Quality and Quantity, 14 (1980) 699–710
Elsevier Scientific Publishing Company, Amsterdam – Printed in the Netherlands

STANDARDIZED VERSUS UNSTANDARDIZED DATA MATRICES: WHICH TYPE IS MORE APPROPRIATE FOR FACTOR ANALYSIS

JOHN C. WOELFEL

Market Dynamics, 101 Birch Street, Falls Church, Virginia 2204, U.S.A.

JOSEPH D. WOELFEL and MARY LOU WOELFEL

State University of New York at Albany, U.S.A.

The issue of whether to use standardized or unstandardized coefficients in regression analysis has been the focus of several papers (Blalock, 1964, 1967; Tukey, 1954; Turner and Stevens, 1959; Wright, 1960). Blalock (1967) suggests that when comparing coefficients across sub-populations or samples, unstandardized regression coefficients are appropriate. His rationale is that the standard deviation of a variable can vary from sample to sample. If this occurs, the standardized form of a variable will also be different across samples. Hence, observed differences in standardized regression coefficients for a variable across samples may be a function of differences in the standard deviation and not real differences in the true value of the coefficient. However, when comparing coefficients of variables measured on different scales within a sub-population or sample, Blalock recommends the use of standardized regression coefficients. The argument here is that standardization transforms each variable to a comparable level of measurement.

While the sociological literature has treated the standardization question with regard to regression analysis, it has not considered as carefully the question in terms of factor analysis. This technique is being used with increasing frequency by sociologists, particularly as a method for scaling variables. In lieu of any guidance from the sociological literature, it might be assumed that the case for factor analysis is analogous to that for regression analysis, and consequently, that the admonishments of Blalock for regression are applicable to factor analysis. However, this would be a faulty assumption. The utilization of standardized variables for factor analysis (either within or across samples) has undesirable properties. It is important that the pitfalls of using standardized variables for factor analysis be addressed, especially since the majority of factor analyses are performed on standardized variables in the form of a correlation matrix (Horst, 1965; Nie et al., 1975).

The purpose of this paper will be twofold. First, it will examine the drawbacks inherent in factor-analyzing standardized variables. This will be done with a mathematical proof and an example from empirical data. Second, the paper will suggest a factor analytic technique which avoids the problems associated with standardized variables.

Factor analysis is essentially placing a set of reference coordinates upon a set of variables and measuring the projection (loading) of each variable on the coordinates. In factor analysis it is helpful to consider each variable as a vector and the entire set of variables as a vector space. In order to factor-analyze a vector space there are two major requirements; the vectors must share a common origin and the length of each vector (its communality) must be known. In the past researchers have been inclined to standardize. The common origin is ordinarily provided by expressing the raw variables as deviation scores by subtracting the variable mean from each variable value. The next step in the analysis is then to standardize these deviation scores. However, this can only be accomplished through a non-linear transformation since each vector is divided by a different value, its own standard deviation [1]. This non-linear transformation is simultaneously effected on each vector length with the vector length (standard deviation) becoming unit length. This non-linear transformation of the vectors produces a non-linear transformation of the factor loadings. The following example offers proof of this. Figure 1 portrays two unstandardized vectors in two-dimensional

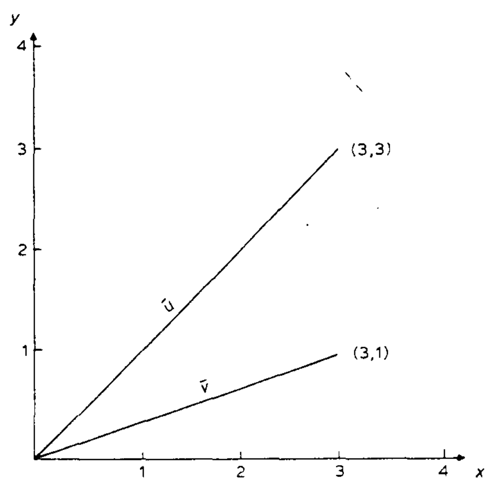


Fig. 1. Graphic portrayal of two unstandardized vectors.

space. The vectors shown here shar
terms of length. The length of these

$|\bar{a}| = \sqrt{x_1^2 + x_2^2 \ldots + x_n^2}$,

where: $|\bar{a}|$ = length of vector $\bar{a}$
x_i = coordinates in the vecto
n = dimension of the vector s

Thus, the length of vector $\bar{u}$ is:

$|\bar{u}| = \sqrt{3^2 + 3^2} = \sqrt{18}$

and the length of $\bar{v}$ is:

$|\bar{v}| = \sqrt{3^2 + 1^2} = \sqrt{10}$

To standardize these vectors (transf
sary to divide each vector by its own

$$\bar{u}' = \frac{1}{\sqrt{18}} (3, 3)$$

and

$$\bar{v}' = \frac{1}{\sqrt{10}} (3, 1)$$

A factor loading is obtained by
which is a projection of the data
axis. Projection of a vector onto ano

