THE CHOICE OF PRINCIPAL VARIABLES FOR COMPUTING HUMAN DEVELOPMENT INDICATORS

TOMSON OGWANG
Department of Economics
University of Northern British Columbia
Regina, Saskatchewan
Canada S4S 0A2

ABDELLA ABDOU
Department of Economics
University of Regina
Regina, Saskatchewan
Canada S4S 0A2

econ@uregina.ca

August 2001

DISCUSSION PAPER #97
THE CHOICE OF PRINCIPAL VARIABLES FOR COMPUTING HUMAN DEVELOPMENT INDICATORS

By

Tomson Ogwang and Abdella Abdou*
Economics Program
University of Northern British Columbia
3333 University Way
Prince George, British Columbia
V2N 4Z9

*The authors wish to thank the UNDP Human Development Report Office for providing the unpublished data for computing the components of GDI and GEM. Thanks are also extended to C. Opio, U.L.G. Rao and R. Wimalasena for their valuable comments. However, the usual disclaimer applies.
Abstract
The present paper investigates the choice of principal variables for computing three human development indicators, namely, the Human Development Index (HDI), the Gender-Related Development Index (GDI), and the Gender Empowerment Measure (GEM). To this end, the principal components variable selection strategy considered by Jolliffe (1972, 1973) is applied to data from the 1999 Human Development Report. The empirical results show that there is statistical justification for selecting only one of the three components of each indicator. In the case of the HDI, the life expectancy index turns out to be the best choice; in the case of the GDI the best choice is the equally distributed educational attainment index; and in the case of the GEM the best choice is the index of parliamentary representation. Accordingly, it seems reasonable to compute simpler HDI, GDI and GEM based solely on the selected indicators without loss of too much information. Also, while our results support the current practice of equally weighting the three components of HDI and GDI, this is not the case for the GEM.
1. Introduction

As evidenced by a spate of recent papers on the subject, there is considerable interest in the concept and measurement of human development. See, for example, Hopkins (1991), McGillivray (1991), Dasgupta and Weale (1992), McGillivray and White (1993), Srivastavan (1994), Streeten (1994), and Pillarisetti and McGillivray (1998), among others. There are several plausible explanations for this. First, it is now widely recognized that GDP (or GNP) per capita is not a fully satisfactory measure of human well-being and that composite indicators, that incorporate more than one attribute of human well-being, are conceptually more appealing. Second, the popularization of the Basic Human Needs (BHN) approach to development in the 1970s led to the search for BHN performance indicators, which have facilitated the construction of human development indicators.¹ The BHN approach emphasizes the satisfaction of basic needs such as good nutrition, health, education, clothing and sanitation. In contrast, the traditional growth strategies emphasize increasing GDP (or GNP) per capita. Third, the annual publication of the Human Development Report (HDR) by the United Nations Development Programme (UNDP) since 1990 has raised awareness of and focused debate on the

¹ Some measures of BHN performance proposed by Hicks and Streeten (1979) have been incorporated in the HDI.
importance of the concept of human development. In fact, over the past decade the UNDP has taken
the lead in conceptualizing and measuring the conditions of human development.

In the inaugural issue of the HDR, the UNDP(1990) proposed the Human Development Index
(HDI) which is a composite index comprising of life expectancy, adult literacy, and real GDP per capita
adjusted for purchasing power parity. The index ranges from zero to one with higher values signifying
higher levels of human development. The HDI is rationalized on the grounds that human beings strive to
lead long healthy lives (captured by life expectancy), to have a descent standard living (captured by real
GDP per capita), and to acquire knowledge (captured by adult literacy), in which case the focus of
development policy should be to expand the choices and capabilities of people in these areas. The HDI
is increasingly being used as a basis for setting human development goals of many countries and more
than 100 countries are already constructing national or sub-national HDIs.

