.0467	0.0340	100.0000
0.0429	0.0386	0.0300	0.0215	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7715	0.6258	0.4621	0.2675	50.0000
0.7626	0.6226	0.4665	0.2682	100.0000
0.7554	0.6193	0.4644	0.2682	200.0000
STD				
0.0589	0.0383	0.0340	0.0207	50.0000
0.0351	0.0289	0.0233	0.0113	100.0000
0.0251	0.0209	0.0155	0.0089	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0030	-0.0068	-0.0021	0.0049	50.0000
-0.0027	-0.0006	0.0014	0.0020	100.0000
0.0024	0.0000	-0.0006	0.0004	200.0000
STD				
0.0989	0.0811	0.0623	0.0376	50.0000
0.0665	0.0520	0.0436	0.0275	100.0000
0.0419	0.0358	0.0284	0.0167	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.6648	0.6145	0.5506	0.4721	50.0000
0.6782	0.6212	0.5560	0.4853	100.0000
0.6810	0.6241	0.5595	0.4874	200.0000
STD				
0.0926	0.0836	0.0757	0.0668	50.0000
0.0615	0.0561	0.0488	0.0400	100.0000
0.0392	0.0364	0.0319	0.0255	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0049	-0.0051	0.0014	0.0011	50.0000
0.0065	0.0005	-0.0035	0.0026	100.0000
-0.0033	0.0014	0.0005	0.0012	200.0000
STD				
0.1415	0.1238	0.1139	0.1037	50.0000
0.0938	0.0930	0.0787	0.0641	100.0000
0.0625	0.0582	0.0505	0.0441	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.5969	0.5160	0.4237	0.2923	50.0000
0.5965	0.5070	0.4201	0.2995	100.0000
0.5986	0.5187	0.4262	0.2995	200.0000
STD				
0.1229	0.0983	0.0803	0.0603	50.0000
0.0714	0.0670	0.0507	0.0364	100.0000
0.0515	0.0429	0.0332	0.0241	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0034	-0.0087	-0.0031	-0.0012	50.0000
-0.0065	-0.0050	-0.0012	-0.0031	100.0000
-0.0016	0.0040	0.0023	-0.0027	200.0000
STD				
0.1286	0.1066	0.0919	0.0675	50.0000
0.0863	0.0774	0.0620	0.0444	100.0000
0.0566	0.0500	0.0399	0.0291	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.5403	0.4314	0.3086	0.1692	50.0000
0.5322	0.4189	0.2955	0.1659	100.0000
0.5244	0.4134	0.2980	0.1621	200.0000
STD				
0.1063	0.0737	0.0531	0.0273	50.0000
0.0612	0.0532	0.0375	0.0189	100.0000
0.0409	0.0333	0.0242	0.0126	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0080	-0.0035	-0.0077	-0.0023	50.0000
-0.0064	-0.0025	0.0030	0.0006	100.0000
-0.0009	0.0030	0.0010	-0.0015	200.0000
STD				
0.1248	0.0979	0.0710	0.0383	50.0000
0.0810	0.0635	0.0497	0.0267	100.0000
0.0488	0.0427	0.0312	0.0170	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	0	0	50
0	0	0	0	100
0	0	0	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.4331	0.4236	0.3940	0.3368	50.0000
0.5065	0.4778	0.4388	0.3993	100.0000
0.5402	0.5046	0.4674	0.4314	200.0000
STD				
0.2321	0.2033	0.1916	0.1641	50.0000
0.1399	0.1161	0.1178	0.1122	100.0000
0.0788	0.0748	0.0681	0.0610	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0138	-0.0058	0.0025	-0.0041	50.0000
-0.0043	-0.0018	-0.0007	0.0025	100.0000
-0.0019	-0.0000	0.0009	-0.0003	200.0000
STD				
0.2401	0.2114	0.1903	0.1907	50.0000
0.1616	0.1516	0.1448	0.1289	100.0000
0.0987	0.0915	0.0910	0.0809	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
2	3	4	6	50
0	2	2	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.3666	0.3409	0.2784	0.1918	50.0000
0.3948	0.3306	0.2856	0.1929	100.0000
0.3965	0.3469	0.2854	0.2008	200.0000
STD				
0.1834	0.1723	0.1406	0.1085	50.0000
0.1473	0.1156	0.0875	0.0620	100.0000
0.0871	0.0730	0.0545	0.0425	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0059	0.0078	0.0063	-0.0013	50.0000
0.0022	-0.0043	-0.0030	-0.0026	100.0000
-0.0011	-0.0041	-0.0004	0.0014	200.0000
STD				
0.2028	0.1622	0.1324	0.0894	50.0000
0.1241	0.1129	0.0861	0.0646	100.0000
0.0790	0.0665	0.0545	0.0387	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
7	2	2	4	50
1	2	1	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 18
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.3307	0.2507	0.1896	0.0908	50.0000
0.3103	0.2478	0.1641	0.0849	100.0000
0.2905	0.2196	0.1561	0.0836	200.0000
STD				
0.1673	0.1207	0.0762	0.0430	50.0000
0.1052	0.0733	0.0505	0.0244	100.0000
0.0565	0.0432	0.0281	0.0152	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0019	0.0012	0.0051	0.0026	50.0000
-0.0020	-0.0007	0.0002	0.0015	100.0000
-0.0033	-0.0013	-0.0007	-0.0030	200.0000
STD				
0.1664	0.1293	0.0814	0.0453	50.0000
0.1030	0.0861	0.0583	0.0269	100.0000
0.0627	0.0481	0.0320	0.0165	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
8	7	11	8	50
4	3	6	4	100
0	0	1	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.8066	0.7184	0.5627	0.3645	50.0000
0.8119	0.7197	0.6128	0.4772	100.0000
0.8170	0.7223	0.6118	0.4780	200.0000
STD				
0.0512	0.0660	0.1778	0.3021	50.0000
0.0321	0.0287	0.0215	0.0175	100.0000
0.0203	0.0186	0.0154	0.0106	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0114	0.0019	0.0064	-0.0159	50.0000
-0.0064	-0.0036	0.0033	-0.0017	100.0000
0.0060	0.0025	-0.0007	-0.0010	200.0000
STD				
0.0995	0.1161	0.1970	0.2683	50.0000
0.0697	0.0575	0.0462	0.0362	100.0000
0.0445	0.0365	0.0315	0.0245	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	3	10	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7986	0.6837	0.5296	0.2930	50.0000
0.7966	0.6893	0.5627	0.3994	100.0000
0.7962	0.6905	0.5643	0.3994	200.0000
STD				
0.0536	0.0722	0.1861	0.2636	50.0000
0.0359	0.0313	0.0249	0.0177	100.0000
0.0239	0.0193	0.0169	0.0117	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0119	-0.0090	-0.0094	-0.0016	50.0000
0.0018	0.0048	0.0020	-0.0019	100.0000
-0.0043	-0.0038	-0.0016	0.0014	200.0000
STD				
0.0945	0.1162	0.1840	0.2580	50.0000
0.0669	0.0542	0.0460	0.0318	100.0000
0.0480	0.0388	0.0318	0.0229	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	6	13	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000.
Number variables = 36
Population loading = 0.8