$\bar{w} = (\bar{u} \cdot \bar{x})(\bar{x})$

where: $\bar{w}$ is the projection of $\bar{u}$ on $\bar{x}$
$|\bar{x}| = 1$

For the following example let axis
nates (1, 0). $\bar{w}$ is the projection o
on $\bar{x}$:

$\bar{w} = [(3, 3) \cdot (1, 0)] [(1, 0)] = (3,$

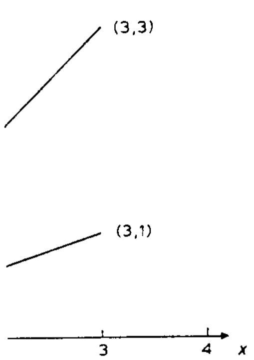
The length of $\bar{w}$ or the loading of $\bar{u}$

$|\bar{w}| = \sqrt{3^2 + 0^2} = 3$

Thus, the loading of $\bar{u}$ on $\bar{x}$ is 3,
Similarly:

$\bar{z} = [(3, 1) \cdot (1, 0)] [(1, 0)] = (3,$

nalysis be addressed, especially since
: performed on standardized variables
(Horst, 1965; Nie et al., 1975).
be twofold. First, it will examine the
zing standardized variables. This will
oof and an example from empirical
est a factor analytic technique which
h standardized variables.
acing a set of reference coordinates
ring the projection (loading) of each
ctor analysis it is helpful to consider
ntire set of variables as a vector space.
r space there are two major require-
ommon origin and the length of each
known. In the past researchers have
common origin is ordinarily provided
s deviation scores by subtracting the
value. The next step in the analysis
ion scores. However, this can only be
: transformation since each vector is
vn standard deviation [1]. This non-
ously effected on each vector length
eviation) becoming unit length. This
ectors produces a non-linear transfor-
e following example offers proof of
dardized vectors in two-dimensional



two unstandardized vectors.

space. The vectors shown here share a common origin (O), but differ in terms of length. The length of these vectors is given by the formula:

$$|\bar{a}| = \sqrt{x_1^2 + x_2^2 \ldots + x_n^2} \tag{1}$$

where: $|\bar{a}|$ = length of vector $\bar{a}$
x_i = coordinates in the vector space
n = dimension of the vector space

Thus, the length of vector $\bar{u}$ is:

$$|\bar{u}| = \sqrt{3^2 + 3^2} = \sqrt{18} \tag{2}$$

and the length of $\bar{v}$ is:

$$|\bar{v}| = \sqrt{3^2 + 1^2} = \sqrt{10} \tag{3}$$

To standardize these vectors (transform to unit length) it is only necessary to divide each vector by its own length such that

$$\bar{u}' = \frac{1}{\sqrt{18}}(3, 3) \tag{4}$$

and

$$\bar{v}' = \frac{1}{\sqrt{10}}(3, 1) \tag{5}$$

A factor loading is obtained by measuring the *length* of a vector which is a projection of the data vector onto the frame of reference axis. Projection of a vector onto another vector is given by the formula:

$$\bar{w} = (\bar{u} \cdot \bar{x})(\bar{x}) \tag{6}$$

where: $\bar{w}$ is the projection of $\bar{u}$ on $\bar{x}$
$|\bar{x}| = 1$

For the following example let axis X be a vector ($\bar{x}$) with the coordinates $(1, 0)$. $\bar{w}$ is the projection of $\bar{u}$ on $\bar{x}$ and $\bar{z}$ the projection of $\bar{v}$ on $\bar{x}$:

$$\bar{w} = [(3, 3) \cdot (1, 0)]\,[(1, 0)] = (3, 0) \tag{7}$$

The length of $\bar{w}$ or the loading of $\bar{u}$ on $\bar{x}$ is:

$$|\bar{w}| = \sqrt{3^2 + 0^2} = 3 \tag{8}$$

Thus, the loading of $\bar{u}$ on $\bar{x}$ is 3, simply its length along the X axis. Similarly:

$$\bar{z} = [(3, 1) \cdot (1, 0)]\,[(1, 0)] = (3, 0) \tag{9}$$

and

$$|\bar{z}| = \sqrt{3^2 + 0^2} = 3 \tag{10}$$

Hence, before standardization the loadings of each vector on $\bar{x}$ are the same.

Now turn to the loadings of the standardized versions of these vectors on $\bar{x}$. Let $\bar{k}$ be the projection of $\bar{u}'$ on $\bar{x}$ and $\bar{l}$ the projection of $\bar{v}'$ on $\bar{x}$:

$$\bar{k} = \left[\left(\frac{3}{\sqrt{18}}, \frac{3}{\sqrt{18}}\right) \cdot (1, 0)\right][(1, 0)] = \left(\frac{3}{\sqrt{18}}, 0\right) \tag{11}$$

and

$$|\bar{k}| = \sqrt{\left(\frac{3}{\sqrt{18}}\right)^2 + 0^2} = \frac{3}{\sqrt{18}} \tag{12}$$

