Since 1994 Inter-University Consortium
ALMALAUREA
Connecting Universities, the Labour Market and Professionals

AlmaLaurea Working Papers – ISSN 2239-9453

ALMALAUREA WORKING PAPERS no. 11

September 2011

The estimation of Human Capital in structural models with flexible specification

by

Pietro Giorgio Lovaglio, Giuseppe Folloni
University of Bicocca-Milan
University of Trento

This paper can be downloaded at:

AlmaLaurea Working Papers series
http://www.almalaurea.it/universita/pubblicazioni/wp

Also available at:

REsearch Papers in Economics (RePEC)

The **AlmaLaurea working paper series** is designed to make available to a wide readership selected works by AlmaLaurea staff or by outside, generally available in English or Italian. The series focuses on the study of the relationship between educational systems, society and economy, the quality of educational process, the demand and supply of education, the human capital accumulation, the structure and working of the labour markets, the assessment of educational policies.
Comments on this series are welcome and should be sent to pubblicazioni@almalaurea.it.

AlmaLaurea is a public consortium of Italian universities which, with the support of the Ministry of Education, meets the information needs of graduates, universities and the business community. AlmaLaurea has been set up in 1994 following an initiative of the Statistical Observatory of the University of Bologna. It supplies reliable and timely data on the effectiveness and efficiency of the higher education system to member universities' governing bodies, assessment units and committees responsible for teaching activities and career guidance.

AlmaLaurea:

- facilitates and improves the hiring of young graduates in the labour markets both at the national and international level;
- simplifies companies' search for personnel, reducing the gap between the demand for and supply of qualified labour (www.almalaurea.it/en/aziende/);
- makes available online more than 1.5 million curricula (in Italian and English) of graduates, including those with a pluriannual work experience (www.almalaurea.it/en/);
- ensures the optimization of human resources utilization through a steady updating of data on the careers of students holding a degree (www.almalaurea.it/en/lau/).

Each year AlmaLaurea plans two main conferences (www.almalaurea.it/en/informa/news) in which the results of the annual surveys on Graduates' Employment Conditions and Graduates' Profile are presented.

AlmaLaurea Inter-University Consortium | viale Masini 36 | 40126 Bologna (Italy)
Website: www.almalaurea.it | E-mail: pubblicazioni@almalaurea.it

---

The opinions expressed in the papers issued in this series do not necessarily reflect the position of AlmaLaurea

© AlmaLaurea 2011
Applications for permission to reproduce or translate all or part of this material should be made to:
AlmaLaurea Inter-University Consortium
email: pubblicazioni@almalaurea.it | fax +39 051 6088988 | phone +39 051 6088919

International Conference on
"Human Capital and Employment in the European and Mediterranean Area"
Bologna, 10-11 March 2011

The estimation of Human Capital in structural models with flexible specification

by

P.G. Lovaglio [*], G. Folloni[**]

Abstract

The present paper focuses on statistical models for estimating Human Capital (HC) at disaggregated level (worker, household, graduates). The more recent literature on HC as a latent variable states that HC can be reasonably considered a broader multi-dimensional non-observable construct, depending on several and interrelate causes, and indirectly measured by many observed indicators. In this perspective, latent variable models have been assuming a prominent role in the social science literature for the study of the interrelationships among phenomena. However, traditional estimation methods are prone to different limitations, as stringent distributional assumptions, improper solutions, and factor score indeterminacy for Covariance Structure Analysis and the lack of a global optimization procedure for the Partial Least Squares approach. To avoid these limitations, new approaches to structural equation modelling, based on Component Analysis, which estimates latent variables as exact linear combinations of observed variables minimizing a single criterion, were proposed in literature. However, these methods are limited to model particular types of relationship among sets of variables. In this paper, we propose a class of models in such a way that it enables to specify and fit a variety of relationships among latent variables and endogenous indicators. Specifically, we extend this new class of models to allow for covariate effects on the endogenous indicators. Finally, an application aimed to measure, in a realistic structural model, the causal impact of formal Human capital (HC), accumulated during Higher education, on the initial earnings for University of Milan (Italy) graduates is illustrated.

[*] Dept. of Quantitative Methods, University of Bicocca-Milan
e-mail: piergiorgio.lovaglio@unimib.it
[**] Department of Economics, University of Trento

1. Introduction

The concept of human capital (henceforth HC) is an old one. Perhaps the first to try to define and measure what we now call HC was Sir William Petty (Petty, 1690).

Since the time of Petty, distinguished statisticians, actuaries, econometricians, and economists have discussed the concept of HC. The second half of the twentieth century shows a major revival of interest in HC, starting with the contributions of Schultz (1961), Mincer (1970) and Becker (1975), who brought to the fore the research interest in this subject. Mincer bypassed the estimation problem by specifying an earnings function (resulting from the reduced form of HC equations for individuals with different years of schooling) as the outcome of mathematically convenient and controversial counterfactual assumptions, such as (*i*) the equality of the HC present values of future earnings at the time training begins and (*ii*) the constancy, by years of education, of their respective flow of earnings throughout their working lives. Substantially, these authors consider the length of school training and the total cost of post-school investment in training, health, and mobility as the main sources of the stock of human capital. However, these authors undertook a thorough study of the concept of HC and analyzed the main forces that contribute to its formation and accumulation. However, they did not engage in any quantitative estimation of it.