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.7997	0.6675	0.4644	0.2871	50.0000
0.7809	0.6610	0.5140	0.3230	100.0000
0.7851	0.6569	0.5138	0.3258	200.0000
STD				
0.0537	0.0883	0.1910	0.2865	50.0000
0.0363	0.0300	0.0239	0.0154	100.0000
0.0216	0.0208	0.0154	0.0105	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0013	-0.0029	-0.0166	0.0228	50.0000
0.0081	-0.0016	0.0014	-0.0022	100.0000
-0.0034	0.0013	-0.0006	-0.0002	200.0000
STD				
0.0983	0.1297	0.2157	0.2798	50.0000
0.0639	0.0585	0.0448	0.0294	100.0000
0.0418	0.0375	0.0291	0.0186	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	1	14	14	50
0	0	0	0	100
0	0	0	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.6315	0.5497	0.4135	0.2479	50.0000
0.6400	0.5727	0.4933	0.4003	100.0000
0.6384	0.5753	0.4961	0.4037	200.0000
STD				
0.0921	0.1143	0.2229	0.3192	50.0000
0.0595	0.0509	0.0460	0.0375	100.0000
0.0419	0.0339	0.0306	0.0240	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0131	-0.0074	0.0037	-0.0246	50.0000
0.0046	0.0024	-0.0029	0.0017	100.0000
-0.0043	0.0019	0.0021	0.0001	200.0000
STD				
0.1328	0.1478	0.2227	0.2749	50.0000
0.0851	0.0779	0.0663	0.0564	100.0000
0.0610	0.0500	0.0458	0.0352	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	1	3	14	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.5911	0.5053	0.3710	0.1871	50.0000
0.5971	0.5130	0.4222	0.2979	100.0000
0.5997	0.5238	0.4254	0.3008	200.0000
STD				
0.0959	0.1132	0.2174	0.2593	50.0000
0.0651	0.0583	0.0461	0.0344	100.0000
0.0436	0.0386	0.0332	0.0235	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0051	-0.0053	0.0037	0.0082	50.0000
-0.0075	0.0037	-0.0053	0.0001	100.0000
0.0029	-0.0007	-0.0006	-0.0004	200.0000
STD				
0.1227	0.1376	0.1834	0.2629	50.0000
0.0870	0.0743	0.0615	0.0426	100.0000
0.0568	0.0460	0.0383	0.0271	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	0	5	15	50
0	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.6