In response to some of the criticisms of the earlier versions of HDI, the UNDP has continued to
refine the underlying analytical framework. To this end, several improvements have been made in the
methodology and quality of data used for constructing the HDI. The UNDP (1995) introduced two
additional measures of human development, namely, the Gender-Related Development index (GDI) and
the Gender Empowerment Measure (GEM). As we shall see in section 2 below, the GDI uses the same
dimensions as the HDI except for the fact that the former takes into account gender disparities in human
development. Specifically, the GDI is constructed such that countries that exhibit greater gender
disparities in human development are penalized by achieving lower GDI values relative to their HDI
counterparts. The GEM is a measure of the degree of women’s participation in the political, economic,
and professional activities. It must be pointed out that the GDI and GEM, like the HDI, lie between
zero and one with greater values signifying higher levels of human development.

The HDI, GDI and GEM have the common characteristic of incorporating more than one attribute of human well-being. Although the addition of extra attributes may be appealing on conceptual grounds, it also introduces new complications with respect to the assignment of weights to the attributes, and the added time and monetary costs of obtaining information on all attributes, among others.\(^2\) Even the proponents of the HDI, GDI and GEM concede that it is impossible to come up with a comprehensive set of human development indicators. See, for example, UNDP (1999, p. 127). In fact, as pointed out by UNDP (1994, p. 91), more indicators is not necessarily better in the sense that there might be an overlap among some indicators as is the case with infant mortality which is already reflected in life expectancy. The fact that it is impossible to come up with a single composite indicator of human development that incorporates all attributes of human well-being raises the question as to whether or not there is statistical justification for including all dimensions of a particular composite indicator of human well-being. In our opinion, it is important to incorporate only attributes that add new information. Accordingly, the most parsimonious, yet informative, composite indicators of human development are recommended.

\(^2\) For a discussion of the difficulties in measuring the literacy component of HDI, see Hopkins (1992, p. 1471).
In a recent study of the HDI, Ogwang (1994) found that there is statistical justification for retaining only one of the three components of HDI. It must be noted that Ogwang’s study was based on data for an earlier version of HDI obtained from the 1991 HDR. Given recent refinements in the methodology for computing the HDI as well as improvements in the quality of the underlying data, it seems reasonable to re-examine this issue using the most recently available and more reliable data for the HDI. Furthermore, to the best of our knowledge, no empirical studies have hitherto examined the issue of choice of principal variables for computing the GDI and GEM. The present paper fills in the gap by examining the choice of principal variables for computing HDI, GDI, and GEM. To this end, the principal components variable selection strategy considered by Jolliffe (1972,1973) is used to analyze data from the 1999 HDR.

The format of the rest of the paper is as follows. In section 2, the methodologies for constructing the HDI, GDI and GEM are briefly discussed. In section 3, the variable selection strategy employed is briefly described. In the penultimate section, the empirical results are presented. The concluding remarks are made in the final section.

2. Construction of the HDI, GDI and GEM

The most recent versions of HDI, GDI and GEM, reported in the 1999 HDR, are each computed as simple arithmetic means of three sub-components. The HDI is computed as a simple arithmetic mean of the life expectancy index (LEI), educational attainment index (EAI), and an index of real GDP per capita adjusted for purchasing power parity (GDPI). Prior to computing the mean, each of the three components is scaled to lie between zero and one. For further details concerning the
construction of the most recent version of HDI, see UNDP (1999, pp. 159-160).

The GDI is computed as a simple arithmetic mean of an equally distributed life expectancy index (ELEI), an equally distributed educational attainment index (EEAI), and an equally distributed income index (EGDPI). Prior to computing the mean, each index is scaled to lie between zero and one. The equally distributed components take into account differential life expectancies, educational attainments and incomes of men and women as well as their differential population shares. For further details concerning construction of the most recent version of GDI, see UNDP (1999, pp. 160-161).

