An Inquiry into the Disparity in Educational Development

Anil K. Yadav*

1. Introduction

Education is fundamental to all-round human development—material as well as spiritual. To realise this, Article 45 of the Constitution of India has the provision of providing free and compulsory education for all children up to the age of 14 years. Keeping this objective of providing education to children in mind, efforts have been made by the government from time to time to enhance the literacy rate and the education levels. Towards this endeavour, Kothari Commission (1966) was constituted to make suitable recommendations so that the goal of Universalisation of Elementary Education and in turn the all-round human development may be achieved. The report suggested around 6 per cent of Gross National Product (GNP) to be spent on education in order to attain the goal of achieving the universalisation of elementary education. Soon after the submission of the Kothari Commission's report various expert groups were formed to make specific recommendations. On the recommendations of these groups the National Policy on Education was implemented in 1968.

Keeping alive the issue of universalisation of elementary education and knowing well the importance of primary education, a new education policy was brought out in 1986. Once again many expert groups were called to deliberate on the recommendation of the New Education Policy. Based on the recommendations of these high level committees Programme of Action (POA) was framed and implemented in 1992. As a result of National Policy on Education, 1986 and the POA, 1992 there has been a considerable emphasis on expansion of educational facilities throughout the country. A strong thrust has been given

* Chief, Institute of Applied Manpower Research, Delhi.

The author is grateful to Indra Kumar and Ashok Kumar Yadav for research inputs and Devendra P. Kohad for secretarial assistance. Special thanks to Dipika Sen for editing the earlier draft.

to education development through this policy. Many milestones have been established in the field of education. In addition to this, it had also been argued during the early 1990s that India should spend at least 3 per cent of its GDP on education, and later it was felt that instead of 3 per cent 6 per cent be spent. To attain this goal, the government has made some efforts. One leap towards this step came in the form of "Sarva Shiksha Abhiyan" (2001). This has been done in order to meet the goal of universalisation of primary and elementary education. The basic foundation of these initiatives was to have equality in the field of education. However, the available data indicate glaring disparities at the state level in access to education, infrastructure facilities, availability of teachers, outcomes and human resource development (Yadav & Srivastava, 2001, 2005), Shariff (1999) and Mehta (2007). The problems of disparity are still more of perennial nature at the district and block levels.

Available evidence indicate that the poor have undoubtedly suffered from inadequate educational allocations, and poor management and institutional mechanisms to ensure equity (Shariff and Ghosh, 1999). Moreover, there is probably a mismatch between what is provided in the form of social services (supply) and the requirements of the target group.

Keeping in view this background, it is felt to develop the Educational Development Index at the district levels in the country to probe deeper into the disparity in educational development. This would subsequently serve to achieve the goal of higher educational development as well as the equity through need-based funding at state and national levels.

An attempt has been made in this paper to make an inquiry into the disparity in the educational development at the district level so that appropriate funding could be realised.

2. Database and Methodology

The data used in this exercise is the District Information System on Education (DISE) data managed by National University of Educational Planning and Administration (NUEPA), New Delhi. The Educational Development Index has been computed at the district level. Therefore, initially it is proposed to cover two districts each from six major states in the country. The names of the districts and their respective states are placed below:

List of Districts in Selected States

S.No.	Name of the State	Name of the Districts
1.	Uttar Pradesh	Baghpat Gorakhpur
2.	Madhya Pradesh	Dhar Betul
3.	West Bengal	Cooch Behar South 24 Parganas
4.	Rajasthan	Dungarpur Bikaner
5.	Karnataka	Dharwad Hassan
6.	Tamil Nadu	Coimbatore Madurai

Methodology

To develop educational development index at district level, following four broad indicators with subdivision have been considered for the study.