Hence, the loading of the standardized version of $\bar{u}$ ($\bar{u}'$) on $\bar{x}$ is $3/\sqrt{18}$. The loading of $\bar{l}$ on $\bar{x}$ is

$$\bar{l} = \left[\left(\frac{3}{\sqrt{10}}, \frac{1}{\sqrt{10}}\right) \cdot (1, 0)\right][(1, 0)] = \left(\frac{3}{\sqrt{10}}, 0\right) \tag{13}$$

and

$$|\bar{l}| = \sqrt{\left(\frac{3}{\sqrt{10}}\right)^2 + 0^2} = \frac{3}{\sqrt{10}}$$

Hence the loading of the standardize

As can be seen from this simple
vectors $\bar{u}$ and $\bar{v}$ are the same in t
tions of vectors $\bar{u}'$ and $\bar{v}'$ are differ
inconsistencies in the factor loadir
problems in factor interpretation.
from the unstandardized variables
taminated by somewhat arbitrary m
to be preferred over the loadings ob
An example from some empirical d
is very well known to most everyon
of this point.

Accordingly the data in Table I w
the distances among 16 major Ame
equivalent to an ordinary sociolog
thought of as variables and the r
numerical entry (s_{ij}) may be though

TABLE I

Separations in Space Among 16 Selected U.S. Cities [a]

	Atlanta	Boston	Chicago	Cleve-land	Dallas	Denver	Detroit	Los Angeles	Miami	New Orleans	New York	Phoenix	Pittsbur
Atlanta	0	1508	944	891	1160	1950	869	2310	972	682	1204	2562	838
Boston		0	1369	886	2496	2846	986	4177	[illegible]019	2186	302	3701	777
Chicago			0	496	1292	1480	383	2807	[illegible]911	1340	1147	2338	660
Cleveland				0	1649	1974	145	3297	[illegible]749	1487	652	2814	185
Dallas					0	1067	1607	2005	[illegible]788	713	2211	1427	1721
Denver						0	1860	1337	[illegible]777	1741	2624	943	2124
Detroit							0	3191	[illegible]54	1511	1258	2719	330
Los Angeles								0	[illegible]764	2692	3944	574	3437
Miami									0	1076	1757	3189	1625
New Orleans										0	1884	2117	1478
New York											0	3451	510
Phoenix												0	2941
Pittsburgh													0
San Francisco													
Seattle													
Washington													

[a] The distances here are standard airline distances measured in kilometers.

$$(10)$$

loadings of each vector on $\bar{x}$ are the

e standardized versions of these vec-
ı of $\bar{u}'$ on $\bar{x}$ and $\bar{l}$ the projection of

$$))] = \left(\frac{3}{\sqrt{18}}, 0\right) \quad (11)$$

$$(12)$$

lized version of $\bar{u}$ ($\bar{u}'$) on $\bar{x}$ is $3/\sqrt{18}$.

$$))] = \left(\frac{3}{\sqrt{10}}, 0\right) \quad (13)$$

and

$$|\bar{l}| = \sqrt{\left(\frac{3}{\sqrt{10}}\right)^2 + 0^2} = \frac{3}{\sqrt{10}} \quad (14)$$

Hence the loading of the standardized version of $\bar{v}$ ($\bar{v}'$) on $\bar{x}$ is $3/\sqrt{10}$.

As can be seen from this simple exercise, whereas the projections of vectors $\bar{u}$ and $\bar{v}$ are the same in the unstandardized case, the projections of vectors $\bar{u}'$ and $\bar{v}'$ are different in the standardized case. These inconsistencies in the factor loadings due to standardization provide problems in factor interpretation. Since the factor loadings obtained from the unstandardized variables represent those loadings least contaminated by somewhat arbitrary mathematical manipulations, they are to be preferred over the loadings obtained from standardized variables. An example from some empirical data where the *real* factor structure is very well known to most everyone provides a forceful underscoring of this point.

Accordingly the data in Table I were assembled. These data represent the distances among 16 major American cities. The data are formally equivalent to an ordinary sociological data set; the columns may be thought of as variables and the rows as cases or individuals. Each numerical entry (s_{ij}) may be thought of as the score of the ith individ-

Cities [a]

o	Cleveland	Dallas	Denver	Detroit	Los Angeles	ami	New Orleans	New York	Phoenix	Pittsburgh	San Francisco	Seattle	Washington
	891	1160	1950	869	2310	972	682	1204	2562	838	3447	3511	874
	886	2496	2846	986	4177	019	2186	302	3701	777	4343	3979	632
	496	1292	1480	383	2807	911	1340	1147	2338	660	2990	2795	961
	0	1649	1974	145	3297	749	1487	652	2814	185	2485	3260	492
		0	1067	1607	2005	788	713	2211	1427	1721	2366	2705	1907
			0	1860	1337	777	1741	2624	943	2124	1524	1643	2404
				0	3191	154	1511	1258	2719	330	3364	3118	637
					0	764	2692	3944	574	3437	556	1543	3701
						0	1076	1757	3189	1625	4174	4399	1485
							0	1884	2117	1478	3099	3381	1554
								0	3451	510	4137	3874	330
									0	2941	1051	1792	3191
										0	3643	3440	309
											0	1091	3929
												0	3849
													0

ıces measured in kilometers.

ual on the jth variable. While formally equivalent to a typical sociological data set, however, the extreme precision and well-known configuration underlying these data make them ideal for the present example.