The GEM is computed as a simple arithmetic mean of an index of parliamentary representation (PRI), index of administrative and managerial, and professional and technical positions (AMPTPI), and an equally distributed income index (EGDPI). Prior to computing the mean, each index is scaled to lie between zero and one. For further details concerning construction of the most recent version of GEM, see UNDP (1999, pp. 161-162).

Note that taking the arithmetic mean of the three components of each indicator amounts to assigning equal weights to these components, implying that they are equally important in that indicator. In the sections below, we will examine the issue of which component best represents the three components of each of the three human development indicators (HDI, GDI and GEM).

3. The selection strategy

Principal components analysis (PCA) technique has traditionally been used to transform a large set of correlated variables into a smaller set of uncorrelated variables, called the principal components, that account for most of the variation in the original set of variables. Since the principal components are
linear combinations of the original variables with mathematically determined characteristic vectors of the covariance (or correlation) matrix of original variables as weights, it can be argued that PCA resolves the problem of arbitrary choice of weighting scheme. The first principal component accounts for the largest proportion of variation in the original set of variables. The second principal component captures the largest proportion of variation which is not accounted for by the first principal component, and so on. For a set of $k$ variables, the maximum number of principal components which can be extracted is equal to $k$. However, if the $k$ variables are highly correlated, only a few principal components capture most of the variation. The proportion of variation attributed to a particular principal component is obtained by dividing the associated characteristic root by the sum of all the characteristic roots.

Dunteman (1989) provides a good introduction to PCA.

In principle, the high inter-correlations among the components of composite indicators could be exploited to construct new composite indicators which are, in fact, first principal components of the original components and then scaling them to lie within the desired range. A drawback of this approach is that information on all the components is required since each principal component is a linear combination of all the constituent variables.\(^3\)

Recently, PCA has also served as a useful tool for selecting a few variables among a wider set of correlated variables. The technique used in the present paper is PCA method B4, described by

---

\(^3\) See, for example, Ram (1982) and UNDP (1993, pp. 109-110).
Jolliffe (1972, p. 164; 1973, p. 22), which is recommended by Jolliffe (1972, p. 168) in cases where retention of best subsets of variables is considered important, which is obviously the case with the present study. Essentially, the method involves first selecting the variable which has the highest correlation with the first principal component. This is followed by the variable which has the highest correlation with the second principal component, where the second stage selection process is limited only to those variables that are discarded in the first stage selection process. This procedure is continued until the required number of variables has been selected. Jolliffe (1972, p. 171) recommended that the number of variables to be selected should be equal to the number of principal components associated with characteristic roots of the correlation matrices of variables which are greater than 0.70. If the covariance matrix of variables is used, the number of variables to be selected is equal to the number of characteristic roots of the covariance matrix which are greater than 0.7\(\bar{\text{r}}\), where \(\bar{\text{r}}\) is the average of the characteristic roots associated with the covariance matrix as suggested in the principal components literature. See Dunteman (1989, p. 22) for further details.

4. Empirical Results

The empirical results presented in this section are based on the most recent and best available data taken from the 1999 HDR. The data for the HDI relate to 174 countries whereas those for the GDI and GEM pertain to 143 and 102 countries, respectively. The choice of the number of countries is dictated by the availability of data as reported in the 1999 HDR.

To provide some insights into the degree of association among the three possible pairs of the three components of each of HDI, GDI, and GEM, the covariances, Pearson's product moment
correlation coefficients and the Spearman's rank correlation coefficients are reported in Table 1. Evidence of strong positive correlation among the three components of each human development indicator is apparent from the table. The lowest inter-correlations are exhibited by the three GEM components.