1) **Access**
   Percentage of Habitations not Served
   Number of Schools per 1000 Population

2) **Infrastructure Gap**
   Average Student-Classroom Ratio (SCR)
   School with SCR > 60
   Percentage of Schools without Drinking Water Facility
   Percentage of Schools with Boys' Toilet
   Percentage of Schools with Girls' Toilet

3) **Teachers**
   Percentage of Female Teachers
   Average Pupil-Teacher Ratio

Percentage of Schools with Pupil-Teacher Ratio> 60

Percentage of Single-Teacher Schools with the Number of Students > 15

Percentage of Schools with three or less Teachers

Percentage of Teachers with Professional Qualifications

(4) **Outcomes**

Gross Enrolment Ratio (GER) – Overall

GER – Scheduled Castes

GER – Scheduled Tribes

Gender Parity Index in Enrolment

Repetition Rate

Dropout Rate

Ratio of Exit Class over Class I Enrolment (Only at Primary Stage)

Percentage of Passed Children to Total Enrolment

Percentage of Appeared Children Passed with > 60 per cent

The Principal Component Analysis method is used for developing the educational development index at the district levels. The details of this method are given in the annexure. The formulae used in preparing the Index is as follows:

Before making the Index, we have to undergo a lot of exercises. The very first exercise is to normalise or standardise the whole distribution. In order to normalise the distribution the following formulae is used. It becomes imperative to mention that for the negative variables sign becomes the best value.

Normalisation Method

$$NVij = 1 - \left(\frac{\{\text{Best } Xi - \text{Observed } Xij\}}{\{\text{Best } Xi - \text{Worst } Xi\}} \right)$$

As soon as the normalisation is done the whole data is placed appropriately in the SPSS data sheet. We go for the Principle Component Analysis. We derive the components and the factor loadings from the run as shown in the annexure. Then, we derive the weights by multiplying the factor loadings and the eigen values. The following formulae is used in order to obtain the weights.

Formulae used to compute weights:

$$\text{Weights} = \sum_{i,j=1}^{n} /Fij/Ej$$

Where Fij is the factor loading
Value of the ith variable in the
jth factor and
Ej is the eigen value of the jth factor

Once we have got the weights of the distribution, the normalised values are multiplied with the weights. These derived values are then divided by the double summation of the factor loadings and the eigen values. The formulae used for doing this is given below. Secondly, in order to remove the negative values, if any, we take the mode of the factor loadings as has been shown in the formulae.

$$I = \frac{\sum_{i=1}^{n} V_i \left(\sum_{j=1}^{n} \left| F_{ij} \right| . E_j \right)}{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \left| F_{ij} \right| . E_j \right)}$$

Where I is the Index value
Vi is the indicator
Fij is the factor loading value of the ith variable on the jth factor
and Ej is the eigen value of the jth factor.

3. Educational Development

Since independence, India has been developing its education system by way of enhancing the facilities. The government has always been concerned about providing the education to all. Many new schools have been opened by the government. A lot more has been done to provide better educational facilities such as, black board, teaching aids (Charts, Audio-Video, etc.), building, desks and carpet (Tat-patti), sports facilities, etc. Yadav and Srivastava (2002) have noted a considerable expansion in education facilities. In this section, we have made an attempt to capture the facilities available in the schools up to elementary level.

i) Primary Level

We have tried here to capture the educational development through an Index. The index as mentioned above has been developed for the four indicators such as Access, Infrastructure, Teachers and the Outcomes. It may be observed through Table 1.1 that the most developed district among all is the Coimbatore District followed by Dharwad and Madurai. The worst among them is the South 24 Parganas and a bit better than this are Dhar and Kooch Bihar. The districts' positions differ according to different indicators. This is quite visible from the index values given in the table. It is interesting to note that in terms of Access the best district is Dhar, while for Infrastructure it is the Baghpat. Similarly, the availability of teachers is the best in Dharwad whereas Coimbatore is leading in Outcomes. This makes one to understand that the districts which are weak in a particular indicator need to be strengthened appropriately. Whereas, the districts which are better off in some indicator may need a moderate help. Let us, for example, see the case of Baghpat District. It requires considerable additional funding for the Access in order to enhance the Access facilities for the children, whereas for infrastructure it does not require much. Similar is the case for other districts.