Instead, many authors have proposed different approaches to measure the 'quantity' of HC embodied in individuals, and to determine its sources. For example, HC in educational attainment, retrospective and prospective methods, has been measured as an aggregate indicator by means of numerous different variables (World Bank, 1995; United Nations, 2002; Wossmann, 2003). These variables are related to different dimensions: educational attainment (schooling, training while working, experience, and educational investment costs), non-cognitive skills, parental and family characteristics and earnings.

However, the OECD report (1998, p.9) defines HC as 'the knowledge, skill, competencies and attributes embodied in individuals that are relevant to economic activity'. HC is a complex, multifaceted phenomenon with various intangible dimensions that are not directly observable and cannot be measured with precision by a single attribute, a set of attributes, or their combined sum on individuals or households.

Statistically speaking, HC is a latent variable (LV). On this basis, Dagum and Slottje (2000) have defined household HC at a micro-economic level as the multidimensional non-observable construct generated by personal ability, home and social environments, investments in the education of the household head and spouse whose effects are indirectly measurable by means of the present value of a flow of earned income throughout an individual's life span.

However, an LV has been defined in various ways in the literature. Structural equation models are used to specify and test hypothesized relationships among manifest variables and latent variables in multivariate data. Two different approaches have been proposed for Structural Equation Models: Covariance Structure Analysis (CSA, Jôreskog, 1970) and Partial Least Squares Path Modelling (PLSPMWold, 1982). Covariance Structure Analysis and PLSPM deal with structural equation models from statistically different perspectives.

In the former, a variable is termed 'latent' if the equations cannot be manipulated so that they can be expressed as a function of manifest/observable variables (Bentler, 1992). Therefore, an LV is seen as a latent cause of manifest indicators and accounts for their variance in a measurement model (typically a factor model). In the latter, within the Component Analysis (CA, Meredith and Millsap, 1985) framework, latent variables are equivalent to (latent) components (LC), extracted by their manifest variables (Velicer and Jackson 1990; Schonemann and Steiger, 1976).

Depending on the perspective, manifest variables may be considered either as the effects of an underlying LV or as the causes defining the LC. Thus, each measurement model may identify either a "formative scheme" (the manifest variables contribute to the formation of the LC scores, following a Principal Component Analysis framework) or a "reflective scheme" (the manifest variables constitute the effects of their LVs, following a Factor Analysis framework).

In CA framework, where formative (reflective) variables are called exogenous (endogenous) variables, the relationship between the exogenous and the manifest endogenous variables are moderated by the presence of LC: LC are obtained as exact linear combinations of formative indicators, and model parameters are estimated by consistently minimizing a single criterion.

Covariance Structure Analysis is the most used model for structural relations with LV, but its drawbacks (non uniqueness of LV scores, normality distribution of LV, strong error hypotheses, lack of sufficient conditions for model identification) has already been shown in depth in previous works (see Vittadini 1989 for a review).

PLSPM is a model, proposed in alternative to the CSA-based models, based on soft hypotheses. PLSPM provides the estimate of structural parameters in a second stage, by using the scores of the LC achieved in the first stage as Wold says "by deliberate approximation as linear aggregation (proxy) of its manifest indicators" (Wold, 1982). In this way, the structural equations become a Path Analysis between sets of estimated linear combinations.

In the Dagum and Slottje (2000) approach, the HC was estimated as a linear combination of several manifest (formative) indicators connected with a person's characteristics and investment to acquire abilities made by the individual or his/her family. To obtain solutions of HC scores the authors choose the PLSPM approach.

Despite a number of benefits of PLSPM for fitting structural equation models, the lack of a global optimization criterion (PLSPM solutions are not optimal in an overall fit) and the logic inconsistence of the algorithm seems to make its use limited (McDonald, 1996). To this end, the method proposed by Dagum and Slottje (2000) can be improved to obtain solutions more consistent with the HC economic theoretical framework (Dagum *et al*., 2007; Vittadini and Lovaglio 2007). Moreover, it can be extended to more complex and realistic situations. To this end, in applications, often the underlying theory may specify the presence in the model of exogenous covariates (e.g. wealth) that do not strictly belong to the formative blocks of LCs, but may have a causal impact on manifest endogenous variables. Statistically speaking, these manifest exogenous factors (directly linked with the reflective indicators of an LC, without being embedded in its formative block) are called concomitant indicators.

In this paper, we propose a new estimation method, within the Component analysis framework, for HC estimation that consistently allows concomitant indicators in the structural model. The proposed algorithm estimates LC as linear combinations of manifest variables providing a global fitting criterion, allowing fitting diverse complex relationships among variables, including direct effects of manifest variables (concomitant indicators) that do not "form" LC, but affect endogenous variables.

The remaining sections of this paper are organized as follows. Section 2 discusses available statistical methods for LC estimation and explains methodological drawbacks of PLSPM. In Section 3, we briefly describe a recent approach, called Extended Redundancy Analysis (ERA) model and in Section 4 we present the estimation procedure allowing concomitant indicators. Section 5 discusses the approach and in Section 6 an example, focused on human capital estimation for University of Milan (Italy) Graduates, is presented for illustration. Section 7 concludes the paper.