Table 1 here

Since the selection strategy employed here requires knowledge of characteristic roots and characteristic vectors of the covariance matrices of the three components of each indicator, the covariances and the associated characteristic roots and vectors of the three HDI components are reported in Table 2. Similar data for the GDI and GEM are reported in Table 3 and Table 4, respectively.\textsuperscript{4} It can be seen from the entries in Table 2 that the first principal component accounts for approximately 85 percent of the variation in the three components of HDI. The first two principal components account for approximately 94 percent of the variation in the three variables. Clearly, the

\textsuperscript{4}The present study uses the covariance matrix of variables rather than the correlation matrix since the units of measurement of the three components of each indicator are reasonably commensurate, as suggested in the principal components literature (for example, Morrison, 1967, p. 223).
first principal component contains most of the statistical information embedded in the three components of HDI.

The components of the characteristic vector associated with the first principal component of the three HDI components show that the first principal component HDI series could be computed as 0.564 (LEI) + 0.620 (EAI) + 0.546 (GDPI). However, this necessitates data for computing all the three components, which can be expensive to collect. Furthermore, since the components of the characteristic vector associated with the first principal component do not sum to unity, the first principal component HDI series will not lie between zero and one without further transformation. Another notable feature of Table 2 is that the first principal component weights attached to the three HDI components are approximately equal which is consistent with the equal weighting scheme adopted by the UNDP in the estimation of the HDI in the 1999 HDR. It is also easy to verify from the entries in Table 2 that only one characteristic root is greater than 0.70, suggesting that only one of the three variables be retained for purposes of computing a simpler HDI.

Table 2 here

Since the decision as to which of the three variables should be retained requires information on the correlations (or loadings) between the variables and the principal components, these correlations are also reported in Table 2. It can be seen from the entries in the table that the life expectancy index (LEI) has the highest correlation with the first principal component. Accordingly, it can be concluded, based

5 For the weights to sum to unity, the weight attached to each indicator must be divided by the sum of all the weights in each case.
on the aforementioned selection criterion, that the life expectancy index (LEI) is the best choice. Thus, a simpler HDI series based solely on the LEI would be cheaper to compute without loss of too much information. In fact, the product moment correlation and rank correlation coefficients between the HDI and LEI are 0.934 and 0.944, respectively, which are very close to one indicating near perfect positive correlation.

Table 3 shows the results of PCA of the three GDI components. Four points about the table are in order. First, the first principal component accounts for approximately 86 percent of the variation in the three components of GDI. The first two principal components account for approximately 94 percent of the variation in the three variables. Just as in the HDI, the first principal component contains most of the statistical information embedded in the three GDI components. Second, the components of the characteristic vector associated with the first principal component of the three GDI components show that the first principal component GDI series could be computed as 0.542 (ELEI) + 0.644 (EEAI) + 0.539 (EGDPI). However, this also necessitates data for computing all the three components, which is an expensive undertaking. Furthermore, since the components of the characteristic vector associated with the first principal component do not sum to unity, the first principal component GDI series will not lie between zero and one without further transformation. Third, the first principal component weights attached to the three GDI components are approximately equal which is consistent with the equal weighting scheme adopted by the UNDP in the estimation of the GDI in the 1999 HDR. Fourth, it is also easy to verify from the entries in Table 3 that only one characteristic root is greater than 0.70 suggesting that only one of the three variables be retained for purposes of computing a simpler GDI. An inspection of the correlations between the variables and the principal components and application of the
aforementioned selection criterion indicates that the selected variable should be the equally distributed educational attainment index (EEAI). Thus, a simpler GDI series based solely on EEAI would be cheaper to compute without loss of too much information. In fact, the product moment correlation and rank correlation coefficients between the GDI and EEAI are 0.924 and 0.907, respectively, which are very close to one indicating near perfect positive correlation. We also note that had the EEAI been used as the sole indicator for constructing the GDI in the 1999 HDR, the number of countries ranked would have increased from 143 to 145.