Table 1 : Education Development Index for Primary Education at District Level

	EDI at Primary Level					
Districts	**Access**	**Infra-structure**	**Teachers**	**Outcomes**	**Primary Level**	**EDI Rank at Primary Level**
BAGHPAT	0.130	0.821	0.399	0.586	0.529	7
BETUL	0.611	0.412	0.489	0.498	0.490	8
BIKANER	0.458	0.550	0.511	0.606	0.541	5
COIMBATORE	0.444	0.720	0.823	0.792	0.726	1
DHAR	0.725	0.274	0.471	0.490	0.462	11
DHARWAD	0.403	0.668	0.874	0.533	0.643	2
DUNGARPUR	0.376	0.449	0.532	0.527	0.483	9
GORAKHPUR	0.516	0.714	0.306	0.571	0.530	6
HASSAN	0.222	0.471	0.793	0.709	0.590	4
KOOCH BIHAR	0.216	0.545	0.588	0.449	0.477	10
MADURAI	0.239	0.606	0.862	0.652	0.632	3
SOUTH 24 PARGANAS	0.304	0.183	0.119	0.419	0.255	12

ii) Upper Primary Level

Table 1.2 gives the index values for various indicators at the upper primary level. Similar to the primary level here also Coimbatore District emerges as the best one followed by Bikaner and then Madurai. The indicator-wise position is similar to earlier one. Although Bikaner stands second, it is weak in terms of Access.

Table 2 : Education Development Index at Upper Primary Level

	EDI at Primary Level					
Districts	**Access**	**Infra-structure**	**Teachers**	**Outcomes**	**Primary Level**	**EDI Rank at Primary Level**
BAGHPAT	0.051	0.959	0.432	0.538	0.477	10
BETUL	0.664	0.296	0.565	0.473	0.508	9
BIKANER	1.139	0.683	0.794	0.327	0.755	2
COIMBTORE	0.811	0.784	0.857	0.786	0.812	1
DHAR	0.733	0.474	0.743	0.276	0.572	7
DHARWAD	0.327	0.751	0.901	0.639	0.654	5
DUNGARPUR	0.375	0.716	0.810	0.244	0.544	8
GORAKHPUR	0.445	0.865	0.499	0.542	0.579	6
HASSAN	0.500	0.790	0.872	0.744	0.725	4
KOOCH BIHAR	0.357	0.502	0.277	0.435	0.386	12
MADURAI	0.509	0.726	0.939	0.744	0.731	3
SOUTH 24 PARGANAS	0.281	0.554	0.495	0.487	0.450	11

iii) Elementary Level

Table 1.3 indicates the index value for total, i.e., elementary education. In this level also Coimbatore District stands first followed by Madurai and Hassan. This makes it clear that districts in Tamil Nadu are placed better than the districts in other states considered here. One may safely state that Tamil Nadu is quite developed in terms of educational facilities and standards.

Table 3 : Educational Development Index for Elementary Education

Districts	Composite (Primary /Upper Primary) EDI			
	EDI at Primary Level	EDI at Upper Primary Level	Composite (Primary+Upper Upper Primary) EDI	Composite (Primary+Upper Upper Primary) Rank
BAGHPAT	0.529	0.477	0.503	9
BETUL	0.490	0.508	0.499	10
BIKANER	0.541	0.755	0.648	4
COIMBATORE	0.726	0.812	0.769	1
DHAR	0.462	0.572	0.517	7
DHARWAD	0.643	0.654	0.648	5
DUNGARPUR	0.483	0.544	0.514	8
GORAKHPUR	0.530	0.579	0.555	6
HASSAN	0.590	0.725	0.657	3
KOOCH BIHAR	0.477	0.386	0.431	11
MADURAI	0.632	0.731	0.681	2
SOUTH 24 PARGANAS	0.255	0.450	0.352	12

4. Conclusion

In this paper, we have tried to highlight the disparity in educational development at the district level. The underpinning of development in education reveals that some districts are far ahead of the other districts. In our case Coimbatore emerges as the best district among all the districts in the composite index. While indicator-wise positions are different. It means that interventions are required in those indicators in which a particular district is weak. In order to uplift the weak districts as per EDI, a resource re-allocation is necessary.