2. Statistical methods

In Structural equation modelling, when the number of variables is very large, as well as, in presence of more than one sets of them playing a logical asymmetrical role (explanatory and response variables), it may be advantageous to find for each set a linear combination of variables (latent variables) having some properties in terms of correlation, covariance or variance. When the goal is to predict a set of dependent variables as well as possible in terms of least square error, an appropriate model is Reduced Rank Regression (RRR, Izenman, 1975). In general, when the goal is to predict more dependent variables by substituting the set of manifest explanatory variables with a fewer sequence of orthogonal latent variables, Dimensional Reduction Methods (DRM, Abraham and Merola, 2001) should be applied. The commonly used DRM methods are Principal Component Regression, Canonical Correlation Regression, RRR and Partial Least Squares regression (see Abraham and Merola, 2001). In presence of more than two blocks of observable indicators, PLSPM is a viable alternative approach. However, these approaches do not support concomitant indicators. To show the iterative steps of PLSPM algorithm, let Figure 1 denote a simple structural equation model involving two vectors of LCs ($\boldsymbol{\eta_j}$), an exogenous $\boldsymbol{\eta_2}$ and an endogenous $\boldsymbol{\eta_1}$ (for the sake of simplicity unidimensional), underlying a block of p and q manifest indicators $\mathbf{V}=(\mathbf{v_1},..,\mathbf{v_p})$ and $\mathbf{Y}=(\mathbf{y_1},\ldots,\mathbf{y_q})$, respectively:

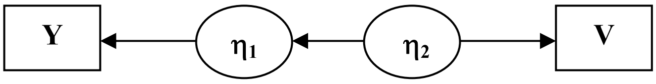

Figure. 1: Path diagram of a simple structural model.

PLSPM defines and estimates each LC as a weighted combination of its manifest variables ($\boldsymbol{\eta_1}= \mathbf{Yw_1}$, $\boldsymbol{\eta_2} =\mathbf{Vw_2}$). However, we will show that the vectors of weights ($\mathbf{w_j}$,=1,2) that define the LC's scores are performed with arbitrariness by means of regressions which do not respect the role of the manifest variables in the structural model.
PLSPM proceeds in two stages (Wold, 1982). In the first stage, the LC scores are estimated as linear combinations of their own manifest indicators by means of a cyclical procedure. After having initially estimated each LC ($\boldsymbol{\eta_{0j}}$) as a linear combination of its manifest indicators, using arbitrary weights ($\mathbf{w_{0j}}$), LC scores are estimated by means of an *iterative algorithm* with respect to both the inner and outer model. The inner model updates $\boldsymbol{\eta_{0j}}$ (inner estimate, η_{0j}^+) as a "weighted sum" of the LCs which, in the structural model, are adjacent to $\boldsymbol{\eta_{0j}}$ (previously estimated by the same criteria used for $\boldsymbol{\eta_{0j}}$). The sum of the adjacent LCs is weighted by the coefficients representing the correlation between the adjacent LCs (factorial scheme) or simply their signs (centroid scheme) or the regression coefficients, following the specified structural model (structural scheme). Differently, in the outer model, the weights defining LC scores are updated depending on the causal links between latent variables and their manifest indicators: for reflective indicators (Figure 1), weights are obtained by simple regression of each manifest indicator on $\boldsymbol{\eta_{0j}}^+$, whereas for formative indicators (reversing the arrows linking observed variables and LCs in Figure 1), the weights are updated by multiple regression of $\boldsymbol{\eta_{0j}}^+$ onto the matrix of manifest indicators. By iterating the steps of this procedure, the convergence of the algorithm presents the final estimation of the weights defining the LCs' scores as linear combinations of their manifest indicators.
The application of these steps is not consistent with the structural model specified in Figure 1 because: i) the initial estimates of $\boldsymbol{\eta_j}$ (η_{0j}) are performed by treating the manifest variables as *causes*, whereas these are *indicators* (effects) of their LCs; ii) since we have only two LCs, the inner estimates for $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ (η_{01}^+, η_{02}^+) are equal, for each adopted scheme; iii) in the outer model, the weights are updated by single regressions of $\mathbf{v_i}$ and $\mathbf{y_i}$ on the the inner estimates of the LCs, whereas the manifest variables are indicators only of their corresponding LCs.
Further, in the simplest case when the structural model involves only a single LC (e.g $\boldsymbol{\eta_1}$) measured by reflective indicators, if we estimate $\boldsymbol{\eta_1}$ as a linear aggregation of its manifest indicators ($\mathbf{y_i}$) we have two logic inconsistencies: firstly, this approach is not consistent with the logical linking between $\boldsymbol{\eta_1}$ and its manifest variables (since $\boldsymbol{\eta_1}$ is a underlying cause of any $\mathbf{y_i}$ endogenous variable) and secondly, it does not allow the specification of a true measurement model, because it results in unit *R*-squared. To illustrate this form of inconsistency we mention the above approach of Dagum and Slottje (2000).
The theoretical framework for HC estimation rested on the central role played by HC in the Income Generating Function (IGF, Dagum, 1994), a revised version of the human capital production function. Specifically, Dagum (1994) specifies, in the IGF equation, that earned income (y_i) for the *i*th economic unity, is a function of its stock of HC (HC_i) and non-human capital or wealth (k_i) plus a random term (e_i)