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.6047	0.4990	0.3576	0.2283	50.0000
0.5801	0.4778	0.3616	0.2176	100.0000
0.5632	0.4708	0.3608	0.2170	200.0000
STD				
0.1043	0.1237	0.2020	0.2467	50.0000
0.0646	0.0532	0.0405	0.0234	100.0000
0.0410	0.0330	0.0255	0.0155	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0010	-0.0051	0.0334	0.0088	50.0000
-0.0041	-0.0019	-0.0003	-0.0010	100.0000
-0.0008	-0.0011	0.0010	0.0007	200.0000
STD				
0.1289	0.1538	0.1953	0.2387	50.0000
0.0820	0.0674	0.0504	0.0311	100.0000
0.0525	0.0436	0.0342	0.0211	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	3	14	18	50
0	0	0	0	100
0	0	0	0	200

Principal Component Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.4340	0.3600	0.2722	0.1991	50.0000
0.4645	0.4166	0.3835	0.3268	100.0000
0.4712	0.4294	0.3872	0.3365	200.0000
STD				
0.1610	0.1925	0.2562	0.3230	50.0000
0.1028	0.0995	0.0802	0.0756	100.0000
0.0691	0.0610	0.0547	0.0472	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0069	-0.0007	0.0055	-0.0025	50.0000
0.0047	-0.0004	-0.0046	0.0015	100.0000
-0.0022	-0.0025	-0.0004	-0.0009	200.0000
STD				
0.1852	0.2068	0.2457	0.2678	50.0000
0.1204	0.1112	0.0984	0.0853	100.0000
0.0802	0.0749	0.0683	0.0588	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	4	13	17	50
0	0	0	0	100
0	0	0	0	200

Principal Factors Factor Analysis - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.3818	0.3131	0.2181	0.1138	50.0000
0.3873	0.3463	0.2785	0.1985	100.0000
0.4010	0.3436	0.2802	0.1997	200.0000
STD				
0.1597	0.1910	0.2312	0.3087	50.0000
0.1025	0.0833	0.0745	0.0499	100.0000
0.0703	0.0574	0.0486	0.0333	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
-0.0090	-0.0195	0.0188	0.0041	50.0000
0.0030	0.0058	-0.0010	-0.0031	100.0000
-0.0013	-0.0013	0.0002	-0.0017	200.0000
STD				
0.1602	0.1736	0.2248	0.2525	50.0000
0.1117	0.0906	0.0756	0.0558	100.0000
0.0677	0.0598	0.0490	0.0349	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
2	4	13	17	50
1	0	0	0	100
0	0	0	0	200

Image Analysis on G - iterations = 20
Population size = 1000
Number variables = 36
Population loading = 0.4

Non-zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.3792	0.2819	0.2060	0.1226	50.0000
0.3813	0.2943	0.2088	0.1157	100.0000
0.3563	0.2853	0.2008	0.1117	200.0000
STD				
0.1718	0.1928	0.2189	0.2455	50.0000
0.0925	0.0730	0.0543	0.0287	100.0000
0.0547	0.0440	0.0329	0.0180	200.0000

Zero loadings

% Measurement Error

0	25	50	75	S. SIZE
0.0089	0.0076	0.0227	-0.0337	50.0000
0.0008	0.0008	-0.0043	0.0004	100.0000
-0.0005	-0.0003	-0.0020	0.0025	200.0000
STD				
0.1824	0.1954	0.2189	0.2886	50.0000
0.0998	0.0807	0.0579	0.0326	100.0000
0.0617	0.0495	0.0351	0.0207	200.0000