References

Mehta, Arun C. (2007), "*Elementary Education in India – Progress towards UEE* ", Analytical Report, 2005-06, NUEPA & MHRD, Government of India.

Ministry of Education (1966), *Report of the Education Commission, 1964-66*, Education and National Development (Kothari Commission), Government of India, Delhi.

Naik, J.P. (1997), *The Education Commission and After*, A.P.H. Publishing Corporation, Delhi.

Shariff, A. and Prabir Ghosh (1999), "Education Expenditure in Indian States", A paper presented at a conference organised by Ed.CIL and ISEC, Banglore, held at Bangalore during August 19-22, 1999 and published in *EPW*, Vol. 35, No. 16, 2000, April 14-21.

____ (1999), *India: Human Development Report, A Profile of Indian States in the 1990s*, Oxford University Press, New Delhi.

Yadav, Anil K. and Madhu Srivastava (2001), *Educational Development Parameters and the Preparation of Educational Development Index*, Institute of Applied Manpower Research, New Delhi, Planning Commission, Government of India.

____ (2002), "Is Universalisation of Primary Education Possible?" *Journal of Educational Planning and Administration*, Vol. XVI, No.1, January, 2002, pp. 99-111.

____ (2005), *Education Development Index in India – An Inter-state Perspective*, Institute of Applied Manpower Research, Manak Publication Pvt. Ltd., Delhi.

Annexure

Data Reducation Techniques

The term 'Data Reduction' is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large data sets into manageable (smaller) information nuggets. Data reduction techniques can include simple tabulation, aggregation (computing descriptive statistics) or more sophisticated techniques like principal components analysis, factor analysis, cluster analysis, etc. Here, mainly principal component analysis (PCA) and factor analysis are covered along with example and software demonstration. At the end, different commands used in SPSS and SAS packages for PCA and factor analyses are given.

Principal Components Analysis

Most of the times the variables under the study are highly correlated and as such they effectively "say the same things". To examine the relationships among a set of *p* correlated variables, it may be useful to transform the original set of variables to a new set of uncorrelated variables called 'principal components'. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that for example, the first principal component accounts for as much as possible of the variation in the original data. Also, PCA is a linear dimensionality reduction technique, which identifies orthogonal directions of maximum variance in the original data, and projects the data into a lower-dimensionality space formed of a sub-set of the highest variance components.

Let $x_1, x_2, x_3, \ldots x_p$ are variables under study, then first principal component may be defined as

$$z_1 = a_{11} x_1 + a_{12} x_2 + \ldots\ldots\ldots + a_{1p} x_p$$

such that variance of z1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + \ldots\ldots + a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then Var(z_1) can be increased simply by multiplying any a1js by a constant factor. The second principal component is defined as

$z_2 = a_{21}x_1 + a_{22}x_2 + \ldots\ldots + a_{2p}x_p$

such that Var (z_2) is as large as possible next to Var(z_1) subject to the constraint that

$a_{21}^2 z + a_{22}^2 + \ldots\ldots a_{2p}^2 = 1$ and cov ($z_{1,2}$) = 0 and so on.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus, reducing the effective dimensionality of the problem. An analysis of principal components often reveals the relationships that were not previously suspected thereby allowing interpretation that would not ordinarily result. However, Principal Component Analysis is more of a means to an end rather than an end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique which does not require user to specify the statistical model or assumption about the distribution of original varieties. It may also be mentioned that principal components are artificial variables as often it is not possible to assign physical meaning to them. Further, since Principal Component Analysis transforms original set of variables to a new set of uncorrelated variables, it is worth stressing that if original variables are uncorrelated, then there is no point in carrying out principal component analysis.