$$y_i = \rho HC_i + \tau k_i + e_i \quad (1)$$

where ρ and τ correspond to the rate of return of HC and wealth, respectively.
The authors estimated equation (1) in a regression equation, once the scores of HC_i were previously estimated with PLSPM as linear combination of formative indicators **F** (including wealth). Since the algorithm converged on a solution for the HC coincident with the first principal component of manifest indicators (Wold, 1982), this approach only partially captures the economic definition of HC. In fact, HC is estimated by considering only its formative indicators **F** (corresponding to the HC retrospective definition) without embedding in the measurement model the earned income variable and thus the effects or returns of the investment in HC. To this end, for consistency with the definition by Dagum and Slottje (2000), HC must be simultaneously seen as an unknown function of formative indicators **F** and as a 'latent effect' underlying earned income (reflective indicator). Furthermore, another inconsistency linked to the role of wealth plagues the approach. In the first step, **k** (together with other formative variables) contributes to define the scores of HC_i, whereas in equation (1) it appears as a concomitant indicator. These inconsistencies affect the estimates with serious problems of interpretability.
This cited example, however, remarks the importance of correctly specifying in the model the presence of concomitant indicators. In literature, a few attempts have been made to accommodate their presence. In CA framework, concomitant indicators typically constitute a regressor block of full rank, which are used together with formative indicators to obtain latent scores. Typical methodologies are extensions of Redundancy Analysis (RA, van den Wollenberg, 1977) and RRR model to more than two sets of variables. However, they are limited to model particular types of models (Bougeard et al., 2008) and limited to relationship among three sets of variables (Davis and Tso, 1982; Reinsel and Velu, 1998). A new approach to Structural Equation Modelling based on so-called Extended Redundancy Analysis (ERA, Takane and Hwang, 2005), that generalizes RA and RRR for more than two blocks, was recently proposed in literature.

3. The Extended Redundancy Analysis model.
The ERA model can generally be stated as follows: let **Y** denote an *n* by *p* matrix consisting of manifest endogenous variables. Let **X** denote an *n* by *q* matrix consisting of manifest exogenous variables. Assume that the columns of the matrices are mean centered and scaled to unit variance. Then, the ERA model can be expressed as

$$\mathbf{Y} = \mathbf{XWA'} + \mathbf{E} = \mathbf{FA'} + \mathbf{E} \qquad \text{rank}(\mathbf{WA'}) = D \leq \min(q,p) \qquad (2)$$

where **W** denotes a *q* by *D* matrix of component weights, **A'** denotes a D by *p* matrix of component loadings, **E** denotes an *n* by *p* matrix of residuals, **F** $(=[\mathbf{f}_1,\ldots,\mathbf{f}_D] = \mathbf{XW})$ denotes an *n* by *D* matrix of component scores. For identification, **F** is restricted to be $\text{diag}(\mathbf{F'F}) = \mathbf{I}_D$). Model (2) reduces to the RRR or RA model when no variables are shared by both **X** and **Y**, and depending on the constraint imposed on the rank of **WA'**.

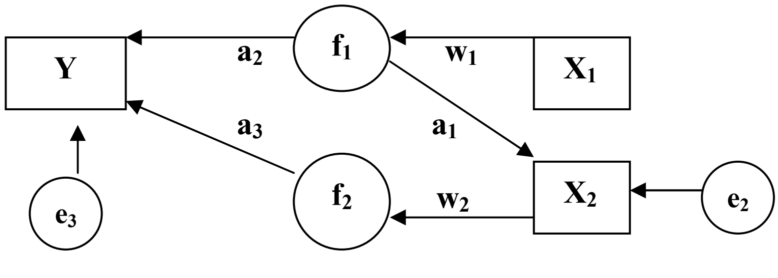

Figure 2 Path diagrams for the ERA model

Another similar approach is the so-called MIMIC (Joreskog and Goldberger 1975) allowing an LV measured by multiple causes (**X**) and multiple indicators (**Y**). However, MIMIC is restrictive because it is based on stringent assumptions about the unitary rank of **A**, the structure of errors covariance matrix (supposed to be diagonal, following the factor model), and the (normal) distribution of errors. Further, it does not admit concomitant indicators. The simple model (2) is useful to specify many different structural models, as that depicted in the path diagram of Figure 2. Including in the ERA specification the rule that when a manifest variable block is exogenous as well as endogenous (e.g. $\mathbf{X_2}$), it is included in both **Y** and **X**, the ERA model associated to Figure 2 becomes

$$(\mathbf{X_2} \mid \mathbf{Y}) = (\mathbf{X_1} \mid \mathbf{X_2}) \begin{bmatrix} w_{11} & 0 \\ \cdots & \\ w_{1q1} & 0 \\ 0 & w_{21} \\ & \cdots \\ 0 & w_{2q2} \end{bmatrix} \begin{bmatrix} a_{11}\cdots & a_{1q1} & a_{21}\cdots & a_{2p} \\ & 0' & a_{31}\cdots & a_{3p} \end{bmatrix} + [\mathbf{e_2} \mid \mathbf{e_3}]$$