Number of Trials with
Problems Distinguishing Factors

% Measurement Error

0	25	50	75	S. SIZE
0	13	18	19	50
0	0	0	0	100
0	0	0	0	200

Appendix D

MATLAB Source Code

```
%
% main program for performing PCA, PFA, and IA
% on samples taken from a population with variance due to the
% measured and unmeasured factors.
% Measurement error is added to the sample,
% analyses are performed, and output is recorded in a diary
%
% Written by Todd S. Woodward, Department of Psychology
% The University of Victoria, Victoria, B.C., CANADA
%
tog=menu('Do you want to clear all previous variables?','YES','NO');
if tog==1
  clear
end

pack
for analysis=1:3
 nummonte=20;   % number of samples to compute
 popsize=1000;  % determining size of the population
 var=36;        % number of variables
 popald=.8;     % population loading
 numfactors=3;  % number of factors
 numsamples=3;  %number of different sample sizes
 numerrors=4;   % number of different error sizes
 errinc=25;     % percent error increase per increment
 sub=50;        % starting number of subjects (doubled each iteration)
 OUTMATNZ=[];
 OUTMATZERO=[];
 OUTMATPROBS=[];
 for samplesiz=1:numsamples
  sub
  errov=0;    % starting amount of error
  for errorsiz=1:numerrors
   errov
   Probsf=0;
   numgood=0;
   numprobs=0;
   cl=clock;
                %
   rand('seed',cl(6)); % setting random number generator seed to current second
                %

   %%%%%%%%%%%%%%%%%%%%% Creating Population


                          %
   [APOP FPOP numvarclust]=varad1(var,popsize,numfactors);  % getting A & F
                                  % for pop DATA making
                          % numvarclust is the number
                          % of variables per cluster
```

```
[DATASPEC ORGTRUE enderr]=adspecer(APOP,FPOP); % adding specific error
                                % to data using A & F


%%%%%%%%%%%%%%%%%%%%% Taking Samples and Analyzing

TOTMAT=[];
totnzmnld=[];
totnzstdld=[];
totzeromnld=[];
totzerostdld=[];
for count=1:nummonte
 count

 [RSAMP]=randsamp(DATASPEC,sub);            % Taking random sample
 [RSAMPMEAS,finerr]=admeserz(RSAMP,errov);      % adding measurement error
 if analysis==1 % do PCA
  [A F]=svdfac(RSAMPMEAS,numfactors);       % performing PCA on cor mat
  ANTYPE=(['Principal Component Analysis - iterations = ',num2str(nummonte),'']);
 elseif analysis==2 % do PFFA
  [A CommVec]=pfsvdfac(RSAMPMEAS,numfactors); % performing PFFA analysis
  ANTYPE=(['Principal Factors Factor Analysis - iterations = ',num2str(nummonte),'']);
 elseif analysis==3  % do image if analysis = 3
  [GSTAR,S]=gstar(RSAMPMEAS); % get rescaled image variance-covariance matrix
                        % rescaled by the inverse of the
                        % square root of the anti-image variances
                        % and by a matrix of the square root of the
                        % anti-image variances (1-R2)
  [LSQURTD]=svdcovim(GSTAR);          % performing image analysis
                              % get loadings from SVD of G*
  APREBIG=S*LSQURTD;          % rescaling loadings by square root of
                        % the anti-image variances (1-R2)

  A=APREBIG(:,1:numfactors);   % scaling down for the number of factors
  ANTYPE=(['Image Analysis on G - iterations = ',num2str(nummonte);'']);
 end   % analysis choice

 [AFIN]=varimkn(A);                    % performing varimax rotation
 [AORDMAT,Probs]=ordbysum(AFIN,APOP,numvarclust);
 COMPMAT=[AORDMAT  APOP];
 if Probs > 0
  Probsf=Probsf+Probs;  % record that factors could not be properly ordered
  numprobs=numprobs+1;
 else
  [mnnzld,stdnzld,mnzerold,stdzerold]=mnloads(AORDMAT,numvarclust);
  totnzmnld=[totnzmnld;mnnzld];
  totnzstdld=[totnzstdld;stdnzld];
  totzeromnld=[totzeromnld;mnzerold];
  totzerostdld=[totzerostdld;stdzerold];
  TOTMAT=[TOTMAT COMPMAT];
  numgood=numgood+1;
 end   % problems loop
```

```
    end   % monte carlo loop
    meannz=mean(totnzmnld);
    stdnz=mean(totnzstdld);
    meanzero=mean(totzeromnld);
    stdzero=mean(totzerostdld);
    diff=meannz-APOP(1,1);
    diary off
    disp(ANTYPE)
    disp('')
    disp('    Count  Problems     Good    Meannz    Stdnz  Meanzero   Stdzero')
    disp('    ----------------------------------------------------------------')
    d=[count numprobs numgood meannz stdnz meanzero stdzero];
    disp(d)
    disp('')
    disp('     Diff  Specific  Meas.Err. Subjects')