Principal Components from Sample Variance-Covariance Matrix

Let $\mathbf{x_1, x_2, \ldots, x_n}$ represent n independent observations from some p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. These data yield the sample mean vector $\bar{\mathbf{x}}$, the sample covariance matrix **S**, and the sample correlation matrix **R**. Here, the objective is to construct uncorrelated linear combination of the measured characteristics that account for much of the variation in the sample. The uncorrelated combinations with the largest variances will be called the sample principal components.

If **S** = $\{S_{ik}\}$ is the p x p sample covariance matrix with eigen value-eigen vector pairs $(\hat{\lambda}_1,\hat{e}_1)$, $(\hat{\lambda}_2,\hat{e}_2),\ldots..,(\hat{\lambda}_p,\hat{e}_p)$ the ith sample principal component is given by

$\hat{z}_i = \hat{e}_i' x = \hat{e}_i' {}_1 x_1 + \hat{e}_i' {}_2 x_2 + \ldots + \hat{e}_i' {}_p x_p, \qquad i = 1,2,\ldots\ldots,p$

Where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \cdots \hat{\lambda}_p \geq 0$

and x is any observation on the variables $x_1, x_2, \ldots .. x_p$.

Also,

sample variance $(\hat{z}_k) = \hat{\lambda}_k$, $k = 1,2,\ldots..,p$.

sample covariance $(\hat{z}_i, \hat{z}_k) = 0$, $i \neq k$.

In addition, the total sample variance $= \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \ldots. + \hat{\lambda}_p$

Computation of Principal Components and Scores

Let us consider the following data on average minimum temperature (x_1), average relative humidity at 8 hrs. (x_2), average relative humidity at 14 hrs. (x_3) and total rainfall in cm. (x_4) pertaining to Raipur district from 1970 to 1986 for Kharif season from 21st May to 7th Octotober.

x_1	x_2	x_3	x_4
25.0	86	66	186.49
24.9	84	66	124.34
25.4	77	55	098.79
24.4	82	62	118.88
22.9	79	53	071.88
07.7	86	60	111.96
25.1	82	58	099.74
24.9	83	63	115.20
24.9	82	63	100.16
24.9	78	56	062.38
24.3	85	67	154.40
24.6	79	61	112.71
24.3	81	58	079.63
24.6	81	61	125.59
24.1	85	64	099.87
24.5	84	63	143.56
24.0	81	61	114.97
Mean 23.56	82.06	61.00	112.97
S.D. 4.13	2.75	3.97	30.06

with the variance -covariance matrix

$$\Sigma = \begin{pmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{pmatrix}$$

Let the eigen values in decreasing order and corresponding eigen vectors are

$\lambda_1 = 916.902$ $a_1 = (0.006, 0.061, 0.103, 0.993)$

$\lambda_2 = 18.375$ $a_2 = (0.955, -0.296, 0.011, 0.012)$

$\lambda_3 = 7.87$ $a_3 = (0.141, 0.485 0.855 -0.119)$

$\lambda_4 = 1.056$ $a_4 = (0.260, 0.820, -0.509, 0.001)$

The principal components for this data will be

$z_1 = 0.006\ x_1 + 0.061\ x_2 + 0.103\ x_3 + 0.993\ x_4$

$z_2 = 0.955\ x_1 - 0.296\ x_2 + 0.011\ x_3 + 0.012\ x_4$

$z_3 = 0.141\ x_1 + 0.485\ x_2 + 0.855\ x_3 - 0.119\ x_4$

$z_4 = 0.26\ x_1 + 0.82\ x_2 - 0.509\ x_3 + 0.001\ x_4$

The variance of principal components will be eigen values, i.e.,

$\text{Var}(z_1) = 916.902$, $\text{Var}(z_2) = 18.375$, $\text{Var}(z_3) = 7.87$, $\text{Var}(z_4) = 1.056$

The total variation explained by original variables is

$= \text{Var}(x_1) + \text{Var}(x_2) + \text{Var}(x_3) + \text{Var}(x_4)$

$= 17.02 + 7.56 + 15.75 + 903.87 = 944.20$

The total variation explained by principal components is

$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 916.902 + 18.375 + 7.87 + 1.056 = 944.20$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated. The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = 0.97$$

Continuing, the first two components account for a proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = 0.99$$

of the total variance.