$$\tilde{\mathbf{Y}} = (\mathbf{X_1} \mid \mathbf{X_2}) \quad [\mathbf{w_1} \mid \mathbf{w_2}] \quad \begin{bmatrix} \mathbf{a_1}' \mid \mathbf{a_2}' \\ \mathbf{0}' \mid \mathbf{a_3}' \end{bmatrix} + [\mathbf{e_2} \mid \mathbf{e_3}] = \mathbf{FWA}' + \mathbf{E} \qquad (2.1)$$

where $\mathbf{F} = (\mathbf{f_1} \mid \mathbf{f_2})$ denotes an n by 2 matrix of (unidimensional) latent component scores and $\mathbf{w_1}=(w_{11},\ldots, w_{1q1})$, $\mathbf{w_2}=(w_{21},\ldots,w_{2q2})$ are component weights for $\mathbf{X_1}$ and $\mathbf{X_2}$ having q1 and q2 manifest variables, respectively; $\mathbf{a_1}'=(a_{11},\ldots,a_{1q1})$, $\mathbf{a_2}'=(a_{21},\ldots,a_{2p})$, $\mathbf{a_3}'=(a_{31},\ldots,a_{3p})$ are component loadings for $\mathbf{f_1}$ on $\mathbf{X_2}$, $\mathbf{f_1}$ on **Y** and $\mathbf{f_2}$ on **Y**, respectively. The components are specified (in formative schemes) to affect the manifest variables, that is: $\mathbf{Y} = \mathbf{f_1}\,\mathbf{a_2}' + \mathbf{f_2}\,\mathbf{a_3}' + \mathbf{e_3}$ and $\mathbf{X_2} = \mathbf{f_1}\,\mathbf{a_1}' + \mathbf{e_2}$.

3.1 The ERA model with concomitant indicators

In ERA, the concomitant indicators **T** (embedded in the *n* by *K* matrix **T** of full rank) are typically supposed to be a subset of the indicators of matrix **X** ($\mathbf{T} \subseteq \mathbf{X}$). However, in real applications, it might be the case that external covariates do not belong to (strictly) formative indicators of a LC, but may have an external effect on these LC. Hence, in a more general way, by constraining to zero some elements in the matrix **W** (equation 2) that correspond to the indicators in **T**, we suppose that **T** does not belong to the formative indicators **X** and has a causal impact (by means of the K by p matrix of regression coefficients $\mathbf{A_Y}'$) on manifest endogenous variables (Figure 3).

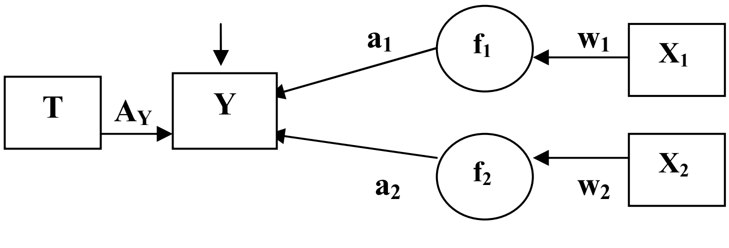

Figure 3 -Structural model with concomitant indicators

As mentioned by authors, concomitant indicators could be accommodated in a new specification of equation (2) that involves three new regressors $\mathbf{X}^{\S}$, weights $\mathbf{W}^{\S}$ and parameters $\mathbf{A}^{\S}$ matrices, with:

$$\mathbf{X}^{\S}_{n \times (K+q)} = [\mathbf{T}_{nxK} \mid \mathbf{X}_{nxq}] \quad \mathbf{W}^{\S}_{(K+q) \times (K+d)} = [\mathbf{G}_{KxK} \mid \mathbf{W}_{qxd}] \quad \mathbf{A}^{\S\prime}_{(K+d) \times p} = \begin{bmatrix} \mathbf{A}'_{\mathbf{Y}\,\mathbf{K}\,\mathbf{x}\,\mathbf{p}} \\ \mathbf{A}'_{\mathbf{d}\,\mathbf{x}\,\mathbf{p}} \end{bmatrix} \qquad (3)$$

where $\mathbf{G}$ is a matrix of fixed elements (zeros or unities), where unity in k-th column and a non zero coefficient in the j-th row (j=1,…,p) of $\mathbf{A_Y}$ selects a causal link between the k-th concomitant indicator and the j-th endogenous indicator. Notice that in the ERA algorithm the structure for $\mathbf{G}$ (the supplementary columns of $\mathbf{W}^{\S}$ consisting of all fixed elements, zeros or unities) has to be considered as fixed elements, whereas the matrix $\mathbf{A_Y}'$ has free parameters to be estimated.
However, the ERA model parameterized as in equation (3) has some limitations in managing direct effects and concomitant indicators. In fact, the LC scores are estimated ignoring the presence of direct effects linking concomitant indicators and the endogenous variables block. In fact, ERA estimates parameters by a sequential algorithm ($\mathbf{W}^{\S}$ for fixed $\mathbf{A}^{\S'}$ and $\mathbf{A}^{\S'}$ for fixed $\mathbf{W}^{\S}$). In each first step, $\mathbf{W}^{\S}$ is estimated by reconstruction of their blocks: firstly, using an estimate of $\mathbf{W}$ (once the effect corresponding to the unit elements $\mathbf{G}$ is subtracted from $\mathbf{W}^{\S}$ and the corresponding columns ($\mathbf{T}$) in $\mathbf{X}^{\S}$ are eliminated) and secondly, replenishing the unit elements in $\mathbf{G}$. Since this is equivalent to fit an ERA model between $\mathbf{Y}$ and $\mathbf{X}$, the estimation of LC scores is obtained ignoring the presence of concomitant indicators (and especially their direct effects on $\mathbf{Y}$).