The data were entered into the SPSS (Nie et al., 1975) version 6.5 factor analysis program and factored by the principal components solution. This solution first centers the data on the mean of each variable by subtracting the mean of each variable from each of its elements. This matrix of deviation scores is then postmultiplied by its transpose to yield a matrix of scalar products. This scalar products matrix is then divided through by the sample size to obtain a variance–covariance matrix. This variance–covariance matrix is then standardized by dividing each cell by the product of the standard deviations of the variables intersecting in that cell. The result is a correlation matrix, or more appropriately, a matrix of cosines where each entry θ_{ij} represents the cosine of the angle between the variable vectors $\overline{i}$ and $\overline{j}$. Since the angle between $\overline{i}$ and $\overline{j}$ where $\overline{i} = \overline{j}$ is 0, and since cos 0 = 1, the diagonal entries of the matrix are unity. This matrix is then orthogonally decomposed to yield a matrix of eigenvectors or factors by a standard eigenvector routine. This solution is equivalent, in principal, to the previous example which transformed the hypothetical vectors $\overline{u}$ and $\overline{v}$ to unit length (eqns. 4 and 5) and then projected them onto $\overline{x}$ (eqns. 11–14).

The standardized output resulting from this principal components analysis is shown in Table II. Some factor analysts would consider this solution two-dimensional since the third eigenvalue is less than unity. However, we realize from our familiarity with these data that there are three dimensions underlying them, an east–west dimension, a north–south dimension, and a third dimension resulting from the curvature of the earth. If we attribute the 3.2 percent of the variance unaccounted for by factors one, two, and three as error variance [2], then it appears that the standardized factor analysis has uncovered the major dimensions underlying these data. Careful inspection of factor one would lead one to identify an east–west attribute quite easily, but it is unlikely that any standard interpretive scheme would lead unambiguously to a north–south interpretation of factor two. Since the pattern of loadings on factor three is not common knowledge, we will not treat it here.

What has taken place here can be made evident by plotting factors one and two as shown in Fig. 2 [3]. Figure 2 depicts a substantially distorted map of the U.S. This distortion is wholly a consequence of the standardization. Each city is constrained to be located one standard unit from the origin, and the result is a semi-circular U.S. with all but one of the cities located on the east or west coast. The consequences of this distortion are very severe, and one can readily notice the dimen-

TABLE II

Factor Loadings, Eigenvalues, and Percent Vari
Standardized Analysis

Cities	Factors	
	One	Two
Atlanta	0.8983	0.3517
Boston	0.9593	−0.1799
Chicago	0.8690	0.2714
Cleveland	0.9538	0.0554
Dallas	0.1935	0.9550
Denver	−0.6091	0.6649
Detroit	0.9235	0.1226
Los Angeles	−0.9192	0.3059
Miami	0.8651	0.2394
New Orleans	0.6689	0.6560
New York	0.9692	−0.1315
Phoenix	−0.8583	0.4572
Pittsburgh	0.9726	0.0194
San Francisco	−0.9664	0.1505
Seattle	−0.9140	0.0152
Washington	0.9863	−0.0298
Eigenvalue	12.0578	2.4337
Percent variance explained	75.4	15.2

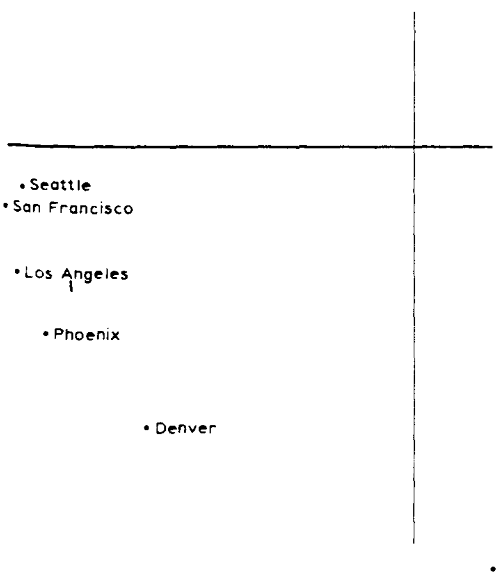


Fig. 2. Plot of factors one and two f

y equivalent to a typical sociolog-
ecision and well-known configura-
ideal for the present example.
PSS (Nie et al., 1975) version 6.5
by the principal components solu-
lata on the mean of each variable
riable from each of its elements.
en postmultiplied by its transpose
This scalar products matrix is then
obtain a variance–covariance ma-
x is then standardized by dividing
ndard deviations of the variables
is a correlation matrix, or more
here each entry θ_{ij} represents the
ble vectors $\bar{i}$ and $\bar{j}$. Since the angle
nce cos 0 = 1, the diagonal entries
is then orthogonally decomposed
factors by a standard eigenvector
n principal, to the previous exam-
cal vectors $\bar{u}$ and $\bar{v}$ to unit length
em onto $\bar{x}$ (eqns. 11–14).
; from this principal components
actor analysts would consider this
third eigenvalue is less than unity.
arity with these data that there are
an east–west dimension, a north–
ion resulting from the curvature of
cent of the variance unaccounted
error variance [2], then it appears
; has uncovered the major dimen-
nspection of factor one would lead
e quite easily, but it is unlikely that
ld lead unambiguously to a north–
nce the pattern of loadings on fac-
ve will not treat it here.
made evident by plotting factors
Figure 2 depicts a substantially dis-
on is wholly a consequence of the
ained to be located one standard
s a semi-circular U.S. with all but
or west coast. The consequences
one can readily notice the dimen-