Hence, in further analysis, the first or first two principal components z_1 and z_2 could replace four variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of xi s in equations of zi s. For above data, the first two principal components for first observation, i.e., for the year 1970 can be worked out as

$z_1 = 0.006 \times 25.0 + 0.061 \times 86 + 0.103 \times 66 + 0.993 \times 186.49 = 197.380$

$z_2 = 0.955 \times 25.0 - 0.296 \times 86 + 0.011 \times 66 + 0.012 \times 186.49 = 1.383$

Similarly for the year 1971

$z_1 = 0.006 \times 24.9 + 0.061 \times 84 + 0.103 \times 66 + 0.993 \times 124.34 = 135.54$

$z_2 = 0.955 \times 24.9 - 0.296 \times 84 + 0.011 \times 66 + 0.012 \times 124.34 = 1.134$

Thus, the whole data with four variables can be converted to a new data set with two principal components.

Note: The principal components depend on the scale of measurement, for example, if in the above example x_1 is measured in Fahrenheit instead of Centigrade and x_4 in mm in place of cm, the data gives different principal components when transformed to original

x's. In very specific situations results are the same. The conventional way of getting around this problem is to use standardised variables with unit variances, i.e., correlation matrix in place of dispersion matrix. But the principal components obtained from original variables as such and from correlation matrix will not be same and they may not explain the same proportion of variance in the system. Furthermore, one set of principal components is not simple function of the other. When the variables are standardised, the resulting variables contribute almost equally to the principal components determined from correlation matrix. Variables should probably be standardised if they are measured on scales with widely differing ranges or if measured units do not commensurate with. Often population dispersion matrix or correlation matrix are not available. In such situations sample dispersion matrix or correlation matrix can be used.

Applications of Principal Components

- The most important use of principal components analysis is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then those components' scores can be utilised in subsequent analysis in place of original variables.

- Plotting of data becomes difficult with more than three variables. Through principal components analysis, it is often possible to account for most of the variability in the data by first two components, and it is also possible to plot the values of first two components' scores for each individual. Thus, principal components analysis enables us to plot the data in two dimensions. Particularly, detection of outliers or clustering of individuals will be easier through this technique. Often, use of principal components analysis reveals grouping of variables which would not be found by other means.

- Reduction in dimensionality can also help in analysis where number of variables is more than the number of observations, for example, in discriminant analysis and regression analysis. In such cases, principal components analysis is helpful in reducing the dimensionality of data.

- Multiple regression can be dangerous if independent variables are highly correlated. Principal components analysis is the most practical technique to solve the problem. Regression analysis can be carried out using principal components as regressors in place of original variables. This is known as principal components regression.

Factor Analysis

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable random quantities called, 'factors'. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations. That is, all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. If variables are uncorrelated factor analysis will not be useful. Factor analysis is done in two parts, first solution is obtained by placing some restrictions and then final solution is obtained by rotating this solution. There are two most popular methods available in literature for parameter estimation, the principal components (and the related principal factor) method and the maximum likelihood method. The solution from either method can be rotated in order to simplify the interpretation of factors. It is always prudent to try more than one method of solution. If the factor model is appropriate for the problem at hand, the solution should be consistent with one another. The estimation and rotation methods require iterative calculations that must be done on a computer.