4. A new estimation method

In this section, we propose a new estimation method that estimates LC as linear combinations of manifest variables, allowing fitting diverse complex relationships among variables, including concomitant indicators that do not "form" latent variables, but affect endogenous variables. In presence of concomitant indicators, model (2) becomes

$$\mathbf{Y} = \mathbf{XWA}' + \mathbf{TA_Y}' + \mathbf{E} = \mathbf{XB} + \mathbf{TA_Y}' + \mathbf{E} \qquad (4)$$

where $\mathbf{B}=\mathbf{WA}'$. We estimate the unknown parameters in model (4), $\mathbf{W}$, $\mathbf{A}$ and $\mathbf{A_Y}'$ in such a way that the sum of squares of the residuals, $\mathbf{E} = \mathbf{Y} - (\mathbf{XB} + \mathbf{TA_Y}')$ is as small as possible. This amounts to minimizing

$$f = SS(\mathbf{Y} - \mathbf{XB} - \mathbf{TA_Y}') = SS(\mathbf{Y} - \mathbf{FA}' - \mathbf{TA_Y}') \qquad (5)$$

with respect to $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{A_Y}$ subject to diag($\mathbf{F}'\mathbf{F}$) $=\mathbf{I_D}$, where SS($\mathbf{X}$) = trace($\mathbf{X}'\mathbf{X}$).
Unfortunately, unlike statistical methods based on Singular Value Decomposition (SVD) or Generalized Singular Value Decomposition (GSVD), minimizing (5) does not provide analytical solutions, due to the particular structure in $\mathbf{W}$ and $\mathbf{A}'$ that contain zeroes, depending on the specified model (see equation 2.1). Hence, we use an iterative method, employing an alternating least squares (ALS) algorithm, developed by Kiers and ten Berge (1989). In the algorithm, parameter matrices $\mathbf{W}$ and $\mathbf{A}'$, $\mathbf{A_Y}'$ are alternately updated until convergence is reached. In the first step, given an initial estimate for $\mathbf{A_Y}'$, $\mathbf{W}$ is updated for fixed $\mathbf{A}$ and $\mathbf{A_Y}'$, whereas in the second step $\mathbf{A}$ is updated for fixed $\mathbf{W}$ and $\mathbf{A_Y}'$.

Estimation of W

Firstly, $\mathbf{A_Y}'$ is initialized with regression parameter referring to a multivariate regression of $\mathbf{Y}$ onto $\mathbf{T}$. Defining $\mathbf{Y}^* = \mathbf{Y} - \mathbf{TA_Y}'$ and let $\mathbf{X} = \mathbf{QR}'$ be portion of the QR decomposition of $\mathbf{X}$, pertaining to the column space of $\mathbf{X}$, where $\mathbf{Q}$ is an n by q orthonormal matrix and $\mathbf{R}'$ is a q by n upper-triangular matrix, equation (5) can be rewritten as follows:

$$SS[\mathbf{Y} - \mathbf{XB} - \mathbf{TA_Y}'] = SS[\mathbf{Y}^* - \mathbf{QR}'\mathbf{B}] = SS[\mathbf{Y}^* - \mathbf{QQ}'\mathbf{Y}^* + \mathbf{QQ}'\mathbf{Y}^* - \mathbf{QR}'\mathbf{B}] \qquad (6.1)$$

since trace[$(\mathbf{Y}^* - \mathbf{QQ}'\mathbf{Y}^*)'\mathbf{Q}(\mathbf{Q}'\mathbf{Y}^* - \mathbf{R}'\mathbf{B})$]= trace[$\mathbf{Y}^{*\prime}\mathbf{QQ}'\mathbf{Y}^* - \mathbf{Y}^{*\prime}\mathbf{QR}'\mathbf{B} - \mathbf{Y}^{*\prime}\mathbf{QQ}'\mathbf{QQ}'\mathbf{Y}^* + \mathbf{Y}^{*\prime}\mathbf{QQ}'\mathbf{Q}\,\mathbf{R}'\mathbf{B}$] is null, equation (6.1) can be rewritten as

$$SS[\mathbf{Y}^* - \mathbf{QQ}'\mathbf{Y}^*] + SS[\mathbf{Q}(\mathbf{Q}'\mathbf{Y}^* - \mathbf{R}'\mathbf{B})] \qquad (6.2)$$