Table 3 here

Table 4 shows the results of PCA of the three GEM components. Four points about the table are worthy of mention. First, the first principal component accounts for approximately 65 percent of the variation in the three GEM components. The first two principal components account for approximately 85 percent of the variation in the three variables. As was the case with the HDI and GDI, the first principal component contains most of the statistical information embedded in the three GEM components. Second, the components of the characteristic vector associated with the first principal component of the three GEM components show that the first principal component GEM series could be computed as 0.786 (PRI) + 0.388 (AMPTPI) + 0.481 (EGDPI). However, this also necessitates data for computing all the three components, which can be expensive to collect. Furthermore, since the components of the characteristic vector associated with the first principal component do not sum to unity, the first principal component GEM series will not lie between zero and one without further transformation. Third, unlike in the cases of the HDI and GDI, the first principal component weights attached to the three GEM components are not approximately equal which is not in conformity with the
equal weighting scheme adopted by the UNDP in the estimation of the GEM in the 1999 HDR. Fourth, it is also easy to verify from the entries in Table 4 that only one characteristic root is greater than 0.70 suggesting that only one of the three variables be retained for purposes of computing a simpler GEM. An inspection of the correlations between the variables and the principal components and application of the aforementioned selection criterion shows that the selected variable should be the index of parliamentary representation (PRI). Thus, a simpler GEM series based solely on the PRI would be cheaper to compute without loss of too much information. In fact, the product moment correlation and rank correlation coefficients between the GEM and PRI are 0.857 and 0.840, respectively, which are very close to one indicating near perfect positive correlation. We also note that had the PRI been used as the sole indicator for constructing the GEM in the 1999 HDR, the number of countries ranked would have increased from 102 to 164.

Table 4 here

5. Concluding Remarks

In this paper, the principal components variable selection strategy considered by Jolliffe (1972, 1973) is used to investigate the choice of principal variables for computing three human development indicators, namely, the HDI, GDI and GEM using data reported in the 1999 HDR. The life expectancy index (LEI) turns out to be the best choice for constructing a simplified HDI; in the case of the GDI, the equally distributed educational attainment index (EEAI) turns out to be the best choice; and in the case of the GEM the index of parliamentary representation (PRI) is the best choice. Furthermore, the
correlation between each composite indicator and the best variable for constructing that indicator is very
close to one. Accordingly, it seems reasonable to compute simpler HDI, GDI and GEM series based
solely on the selected indicators without loss of too much information. Our results also bring into
question the current practice of equally weighting the three GEM components. Finally, it must be
pointed out that our discussion concerning the choice of indicators is based purely on the analysis of the
internal structures of the data and do not preclude the use of other criteria for inclusion of other variables
to these indicators.
References


Table 1
Covariances and correlations among the three components of HDI, GDI and GEM

<table>
<thead>
<tr>
<th>Development indicator</th>
<th>Covariance</th>
<th>Pearson’s correlation</th>
<th>Spearman’s correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDI</td>
<td>LEI and EAI</td>
<td>0.029</td>
<td>0.779</td>
</tr>
<tr>
<td></td>
<td>LEI and GDPI</td>
<td>0.027</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>EAI and GDPI</td>
<td>0.027</td>
<td>0.73</td>
</tr>
<tr>
<td>GDI</td>
<td>ELEI and EEAI</td>
<td>0.03</td>
<td>0.776</td>
</tr>
<tr>
<td></td>
<td>ELEI and EGDPI</td>
<td>0.027</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>EEAI and EGDPI</td>
<td>0.03</td>
<td>0.761</td>
</tr>
<tr>
<td>GEM</td>
<td>PRI and AMPTPI</td>
<td>0.018</td>
<td>0.477</td>
</tr>
<tr>
<td></td>
<td>PRI and EGDPI</td>
<td>0.019</td>
<td>0.438</td>
</tr>
<tr>
<td></td>
<td>EGDPI and AMPTPI</td>
<td>0.01</td>
<td>0.319</td>
</tr>
</tbody>
</table>
Table 2
Results of PCA of the three HDI components using their covariance matrix