TABLE II

Factor Loadings, Eigenvalues, and Percent Variance Explained for the First Five Factors of the Standardized Analysis

Cities	Factors				
	One	Two	Three	Four	Five
Atlanta	0.8983	0.3517	−0.1253	0.1712	−0.0859
Boston	0.9593	−0.1799	0.1489	−0.0149	0.1334
Chicago	0.8690	0.2714	0.3833	−0.0049	−0.0999
Cleveland	0.9538	0.0554	0.2840	0.0362	−0.0342
Dallas	0.1935	0.9550	−0.1016	−0.1184	0.0454
Denver	−0.6091	0.6649	0.3648	−0.1000	0.1268
Detroit	0.9235	0.1226	0.3316	0.0400	−0.1039
Los Angeles	−0.9192	0.3059	0.0642	0.2308	−0.0094
Miami	0.8651	0.2394	−0.4081	−0.0109	−0.0319
New Orleans	0.6689	0.6560	−0.3153	−0.0453	−0.0698
New York	0.9692	−0.1315	0.1153	0.0006	0.1441
Phoenix	−0.8583	0.4572	0.0805	0.1364	0.1177
Pittsburgh	0.9726	0.0194	0.2166	0.0404	0.0043
San Francisco	−0.9664	0.1505	0.1493	0.0856	−0.0382
Seattle	−0.9140	0.0152	0.3274	−0.1362	−0.1640
Washington	0.9863	−0.0298	0.1160	0.0426	0.0700
Eigenvalue	12.0578	2.4337	0.9986	0.1600	0.1391
Percent variance explained	75.4	15.2	6.2	1.0	0.9

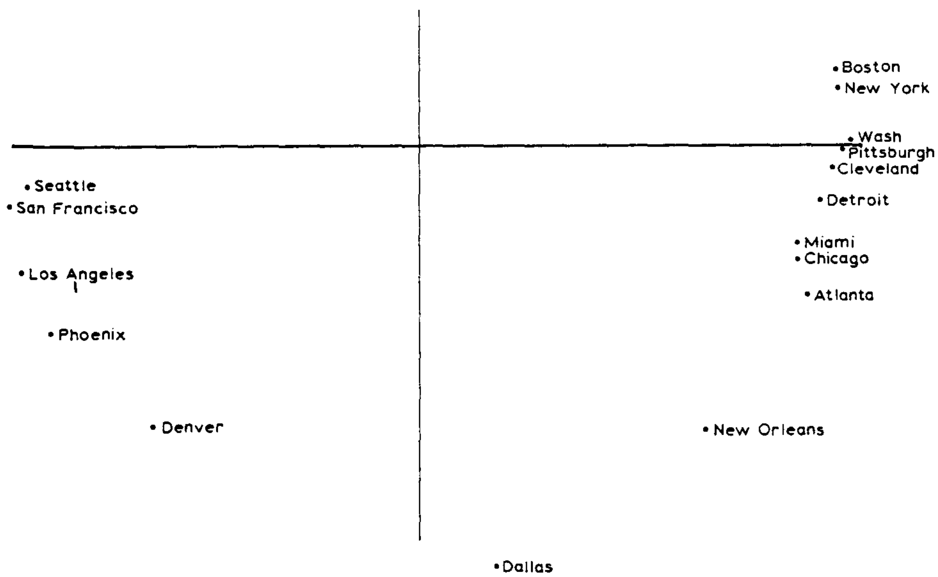


Fig. 2. Plot of factors one and two from the standardized analysis.

sions even contain order inversions. Washington and Pittsburgh, for example, are incorrectly portrayed as farther east than Boston and New York. Miami is incorrectly portrayed as farther west than Cleveland, Detroit, Atlanta, and Chicago. Factor two, which should represent a north–south dimension, contains even greater distortion. The most extreme distortion places Miami north of Chicago. In other instances Washington is north of Seattle and Phoenix is north of Denver.

Lest one believe that distorted as it may be, this factor analytic picture is still the best that might be hoped for, the same data were entered into a metric multidimensional scaling program, Galileo version 3.9 (Gilham and Woelfel, 1976; Woelfel, 1976). This program, like SPSS version 6.5, first centers the data on the mean of the variables by subtracting the mean of each variable from each of its elements. This matrix of deviation scores is then postmultiplied by its transpose to yield a matrix of scalar products. The scalar products matrix is then divided through by the sample size to obtain a variance–covariance matrix. However, instead of standardizing this matrix like the SPSS program, the Galileo program factors this variance–covariance matrix. Its output, therefore, may be interpreted directly as an unstandardized factor analysis [4] and is similar, in principle, to the previous example which measured the projection of the unstandardized vectors $\bar{u}$ and $\bar{v}$ on $\bar{x}$ (eqns. 7–10). The results of this unstandardized analysis are shown in Table III.