    disp('     ---------------------------------')
    d=[diff enderr finerr sub];
    disp(d)
    disp('')
    OUTMATNZ(samplesiz,errorsiz)=meannz;
    OUTMATZERO(samplesiz,errorsiz)=meanzero;
    OUTMATNZSD(samplesiz,errorsiz)=stdnz;
    OUTMATZEROSD(samplesiz,errorsiz)=stdzero;
    OUTMATPROBS(samplesiz,errorsiz)=numprobs;
    errov=errov+errinc;  % incrementing error
  end   % error size loop
  OUTMATNZ(samplesiz,numerrors+1)=sub;
  OUTMATZERO(samplesiz,numerrors+1)=sub;
  OUTMATNZSD(samplesiz,numerrors+1)=sub;
  OUTMATZEROSD(samplesiz,numerrors+1)=sub;
  OUTMATPROBS(samplesiz,numerrors+1)=sub;
  sub=sub*2; % incrementing sample size
end   % sample size loop
diary on
%
% beginning output writing to diary
%
disp('')
disp(ANTYPE)
di=(['Population size = ',num2str(popsize),'']);
disp(di)
dsp=(['Number variables = ',num2str(var),'']);
disp(dsp)
dld=(['Population loading = ',num2str(popald),'']);
disp(dld)
disp('')
disp('')
disp('                Non-zero loadings')
disp('')
disp('         Measurement Error              ')
disp('')
d=(['          ',num2str(0),'       ',num2str(25),'       ',num2str(50),'       ',num2str(75),'| S. SIZE']);
disp(d)
```

```
disp('-------------------------------------------------')
disp('')
disp(OUTMATNZ)
disp('          STD')
disp(OUTMATNZSD)
disp('')
disp('                  Zero loadings')
disp('')
disp('          Measurement Error                ')
disp('')
d=(['          ',num2str(0),'       ',num2str(25),'       ',num2str(50),'       ',num2str(75),'|  S. SIZE']);
disp(d)
disp('-------------------------------------------------')
disp('')
disp(OUTMATZERO)
disp('          STD')
disp(OUTMATZEROSD)
disp('')
disp('')
disp('        Number of Trials with')
disp('   Problems Distinguishing Factors')
disp('')
disp('            Measurement Error                ')
disp('')
d=(['    ',num2str(0),'    ',num2str(25),'    ',num2str(50),'    ',num2str(75),'|  S. SIZE']);
disp(d)
disp('---------------------------------')
disp('')
disp(OUTMATPROBS)
diary off
save lastrun
clear
end  % analysis loop

================== ==========================================
==

%
% file to get true A and F
% function[ATRUE,FTRUE,numvarclust]=varad1(v,s,f)
%
function[ATRUE,FTRUE,numvarclust]=varad1(v,s,f)

FP=rand(s,1);
for count=1:(f-1)
  FUNCORR=fnderrvc([FP  (rand(s,1))],count,s);
  FP=[FP FUNCORR];
end
FPRE=FP;


APRE=[.8 0 0;.8 0 0;.8 0 0;
    0 .8 0;0 .8 0;0 .8 0;
```

```
    0 0 .8;0 0 .8;0 0 .8];
numvarclust=3;

%APRE=[.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;
%      0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;
%      0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8];
%numvarclust=6;

%APRE=[.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0;.8 0 0
%      0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;0 .8 0;
%      0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8;0 0 .8];
%numvarclust=12; % number of variables per cluster


ATRUE=APRE;
FTRUE=zscore(FPRE);

=================== ============================== ==
==

%
% file to make dataset with error for factor analysis
% limitation - only works if all loadings are the same
% function [DATA,TRUE,deferr]=adspecer(APOP,FPOP)
function [DATA,TRUE,deferr]=adspecer(A,F)
uninor;
TRUE = (A * F')';
[subs vars]=size(TRUE);
ERMAT=[];
for count=1:vars  % ensuring each error vector is uncorrelated to the others
              % and to the true part of the X's
                                    %
   [ERRVC]=fndfervc(TRUE,ERMAT,subs,vars);  % getting vector of
                                    % Y - YHAT
   ERMAT=[ERMAT ERRVC];
end
erromlt=(sqrt(1.0-(A(1,1).^2)));  % This formula is used by Snook and Gorsuch
                          % to find a number to weight the specific error matrix
                          % by to create Zscores when added to the FCM
                          % hence the 1.0 (variance) in the equation

ZERR=zscore(ERMAT);
ERROR=ZERR*(diag(ones(vars,1).*erromlt));  % multiplying each column by
                                  % multiplier to make desired
                                  % variance per column

DATA=TRUE+ERROR;   % adding error variance to true part to make desired
              % percentage of error per variable

varerr=mean(variance(ERROR));
vartrue=mean(variance(TRUE));

deferr=varerr/(vartrue+varerr);
```

```
==========
%
% File to find Error vector to use as uncorrelated error