<table>
<thead>
<tr>
<th></th>
<th>Principal component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
</tr>
<tr>
<td>Characteristic root</td>
<td>0.090849</td>
</tr>
<tr>
<td>Percentage of variance explained individually</td>
<td>84.8</td>
</tr>
<tr>
<td>Percentage of variance explained cumulatively</td>
<td>84.8</td>
</tr>
<tr>
<td>Characteristic vectors*</td>
<td></td>
</tr>
<tr>
<td>LEI</td>
<td>0.564</td>
</tr>
<tr>
<td>EAI</td>
<td>0.620</td>
</tr>
<tr>
<td>GDPI</td>
<td>0.546</td>
</tr>
<tr>
<td>Correlation coefficient with*</td>
<td></td>
</tr>
<tr>
<td>LEI</td>
<td>0.931</td>
</tr>
<tr>
<td>EAI</td>
<td>0.922</td>
</tr>
<tr>
<td>GDPI</td>
<td>0.908</td>
</tr>
</tbody>
</table>

* LEI, EAI, and GDPI denote the life expectancy index, the educational attainment index, and the index of real GDP per capita adjusted for purchasing power parity, respectively.
Table 3  
Results of PCA of the three GDI components using their covariance matrix

<table>
<thead>
<tr>
<th></th>
<th>Principal component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
</tr>
<tr>
<td>Characteristic root</td>
<td>0.09656</td>
</tr>
<tr>
<td>Percentage of variance explained individually</td>
<td>85.5</td>
</tr>
<tr>
<td>Percentage of variance explained cumulatively</td>
<td>85.5</td>
</tr>
<tr>
<td>Characteristic vectors*</td>
<td></td>
</tr>
<tr>
<td>ELEI</td>
<td>0.542</td>
</tr>
<tr>
<td>EEAI</td>
<td>0.644</td>
</tr>
<tr>
<td>EGDPI</td>
<td>0.539</td>
</tr>
<tr>
<td>Correlation coefficient with*</td>
<td></td>
</tr>
<tr>
<td>ELEI</td>
<td>0.922</td>
</tr>
<tr>
<td>EEAI</td>
<td>0.934</td>
</tr>
<tr>
<td>EGDPI</td>
<td>0.915</td>
</tr>
</tbody>
</table>

* ELEI, EEAI, and EGDPI denote the equally distributed life expectancy index, the equally distributed educational attainment index, and the equally distributed income index, respectively.
Table 4
Results of PCA of the three GEM components using their covariance matrix

<table>
<thead>
<tr>
<th>Characteristic root</th>
<th>Principal component</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.075832</td>
<td>0.0241</td>
<td>0.017712</td>
</tr>
<tr>
<td>Percentage of variance explained individually</td>
<td>64.5</td>
<td>20.4</td>
<td>15.1</td>
<td></td>
</tr>
<tr>
<td>Percentage of variance explained cumulatively</td>
<td>64.5</td>
<td>84.9</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Characteristic vectors*</td>
<td>PRI</td>
<td>0.786</td>
<td>-0.450</td>
<td>-0.424</td>
</tr>
<tr>
<td></td>
<td>AMPTPI</td>
<td>0.388</td>
<td>-0.174</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>EGDPI</td>
<td>0.481</td>
<td>0.876</td>
<td>-0.038</td>
</tr>
<tr>
<td>Correlation coefficient with*</td>
<td>PRI</td>
<td>0.924</td>
<td>-0.297</td>
<td>-0.241</td>
</tr>
<tr>
<td></td>
<td>AMPTPI</td>
<td>0.655</td>
<td>-0.165</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>EGDPI</td>
<td>0.698</td>
<td>0.716</td>
<td>-0.027</td>
</tr>
</tbody>
</table>

* PRI, AMPTPI, and EGDPI denote the index of parliamentary representation, the index of administrative and managerial, and technical positions, and the equally distributed income index, respectively.