The Orthogonal Factor Model

Let **X** be the observable random vector, with *p* components, has mean μ and covariance matrix **Σ**. The factor model postulates that **X** is linearly dependent upon a few unobservable random variables F_1, F_2, F_3,..., F_m called common factors, and *p* additional sources of variation ε_1,ε_2, ε_3,..., ε_p, called errors or sometimes 'specific factors'. In particular, the factor analysis model is

$$X_1-\mu_1 = l_{11}F_1 + l_{12}F_2 + l\,F_{13} + \ldots + l_{1m}F_m + \varepsilon_1$$

$$X_2-\mu_2 = l_{21}F_1 + 1_{22}F_2 + l_{23}\,F_3 + \ldots + l_{2m}F_m + \varepsilon_2$$

$$X_p-\mu_p = 1_{p1}F_1 + 1p_2F_2 + 1p_3F_3 + \ldots + l_{pm}F_m + \varepsilon_p$$

or, in matrix notation,

$$\mathbf{X-\mu = LF + \varepsilon}$$

Where $\mathbf{X}$ is a matrix of order pxl, $\mathbf{L}$ is of order pxm, $\mathbf{F}$ is of order mxl and $\boldsymbol{\varepsilon}$ is of order px1. The coefficient l_{ij} is called the 'loading of the ith variable on the jth factor, so the matrix $\mathbf{L}$ is the matrix of factor loadings'.

The unobservable random vectors $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ satisfy

$\mathbf{F}$ and ε are independent

$E(\mathbf{F}) = \mathbf{0}$, $\text{cov}(\mathbf{F}) = \mathbf{I}$

$E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}) = \psi$, where ψ is a diagnonal matrix

Covariance structure for the orthogonal factor model

The orthogonal factor model implies a covariance structure for $\mathbf{X}$. It can be seen that $\text{cov}(\mathbf{X}) = \mathbf{LL}' + \psi$ or

$$\text{Var}(X_i) = l_{i1}^2 + l_{i2}^2 + \ldots l_{im}^2 + \psi_i \text{ and } \text{cov}(X_1, X_k) = l_{il}\, l_{kl} + \ldots l_{im}\, l_{km}.$$

That proportion of the variance of the ith variable contributed by the *m* common factors is called the ith 'communality'. That proportion of $\text{var}(X_i) = \sigma_{ii}$ due to the specific factors is often called the 'uniqueness, or specific variance'. Denoting the ith communality by $h_i^2 = l_{i1}^2 + l_{i2}^2 + \ldots l_{im}^2$ and $\sigma_{ii} = h_i^2 + \psi_i$ ($i = 1,2\ldots,p$).

Methods of Estimation

Given the observations $\mathbf{x_1, x_2, \ldots x_n}$ and p generally correlated variables, factor analysis seeks to answer the question, "Does the factor model with a small number of factors, adequately represent the data?" In essence, this statistical model building problem can be tackled by the covariance relationship.

Let the sample covariance matrix $\mathbf{S}$ be an estimator of the unknown population covariance matrix Σ. If the off-diagonal elements of $\mathbf{S}$ are small or those of the sample correlation matrix $\mathbf{R}$ essentially zero, the variables are not related and a factor analysis will not prove useful. In these circumstances, the specific factors play the dominant role, whereas the major aim of the factor analysis is to determine a few important common factors.

If Σ appears to deviate significantly from a diagonal matrix then a factor model can be entertained and the initial problem is one of estimating the factor loadings l_{ij} and specific variances ψ_i. Two most popular methods, viz., the

principal components (and the related principal factor) method and the maximum likelihood method are available in literature for the estimation of factor loadings and specific variances. The solution from either method can be rotated in order to simplify the interpretation of factors. It is always prudent to try more than one method of solution. If the factor model is appropriate for the problem at hand, the solution should be consistent with one another.