The first term of the right-hand side in (6.2) does not depend on **B,** and minimizing (6.2) reduces to minimizing

$$f = SS[\mathbf{Q}(\mathbf{Q}'\mathbf{Y}^* - \mathbf{R}'\mathbf{B})] = SS[\mathbf{Q}'\mathbf{Y}^* - \mathbf{R}'\mathbf{W}\mathbf{A}'] = SS[vec(\mathbf{Q}'\mathbf{Y}^*) - vec(\mathbf{R}'\,\mathbf{W}\mathbf{A}')]$$

$$= SS\,[vec(\mathbf{Q}'\mathbf{Y}^*) - (\mathbf{A} \otimes \mathbf{R}')\,vec(\mathbf{W})] = SS[vec(\mathbf{Q}'\mathbf{Y}^*) - \boldsymbol{\Omega}\mathbf{w}] \qquad (7)$$

where $\mathbf{w}$=vec($\mathbf{W}$) denotes a supervector consisting of all columns of **W**, one below another, ⊗ denotes a Kronecker product and $\boldsymbol{\Omega} = \mathbf{A} \otimes \mathbf{R}'$.

To compute an initial estimate of **W**, $\mathbf{A}'$ is initialized with arbitrary values (Section 5 discusses possible consistent initial estimates of **W** and **A**). Then, we must estimate $\mathbf{W}'$ without destroying its structure –the zero elements in vec(**W**)–, depending on the existing links between manifest variables and latent composites. Let $\mathbf{w}_*$ denote the vector that selects non-zero elements from **w** and let $\boldsymbol{\Omega}_*$ denote the matrix formed by eliminating the columns of $\boldsymbol{\Omega}$ corresponding to the zero elements in **w**. Then, after having initialized **A** with initial arbitrary values, we obtain the least squares estimate of $\mathbf{w}_*$ by

$$\mathbf{w}^* = (\boldsymbol{\Omega}_*'\,\boldsymbol{\Omega}_*)^{-1}\,\boldsymbol{\Omega}_*'\,vec(\mathbf{Q}'\mathbf{Y}^*) \qquad (8)$$

We can simply reconstruct the updated **w** from $\mathbf{w}_*$ by putting back the zero elements to their original positions, and then the updated **W** from **w**. We then obtain **F**=**XW** and normalize it so that diag($\mathbf{F}'\mathbf{F}$) = **I**.

Estimation of A

As explained, regression parameters **A** have to be estimated, net of influence of **T** onto the matrix of endogenous manifest variables. Thus, in the next step, $\mathbf{A}'$ is updated for fixed $\mathbf{A}_Y'$ and **W**. For given $\mathbf{A}_Y'$ and **W**, the loss function (5) becomes

$$SS[\mathbf{Y} - \mathbf{F}\mathbf{A}' - \mathbf{T}\,\mathbf{A}_Y'] = SS[vec(\mathbf{Y}^*) - vec(\mathbf{F}\mathbf{A}')] = SS[vec(\mathbf{Y}^*) - (\mathbf{I} \otimes \mathbf{F})\,vec(\mathbf{A}')] \qquad (9)$$

Defining $\mathbf{a}$=vec($\mathbf{A}'$), $\boldsymbol{\Gamma}=\mathbf{I}\otimes\mathbf{F}$ and $\mathbf{a}_*$ and $\boldsymbol{\Gamma}_*$ in a way similar to $\mathbf{w}_*$ and $\boldsymbol{\Omega}_*$, the least squares estimate of $\mathbf{a}_*$ is

$$\mathbf{a}_* = (\boldsymbol{\Gamma}_*'\,\boldsymbol{\Gamma}_*)^{-1}\,\boldsymbol{\Gamma}_*'\,vec(\mathbf{Y}^*) \qquad (10)$$

and we easily recover the updated **a** and $\mathbf{A}'$ from $\mathbf{a}_*$. The above two steps are alternated until convergence is reached, that is, until the decrease in the function value falls below a certain threshold value.

Note that each iteration starts by updating the new values of $\mathbf{A}_Y'$. For example, in the second iteration, $\mathbf{A}_Y'$ is updated by equation (5), by regressing $\mathbf{Y} - \mathbf{X}\mathbf{B}^\#$ onto **T**, where $\mathbf{B}^\# = \mathbf{W}\mathbf{A}'$ are estimates of **W** and $\mathbf{A}'$, obtained in the first iteration.

Finally, since the algorithm solves a global optimization problem by consistently minimizing (5), the total fit (Ψ) of a hypothesized model is measured, as in ERA, by the total variance of the manifest endogenous variables explained by the exogenous (formative and concomitant) variables.

5. Discussion

Observing the estimated weights in equation (8), the latent composites have a meaningful interpretation: weights defining **F** scores are obtained as projection of a particular endogenous matrix onto the space spanned by vec($\boldsymbol{\Omega}_*$), the principal directions of an orthonormal basis for the space spanned by column of **X** (QR decomposition). Specifically, we project **Q'Y***, the projection of endogenous manifest variables whose scores have been corrected by the effect of **T** on **Y**. Hence, contrary to ERA approach, the weights to obtain latent composites scores **F** are obtained projecting the endogenous matrix of indicators that marginalizes the direct effect of **T** onto **Y**.