% ERRVC will be uncorrelated to IVIN and ERMAT
% function[ERRVC]=fndfervc(IVIN,ERMAT,sub,var)
%
function[ERRVC]=fndfervc(IVIN,ERMAT,sub,var)
[m n]=size(ERMAT);
uninor
DVDAT = rand(sub,1);
if m==0
  IVBEG=IVIN;
else
  IVBEG=[IVIN ERMAT];
end
IVDAT = [IVBEG (ones(sub,1))];
BETCONST = (pinv(IVDAT'*IVDAT))*(IVDAT'*DVDAT);   % finding weights and constant
YHAT = IVDAT * BETCONST;
ERRVC=DVDAT-YHAT;

==========
% function file to perform conversion to Z scores
function [Z] = zscore(X)
[m,n] = size(X);
ST = std(X);
MN = mean(X);
for j=1:n
    for i=1:m
      Z(i,j)=(X(i,j)-MN(1,j))/ST(1,j);
    end
end

==========
%
% file to change random distribution to normal if it is uniform
% 'uninor.m'
%
presdist=rand('dist');
distval=strcmp(presdist,'uniform');
if distval==1
  rand('normal');
end

==========
```

```
%
% file to take a random sample of scores (by row) from the population
% function[RSAMP]=randsamp(X,sizesamp);
%
function[RSAMP]=randsamp(X,sizesamp);
[m,n]=size(X);
alreadypicked=[];
noruni;
q=rand(1,1)*m;
r=ceil(q);
for k=1:sizesamp
  while length(find(alreadypicked==r))~=0 % checks if r has already been picked
                                  % find returns a vector position if
                                  % the number r is in the vector alreadypicked
     q=rand(1,1)*m; % generate new random number
     r=ceil(q);
  end
  alreadypicked=[alreadypicked;r]; % add q to already picked vector
  RSAMP(k,:)=X(r,:);
end
rand('normal');
uninor
==========================================================
==

%
% file to change random distribution to uniform if it is normal
% 'noruni.m'
%
presdist=rand('dist');
distval=strcmp(presdist,'normal');
if distval==1
  rand('uniform');
end

==========================================================
==


%
% file to add a specified amount of measurement error
% to dataset with only specific error already added
% and to end up with standardized variables
% function [DATA,enderr,ERROR,TRUENOR]=admeserz(PREDATA,errov)
%
function [DATA,enderr,ERROR,TRUENOR]=admeserz(PREDATA,errov)
[subs vars]=size(PREDATA);
errper=errov/100; % converting error desired to a percentage
datmult = sqrt(1.0-errper); % finding number to reduce variance of data by so
                    % that error, when added, will give variance of 1
TRUENOR=((zscore(PREDATA)).*datmult); % reducing variance

ERMAT=[];
```

```
for count=1:vars  % ensuring each error vector is uncorrelated to the others
            % and to the true part of the X's
   [ERRVC]=fndfervc(TRUENOR,ERMAT,subs,vars);  % getting vector of
                                    % Y - YHAT
   ERMAT=[ERMAT ERRVC];
end
erromlt=(sqrt(1.0-(datmult.^2)));  % This formula is used by Snook and Gorsuch
                        % to find a number to weight the error matrix
                        % by to make Zscores when added to true part
                        % hence the 1.0 (variance) in the equation

ZERR=zscore(ERMAT);
ERROR=ZERR*(diag(ones(vars,1).*erromlt));  % multiplying each column by
                                % multiplier to make desired
                                % variance per column

DATA=TRUENOR+ERROR;   % adding error variance to true part to make desired
                % percentage of error per variable

varerr=mean(variance(ERROR));
vartrue=mean(variance(TRUENOR));

enderr=varerr/(vartrue+varerr);

================================================================
==

%
% file to perform principal component analysis by singular value decomposition
% f is number of factors, X is raw data. Decomposes corr. Mat.
% function[A1ST,F1ST]=svdfac(X,f)
%
function[A1ST,F1ST]=svdfac(X,f)
CX=corrcoef(X);
[m,n]=size(CX);
[U,S,V]=svd(CX,0);
%
% scaling down for number of factors
%
U=U(:,1:f);
V=V(:,1:f);
S=S(1:f,1:f);
%
% Returning scale to the loading matrix
%
A1ST=U*sqrt(S);
%
% transforming factor scores
%
ZX=zscore(X);
Fp=pinv(A1ST)*ZX';
F1ST=Fp';
```

```
==========
==

%
% file for performing principal factors factor analysis
% by estimating communalities and the performing SVD on
% correlation matrix with adjusted communalities.
% function [A1ST,F1ST,CommVec]=pfsvdfac(X,f)
%
function[A1ST,CommVec]=pfsvdfac(X,f)
PRECX=corrcoef(X);
[CX,CommVec]=pfdat(PRECX,X,f);
[m,n]=size(CX);
[U,S,V]=svd(CX,0);
%
% scaling down for number of factors
%
U=U(:,1:f);
V=V(:,1:f);
S=S(1:f,1:f);
A1N=U;
A2N=V;
%
% Returning scale to the loading matrix
%
A1ST=A1N*sqrt(S);