Since the data are not standardized, each of the columns (factors) represents the distance in kilometers of the cities' projections on the factors from the origin of the space. Note that proportionately more of the variance lies on the first three factors, with only 1.77 percent unaccounted for by them. This is most likely a better representation of the error in the data than the estimate from the principal components analysis. In addition, since the curvature of the earth should account for only about 1 percent of the variance in these data, the estimate of 3.88 percent obtained by factor three is a truer estimate than the corresponding 6.2 percent yielded by the standardized analysis. Hence, in the category of variance explained, the unstandardized version presents a slightly better description of these data than the standardized version.

The major difference, however, between these two techniques lies in the pattern of loadings they yield. Figure 3 [5] presents the plot of factors one and two from the unstandardized version. It is clearly a nearly perfect map of the U.S. cities. Little of the distortion due to standardization can be found here. On factor one there is only one minor inversion with Miami placed east of Pittsburgh. Actually, Pittsburgh is

TABLE III

Factor Loadings, Eigenvalues, and Percent Va
Unstandardized Analysis

Cities	Factors	
	One	Two
Atlanta	−680.8	620.
Boston	−1730.0	−753.
Chicago	−413.4	−390.
Cleveland	−909.6	−424.
Dallas	320.3	653.
Denver	1043.6	−124.
Detroit	−760.0	−454.
Los Angeles	2257.1	535.
Miami	−1374.2	1244.
New Orleans	−339.9	927.
New York	−1548.5	−457.
Phoenix	1747.1	500.
Pittsburgh	−1073.7	−336.
San Francisco	2574.5	−156.
Seattle	2277.7	−1227.
Washington	−1390.2	−156.
Eigenvalue	3.4 $\times 10^7$	6. $\times 10^6$
Percent variance explained	78.6	15.

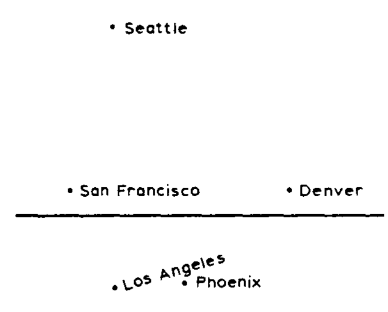


Fig. 3. Plot of factors one and tv

. Washington and Pittsburgh, for
as farther east than Boston and
·trayed as farther west than Cleve-
. Factor two, which should repre-
tains even greater distortion. The
iami north of Chicago. In other
Seattle and Phoenix is north of

it may be, this factor analytic pic-
ɜ hoped for, the same data were
nal scaling program, Galileo version
'oelfel, 1976). This program, like
ata on the mean of the variables by
ɔle from each of its elements. This
postmultiplied by its transpose to
The scalar products matrix is then
ɜ to obtain a variance–covariance
ırdizing this matrix like the SPSS
:s this variance–covariance matrix.
rpreted directly as an unstandard-
nilar, in principle, to the previous
:tion of the unstandardized vectors
sults of this unstandardized analysis

zed, each of the columns (factors)
rs of the cities' projections on the
Note that proportionately more of
factors, with only 1.77 percent un-
ɔst likely a better representation of
ate from the principal components
vature of the earth should account
iance in these data, the estimate of
ɜe is a truer estimate than the corre-
ıe standardized analysis. Hence, in
the unstandardized version presents
data than the standardized version.
etween these two techniques lies in
igure 3 [5] presents the plot of fac-
ırdized version. It is clearly a nearly
ɜ of the distortion due to standard-
r one there is only one minor inver-
Pittsburgh. Actually, Pittsburgh is

TABLE III

Factor Loadings, Eigenvalues, and Percent Variance Explained for the First Five Factors of the Unstandardized Analysis

Cities	Factors One	Two	Three	Four	Five
Atlanta	−680.8	620.8	816.2	−2.6	−5.0
Boston	−1730.0	−753.0	77.7	3.8	−158.4
Chicago	−413.4	−390.0	−42.5	13.1	55.8
Cleveland	−909.6	−424.6	−13.9	17.6	24.1
Dallas	320.3	653.5	−276.0	−1.8	−27.3
Denver	1043.6	−124.2	−167.7	1.8	108.3
Detroit	−760.0	−454.7	27.2	497.8	−38.5
Los Angeles	2257.1	535.6	750.4	−31.9	8.9
Miami	−1374.2	1244.1	−275.9	−10.3	−163.7
New Orleans	−339.9	927.7	−278.3	−8.0	−85.5
New York	−1548.5	−457.7	14.8	−479.4	38.8
Phoenix	1747.1	500.3	−322.5	−12.9	21.1
Pittsburgh	−1073.7	−336.8	−21.9	17.9	−1.1
San Francisco	2574.5	−156.5	−248.9	−11.1	139.8
Seattle	2277.7	−1227.8	−17.8	−63.4	−234.0
Washington	−1390.2	−156.5	−21.0	69.4	316.7
Eigenvalue	3.4 × 10^7	6.8 × 10^6	1.7 × 10^6	4.9 × 10^5	2.5 × 10^5
Percent variance explained	78.6	15.8	3.9	1.1	0.6