The Principal Component (and Principal Factor) Method

Let have eigen value-eigen vector pairs $(\lambda_1, \mathbf{e}_1)$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \ldots \hat{\lambda}_p \geq 0$. Then

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}'_1 + \lambda_2 \mathbf{e}_2 \mathbf{e}'_2 \ldots + \lambda_p \mathbf{e}_p \mathbf{e}'_p$$

$$= \left[\sqrt{\lambda_1}\mathbf{e}_1 \vdots \sqrt{\lambda_2}\mathbf{e}_2 \mathbf{e}'_2 \vdots \ldots \vdots \sqrt{\lambda_p}\mathbf{e}_p \right] \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}'_1 \\ \cdots \\ \sqrt{\lambda_2}\mathbf{e}'_2 \\ \cdots \\ \vdots \\ \cdots \\ \sqrt{\lambda_p}\mathbf{e}'_p \end{bmatrix}$$

This fits the prescribed covariance structure for the factor analysis model having as many factors as variables ($m = p$) and specific variances $\psi_i = 0$ for all i. The loading matrix has jth column given by $\sqrt{\lambda_j}\, e_j$. That is, one can write

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{0} = \mathbf{LL}'$$

In the above equation, it is assumed that the specific factors ε are of minor importance and can also be ignored in the factoring of $\mathbf{\Sigma}$. If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\mathbf{\Sigma} - \mathbf{LL}'$. Since the population parameter $\mathbf{\Sigma}$ is not known, its unbiased estimate $\mathbf{S}$ is considered to obtain principal component solution of the factor model.

The principal component factor analysis of the sample covariance matrix S is specified in terms of its eigen value-eigen vector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \lambda_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3 \ldots \hat{\lambda}_p \geq 0$. Let $m < p$ be the number of common factors. The matrix of estimated factor loadings $\{\tilde{l}_{ij}\}$ is given by

$$\tilde{L} = [\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_m}\mathbf{e}_m]$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S}-\mathbf{LL}'$, so $\tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$. The communalities are estimated as $\tilde{h}_i^2 = \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \cdots \tilde{l}_{im}^2$. The principal components factor analysis of the sample correlation matrix is obtained by starting with R in place of S.

Example: In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The response on a 5-point semantic differential scale was tabulated and the attribute correlation matrix constructed which is given below.

It is clear from the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is "closer" to the (2,5) group than (1,3) group. Observing the results and smaller number of variables, one can expect that the apparent linear relationships between the variables can be explained in terms of, at the most, two or three common factors.

Attribute **Correlation matrix**

		1	2	3	4	5
Taste	**1**	1	.02	.96	.42	.01
Good buy for money	**2**	.02	1	.13	.71	.85
Flavour	**3**	.96	.13	1	.5	.11
Suitable for snack	**4**	.42	.71	.50	1	.79
Provides energy	**5**	.01	.85	.11	.79	1

SPSS Syntax for Principal Components Analysis

```
matrix.

Read x

/file='c: /princomp.dat'

/field= 1 to 80
```

```
/size={17,4}.
compute x1 =ssco(x).
compute x2 =csum(x).
compute x3= nrow(x).
compute x4 = x2/x3.
compute  x5= t(x4)*x4.
compute x6= (x1-(x3*x5)/16.
call eigen (x6,al,a2).
compute x7=x*al.
print x.
print x1.
print x2.
print x3.
print x4.
print x5.
print x6.
print a1.
print a2.
print x7.
end matrix.
```

Factor Analysis (Using SPSS)

File —> New —> Data —> Define Variables —> Statistics —> Data reduction—> Factor —> Variables (mark variables and click ' ▸ ' to enter variables in the box) **Descriptive** Statistics (' √ 'univariate descriptive, ' √ ' initial solution) —> Correlation matrix (' √ 'coefficients) —> **Extraction** —> Method (principal components or maximum likelihood) —> Extract (Number of factors as 2) —> Display (Unrotated factor solution) —> Continue **Rotation** Method (varimax)→ Display (loading plots) —> Continue —> **Scores** —> Display factor score coefficient matrix —> Continue —> OK.