==========
==

%
% file for making data set for principal factors factor analysis
% by estimating communalities
% function [PFCMAT,CommVec]=pfdat(CMAT,DATA,numf)
%
function [PFCMAT,CommVec]=pfdat(CMAT,DATA,numf);
[subs vars]=size(DATA);
NWCMAT=CMAT;
for rsqui=1:vars  % putting Rsquareds on the diagonals of the correlation matrix
  DV=DATA(:,rsqui);
  if rsqui==1
    IV=DATA(:,2:vars);
  else
    IV=[DATA(:,1:rsqui-1) DATA(:,rsqui+1:vars)];
  end
  FINDAT=[IV DV];
  [dum1 dum2 Rsqu]=perfreg(FINDAT,(vars-1),subs);
  NWCMAT(rsqui,rsqui)=Rsqu;
end
for iters=1:7  % -- number of iterations to estimate communalities
  [A F]=svdcorpf(NWCMAT,numf); % performing PC analysis
  ACOM=(A.^2)';
  Coms=sum(ACOM);  % calculating communalities
```

```
 for cmcnt=1:vars
   NWCMAT(cmcnt,cmcnt)=Coms(cmcnt);  %placing communalities on the diagonal
 end;
end
CommVec=Coms;
PFCMAT=NWCMAT;


==========
==

%
% File to find BETA's, percentage of error, and RSquared
% function[BETCONST,IVBEG,Rsqu,MSE,IVBEG,DVDAT,YHAT]=perfreg(FINDAT,var,sub)
%
function[BETCONST,IVBEG,Rsqu,DVDAT,YHAT]=perfreg(FINDAT,var,sub)
IVBEG = FINDAT(:,(1:var));
IVDAT = [IVBEG (ones(sub,1))];
DVDAT = FINDAT(:,(var+1));
BETCONST = (pinv(IVDAT'*IVDAT))*(IVDAT'*DVDAT);  % finding weights and constant
YHAT = IVDAT * BETCONST;
RNSQ = corrcoef([YHAT DVDAT]);
Rsqu=(RNSQ(2,1)).^2;
NMT=((DVDAT-YHAT).^2);
NMTSTERR=sum(NMT);

==========
==

%
% file to perform factor analysis by singular value decomposition with
% the correlation matrix as input
% f is number of factors
% function[A1ST,F1ST]=svdcorpf(CX,f)
%
function[A1ST,F1ST]=svdcorpf(CX,f)
[m,n]=size(CX);
[U,S,V]=svd(CX,0);
%
% scaling down for number of factors
%
U=U(:,1:f);
V=V(:,1:f);
S=S(1:f,1:f);
A1N=U;
A2N=V;
%
% Returning scale to the loading matrix
%
A1ST=A1N*sqrt(S);

==========
==
```

```
%
% File to adjust correlation matrix for use in image analysis
% input is real data, adjustment made to CORR matrix to get
% image covariance matrix
% input is raw data
% function[GSTAR,S,G,SANTSQU]=gstar(DATA)
%
function[GSTAR,S,G,SANTSQU]=gstar(DATA)
[subs vars]=size(DATA);
CORRDAT=corrcoef(DATA);
COL=diag(diag(inv(CORRDAT)));   % [diagR-1]
SANTSQU=inv(COL);               % inv ^    , the anti-image variance, 1-R2
S=sqrt(SANTSQU);
COVANTIS=(SANTSQU*(pinv(CORRDAT))*SANTSQU);  %  S2*R-1*S2
G=CORRDAT + COVANTIS - 2*(SANTSQU);          % G = R + ^ -2*S2
GSTAR=pinv(S)*G*pinv(S);


================================================
==

%
% file to perform factor analysis by singular value decomposition
% for image analysis, covariance for input
% function[LSQURTD,EIGS]=svdcovim(X)
%
function[LSQURTD,EIGS]=svdcovim(COVX)
[m,n]=size(COVX);
[U,S,V]=svd(COVX,0);
EIGS=diag(S);
%
% Returning scale to the loading matrix
%
LSQURTD=U*sqrt(S);

================================================
==
%
% file to perform varimax rotation with Kaiser normalization
%
% function[AFIN]=varimkn(INPMAT)
function[AFIN]=varimkn(INPMAT)
SUMI=sum((INPMAT.^2)');
SUMINP=sqrt(SUMI);
[m n]=size(INPMAT);
for count1=1:m
   for count2=1:n
      KAISNOR(count1,count2)=INPMAT(count1,count2)./SUMINP(count1);
   end;
end;
VARK=varim(KAISNOR);
for count1=1:m
   for count2=1:n
```

```
      AFIN(count1,count2)=VARK(count1,count2).*SUMINP(count1);
   end;
end;

==========================================================
==

function [B,T]=varim(A);
% produces varimax rotated version of A and rotation matrix T
% Based on nevels 1986; see also pascal source
% NOTE - received from INTERNET - T.W.
% input: A    (matrix to be rotated)
% output: B (rotated A)
%     T (rotation matrix)
%     f (varimax function)


conv=.000001;

[m,r]=size(A);
T=eye(r);
B=A;

f=ssq((A.*A)-ones(m,1)*mean(A.*A));
fold=f-2*conv*f;
iter=0;

while f-fold>f*conv
  fold=f;iter=iter+1;
  for i=1:r
    for j=i+1:r
      x=B(:,i);y=B(:,j);
      xx=T(:,i);yy=T(:,j);
      u=x.^2-y.^2;v=2*x.*y;
      u=u-ones(m,1)*mean(u);v=v-ones(m,1)*mean(v);
      a=2*r*sum(u.*v);b=r*sum(u.^2)-r*sum(v.^2);c=(a^2+b^2)^.5;
      if a>=0; sign=1; end;
      if a<0; sign=-1; end;
      if c<.00000000001
        disp(' No rotation anymore');
        cos=1;sin=0;
      end;
      if c>=.00000000001
        vvv=-sign*((b+c)/(2*c))^.5;
        sin=(.5-.5*vvv)^.5;cos=(.5+.5*vvv)^.5;
      end;
      v=cos*x-sin*y;w=cos*y+sin*x;
      vv=cos*xx-sin*yy;ww=cos*yy+sin*xx;
      if vvv>=0       % prevent permutation of columns