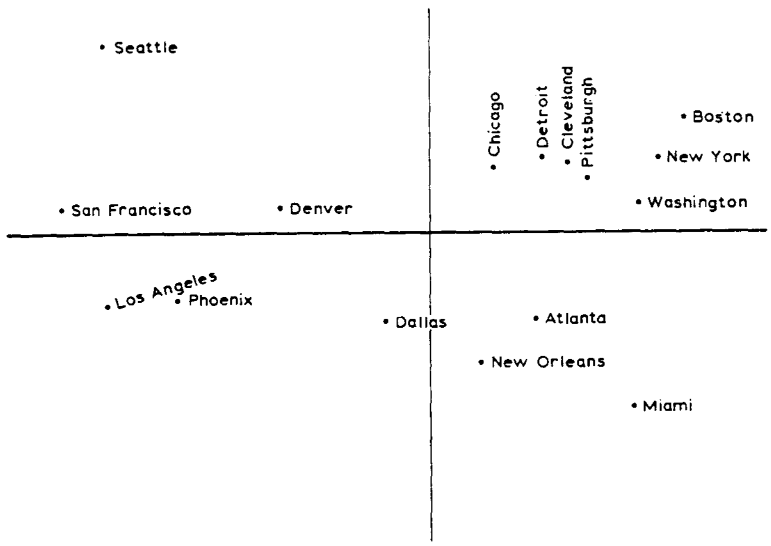


Fig. 3. Plot of factors one and two from the unstandardized analysis.

aproximately 30 kilometers east of Miami. On factor two there are several inversions, but they too are relatively small. The worst of these inversions places San Francisco north of Denver. In reality Denver is about 228 kilometers north of San Francisco.

The comparison between the standardized and unstandardized factor analyses conducted here has revealed striking differences. While both types of analysis yielded similar estimates of variance explained per factor, the unstandardized analysis provided an unquestionably better portrayal of the factor loadings. The major problem with the standardized version was that it constrained the communalities of each variable to be the same (unit length). As a result, the pattern of loadings was necessarily semi-circular. The obvious implication from this is to avoid standardizing variables which are to be factor analyzed. However, a more far-reaching implication emerges from these findings. As we noted previously, there are two critical requirements for factor analysis, a common origin for the vectors and knowledge of the vector lengths or communalities. Our solution has been to factor a variance–covariance matrix using the variance of each vector as its communality. However, variance is largely a function of the unit of measurement chosen for a variable. If one attempts to factor variables measured on *different scales* by means of factor-analyzing a variance–covariance matrix, then one runs the risk of biasing the outcome since the variables measured on the larger scales will most likely have the larger vector lengths, and, consequently, larger factor loadings. As we have shown, standardization is not a legitimate method for circumventing this problem. The *only* solution is to measure all variables on common scales. For example, in the case of attitude measures, a researcher may employ all Likert items as the measure of attitudes. However, for the case where the variables are not measured on the same scale, the researcher cannot hope to achieve mathematical miracles by standardizing his/her variables prior to factor analysis.

One more point must be considered before concluding. This paper has examined the factor analyzing of correlational matrices with unities in the diagonals. A technique sometimes employed in factor analysis is to remove the unities from the diagonal of the correlation matrix and replace them with some other values. These replacement values are quite arbitrary with two of the more common replacements being each variable's highest correlation with any other variable in the matrix or the squared multiple correlation of each variable with all other variables in the matrix (Harmon, 1960). This replacement technique *does not offer* any improvement over the inadequate method of using the unaltered correlation matrix. Replacing the diagonal elements has two

drawbacks. First, this replacement
mation of the data to the initial r
standardization. Hence, this data
further from the actual data repor
factor solution obtained after repla
gent on the replacement values. Sir
know the factor pattern underlyin
being confident of the solution obta

No

1 The only instance where this would be
dard deviations were the same.
2 This error would be due to measuren
kilometers. Actually, 3.2 percent is pr
tained in these data. About 2 percent is
3 North and south are inverted in Table
ple, Dallas appears to be the most nortl
pretation, the signs on factor two were
4 The Galileo program was used because
ability. The user without access to a
could obtain essentially identical res
matrix.
5 Both the east–west and north–south c
sign reversal. As was the case for the s
tion, the signs on factors one and two w

Refer

Blalock, H. (1964). *Causal Inference in*
University of North Carolina Press.
Blalock H. (1967). "Path coefficients v
Journal of Sociology 72: 675–676.
Gilham, J. and Woelfel, J.D. (1976). "Th
precision, stability, and equivalence tc
cation Research (Fall 1976).
Harmon, H. (1960). *Modern Factor Analy*
Horst, P. (1965). *Factor Analysis of Dat*
Winston.
Nie, N., Hull, C.H., Jenkins, J., Steinbrer
tical Package for the Social Sciences. N
Tukey, J. (1954). "Causation, regression
al., eds., *Statistics and Mathematics in*
Turner, M. and Stevens, C. (1969). "T
Biometrics 15: 236–258.