        B(:,i)=v;B(:,j)=w;T(:,i)=vv;T(:,j)=ww;
      end;
      if vvv<0
        B(:,j)=v;B(:,i)=w;T(:,j)=vv;T(:,i)=ww;
```

```
    end;
   end;
  end;
  f=ssq((B.*B)-ones(m,1)*mean(B.*B));
end;

==========

%
% file to match loadings to a reference set where the sum of
% three loading is highest
%  [ORDMATRIX]=ordbysum(UNORDMATRIX,REFMATRIX,numvarclust)
%
function [ORDMATRIX,Probs]=ordbysum(UNORDMATRIX,REFMATRIX,numvarclust)
[m n]=size(REFMATRIX);
[o p]=size(UNORDMATRIX);
numclust=3; % assigning number of clusters per factor
% numvarclust=12; % assigning number of variables per loading cluster
WORKUNORD=UNORDMATRIX;
ABSUNORD=abs(UNORDMATRIX);
Listclusinds=[];
Probs=0;
for count1=1:p % number of factors
  [garb bigfacind]=max(sum(ABSUNORD)); % finding largest column of loadings
  index=0;
  Totclust=[];
  for count3=1:numclust % number clusters per factor
    sumclust=0;
    for count2=1:numvarclust % number of variables per loading cluster
      sumclust=sumclust+(WORKUNORD((count2+index),bigfacind)); % sum of cluster for that column
    end
    Totclust=[Totclust;sumclust];

  index=index+numvarclust;
  end
  ABSUNORD(:,bigfacind)=zeros(ABSUNORD(:,bigfacind)); %change largest last row to
                                        % zeros so it is ignored
                                        % next time around
  [garb bigclustind]=max(abs(Totclust)); % although calculated on clusters,
                                % 1 is 1st factor, 2 2nd etc..
  if (length(find(Listclusinds==bigclustind))~=0) % if index already used
    disp('')
    disp('Problem ordering factors')
    Probs=Probs+1;
    disp('')
    if (length(find(Listclusinds==1))==0) % if 1 not used
      bigclustind=1;
    elseif (length(find(Listclusinds==2))==0) % if 2 not used
      bigclustind=2;
    else
      bigclustind=3;
    end
```

```
 end

 Listclusinds=[Listclusinds bigclustind];

 ORDMATRIX(:,bigclustind)=UNORDMATRIX(:,bigfacind); % assigning factor to
                                        % population partner
end

==========

%
% File to find mean loading for a number of samples
% must set number of variables to load on each factor
%
% function[mnnzld,stdzld,mnzerold,stdzerold]=mnloads(AORD,numvarclust);
%
function[mnnzld,stdzld,mnzerold,stdzerold]=mnloads(AORD,numvarclust);
[vars facs]=size(AORD);
FXEDMAT=fixsigns(AORD,numvarclust);
indic=0;
sumnznum=[];
ZLDSMAT=ones((vars-numvarclust),facs); % beginning with ones for Nzeros
for factornum=1:facs;
   for countnzlds=1:numvarclust
      nznum=FXEDMAT((indic+countnzlds),factornum);
      sumnznum=[sumnznum;nznum];
   end

ZLDSMAT(:,factornum)=([FXEDMAT(1:indic,factornum);FXEDMAT(indic+numvarclust+1:vars,factorn
um)]);
                        % assigning nonzero loading to their own matrix
   indic=indic+numvarclust;
end
sumzeronum=ZLDSMAT(:);
mnnzld=mean(sumnznum);
stdzld=std(sumnznum);
mnzerold=mean(sumzeronum);
stdzerold=std(sumzeronum);

==========

%
% file to give correct signs to loadings
% INPUT is unfixed loadings
% numvarclust is the number of loadings that cluster per factor
% function[FIXED]=fixsigns(AORD,numvarclust)
%
function[FIXED]=fixsigns(AORD,numvarclust)
[vars facs]=size(AORD);
FIXED=AORD;
indic=0;
```

```
for factornum=1:facs;
   sumnum=[];
   for countnzlds=1:numvarclust
      num=AORD((indic+countnzlds),factornum);
      sumnum=[sumnum;num];
   end
   if (sum(sumnum)) < 0
      FIXED(:,factornum)=(FIXED(:,factornum)*(-1)); % switching sign if
                                                    % targets are negative
   end
   indic=indic+numvarclust;
end
```

VITA

Surname: Woodward Given Names: Todd Stephen

Place of Birth: Saskatoon, Saskatchewan Date of Birth:June 17, 1967

Educational Institutions Attended:

University of Victoria	1992 to 1993
University of Western Ontario	1991 to 1992
University of Victoria	1985 to 1989

Degrees Awarded:

B.Sc. (Honours)	University of Victoria	1989

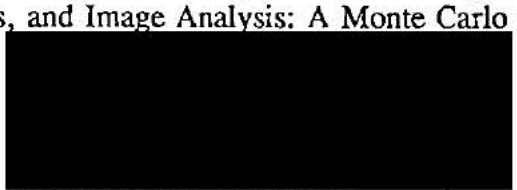
Honours and Awards:

Natural Sciences and Engineering Research Council of Canada Post Graduate Scholarship	1991 to 1993
University of Western Ontario Entrance Scholarship	1991
President's Scholarship for Part-Time Students	1988

Partial Copyright License

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis: A Comparison of Principal Component Analysis, Common Factor Analysis, and Image Analysis: A Monte Carlo Study.

Author: 

TODD S. WOODWARD

30 September, 1993