Miami. On factor two there are sev-
latively small. The worst of these
·th of Denver. In reality Denver is
'rancisco.
dardized and unstandardized factor
ed striking differences. While both
nates of variance explained per fac-
vided an unquestionably better por-
ajor problem with the standardized
ommunalities of each variable to be
the pattern of loadings was neces-
plication from this is to avoid stan-
factor analyzed. However, a more
m these findings. As we noted pre-
rements for factor analysis, a com-
vledge of the vector lengths or com-
factor a variance–covariance matrix
its communality. However, variance
measurement chosen for a variable.
s measured on *different scales* by
e–covariance matrix, then one runs
nce the variables measured on the
ne larger vector lengths, and, conse-
e have shown, standardization is not
ing this problem. The *only* solution
ion scales. For example, in the case
may employ all Likert items as the
the case where the variables are not
researcher cannot hope to achieve
zing his/her variables prior to factor

ered before concluding. This paper
f correlational matrices with unities
times employed in factor analysis is
gonal of the correlation matrix and
lues. These replacement values are
re common replacements being each
any other variable in the matrix or
f each variable with all other vari-
). This replacement technique *does*
he inadequate method of using the
cing the diagonal elements has two

drawbacks. First, this replacement adds another non-linear transformation of the data to the initial non-linear transformation caused by standardization. Hence, this data matrix is removed even one step further from the actual data reported by the respondent. Second, the factor solution obtained after replacing the diagonals is highly contingent on the replacement values. Since the researcher typically does not know the factor pattern underlying his/her data, there is no way of being confident of the solution obtained.

Notes

1 The only instance where this would be a linear transformation is if all the standard deviations were the same.

2 This error would be due to measurement error and conversion from miles to kilometers. Actually, 3.2 percent is probably a liberal estimate of the error contained in these data. About 2 percent is the maximum we would expect.

3 North and south are inverted in Table II simply due to sign reversal. For example, Dallas appears to be the most northerly of the sixteen cities. For aid in interpretation, the signs on factor two were reversed before plotting.

4 The Galileo program was used because of its convenient format and ready availability. The user without access to a metric multidimensional scaling program could obtain essentially identical results by factoring a variance–covariance matrix.

5 Both the east–west and north–south dimensions in Table III are inverted due to sign reversal. As was the case for the standardized version, for aid in interpretation, the signs on factors one and two were reversed.

References

Blalock, H. (1964). *Causal Inference in Non-Experimental Design*. Chapel Hill: University of North Carolina Press.

Blalock H. (1967). "Path coefficients versus regression coefficients," *American Journal of Sociology* 72: 675–676.

Gilham, J. and Woelfel, J.D. (1976). "The Galileo system: preliminary evidence for precision, stability, and equivalence to traditional measures," *Human Communication Research* (Fall 1976).

Harmon, H. (1960). *Modern Factor Analysis*. Chicago: University of Chicago Press.

Horst, P. (1965). *Factor Analysis of Data Matrices*. New York: Holt, Rinehart and Winston.

Nie, N., Hull, C.H., Jenkins, J., Steinbrenner, K. and Bent, D. (1975). *SPSS: Statistical Package for the Social Sciences*. New York: McGraw-Hill.

Tukey, J. (1954). "Causation, regression, and path analysis," in O. Kempthorne et al., eds., *Statistics and Mathematics in Biology*. Ames: Iowa State College Press.

Turner, M. and Stevens, C. (1969). "The regression analysis of causal paths," *Biometrics* 15: 236–258.

Woelfel, J.D. (1976). "Foundations of cognitive theory," unpublished manuscript, Michigan State University.
Wright, S. (1960). "Path coefficients and high regressions: alternative or complimentary concepts?", *Biometrics* 16: 189–202.

Quality and Quantity, 14 (1980) 711–72
Elsevier Scientific Publishing Company, A

ALTERNATIVE INFERENCE M
FOR A WEIGHTED INDEX OF
AGRE

LAWRENCE

University of California,

1. Intro

The kappa (κ) statistic develope
known in psychology for measurin
(see Fleiss, 1973). In a typical ap
the same set of n objects using T u
If the resulting data are formalized
as that given in Table I, then in ter
defined by

$$\kappa = (P_o - P_e)/(1 - P_e)$$

where

$$P_o = \sum_{u=1}^{T} n_{uu}/n$$

is the observed proportion of agreem

$$P_e = \sum_{u=1}^{T} n_{u\cdot}n_{\cdot u}/n^2$$

is the expected proportion under t
and fixed row and column marginal

Although the expression for P_e
chology, the field of sociology gen
sion suggested by Scott (1955).
dorff (1970), and others define

$$P_e = \sum_{u=1}^{T} (n_{u\cdot} + n_{\cdot u})^2/4n^2$$