As noticeably known, ALS algorithms do not guarantee that the obtained minimum is a global minimum (ten Berge, 1993). However, this problem may be avoided in different ways. Among others, when we choose good initial values, the function value is likely to start near to the global minimum, and it is more likely to obtain the global minimum. To this end, Multiblock Redundancy Analysis (MRA, Bougeard et al., 2008) appears as a rational choice. For MRA, explanatory formative blocks $\mathbf{X_k}$ (k=1,...,K) are merged in the matrix **X**, defined as $\mathbf{X}=[\mathbf{X_1}|\ldots|\mathbf{X_K}]$, that is summarized with global components **t**, a linear combination of all the variables derived from the **X** matrix. The first component summarizes K partial components $\mathbf{t_k}^1=\mathbf{X_k g_k}^1$ (with $\|\mathbf{g_k}^1\|=1$), respectively associated with the K blocks $\mathbf{X_k}$. The first global component $\mathbf{t}^1 = \mathbf{Xg}^1 = \Sigma_k \mathbf{g_k}^1\, \mathbf{t_k}^1$ is closely related to the first component $\mathbf{u}^1=\mathbf{Yv}^1$ (with $\|\mathbf{v}^1\|=1$) which summarizes the endogenous manifest variables and is a linear combination of the **Y** variables. MRA components are computed while maximizing a criterion based on the squared covariance between the global component $\mathbf{t}^1$ and the component $\mathbf{u}^1$. The unidimensional solution $\mathbf{v}^1$ is given by the eigenvector associated with the largest eigenvalue of the matrix $\Sigma_k\, \mathbf{Y'X_k(X_k'\,X_k)^{-1}X_k'Y}$.

We suggest to compute a MRA and use the principal coordinates as the rational start for **W** (whereas initial values of **A'** are simply obtained by the least squares estimate, given **W**). Hence, starting values for **W** are obtained by horizontal concatenation of the $\mathbf{g_k}^1$ vectors in the matrix $\mathbf{G}^1=[\mathbf{g_1}^1,\ldots,\mathbf{g_k}^1,\ldots,\mathbf{g_K}^1]$ via a MRA model in which $\mathbf{X}=[\mathbf{X_1}|\ldots|\mathbf{X_K}]$ are the regressor blocks and the dependent block is $\mathbf{Y^*}= \mathbf{Y} - \mathbf{T\hat{A}_Y}'$ (where $\mathbf{\hat{A}_Y}'$ is an estimation of $\mathbf{A_Y}'$).

6. An application on Human capital University Graduates

The aim of the present application is to measure the causal impact of formal Human capital (HC), accumulated during Higher education, on the initial earnings for University of Milan (Italy) graduates. To illustrate the proposed approach, we have to specify a structural model consistent with realistic economic process of HC accumulation.

The structural schemes hypothesized in recent researches (Dagum *et al.*, 2007; Lovaglio, 2008) and proposed in coherence with well-established economic theory, suggest that HC can be composed by two latent dimensions: the Education HC ($\mathbf{f_1}$= EduHC), whose unobservable scores depend on formative indicators $\mathbf{X_1}$ measuring the amount of investment in education and attributes embodied in individuals and the HC that refers to the performance in the labour market during the working career ($\mathbf{f_2}$=JobHC, generated by its formative indicators $\mathbf{X_2}$). As far as causal relations are concerned, EduHC may have a causal impact on formative indicators of JobHC and both HC latent composites are supposed to have a causal effect on a set of endogenous manifest indicators (**Y**), dealing with the economic performance of graduates in the labour market. Moreover, a realistic model aimed to measure HC dimensions would take into account exogenous socio-demographic factors that do not belong to the HC dimensions, but may have a causal impact on endogenous indicators. The specified concomitant indicators (**T**) typically reflect opportunity factors of formal HC, cultural elements and environmental factors of the parents.

The proposed structural model is depicted in the path diagram of Figure 2, imaging an additional block of concomitant indicators **T** affecting **Y**. In this specification, the associated model involves $\mathbf{\tilde{Y}} = (\mathbf{X_2}|\mathbf{Y})$ as dependent block, $\mathbf{X} = (\mathbf{X_1}|\mathbf{X_2})$ as exogenous block, **T** as a concomitant indicator block and $\mathbf{f_1}$ = EduHC, $\mathbf{f_2}$=JobHC, as latent composites.

Data are drawn by the combination of three institutional administrative archives.

The *Administrative Archive of Milan University* is our main source, which includes socio-demographic characteristics and academic performance for all students gaining a University degree between 2001 and 2007 (47,693 records). The archive of *Employment Centres of the Province of Milan* is the second data source we employ, which offers information on the workers who are active

in the labour market in the area surrounding Milan. It provides fine-grained information on many contextual and descriptive considerations for workers (including contract type, duration of vocational experiences, unemployment periods, as well as demographic characteristics) from 2000 to 2007. Our third and final source is the *Italian Internal Revenue Service Archive,* which adds data on annual gross earned incomes (drawn from annual declarations from 2000 to 2006) for workers residing in the Province of Milan.
These three archives cover differing subpopulations and temporal periods. In this application, we selected only post-graduation earnings (starting one year later than matriculation). Since the last reported data on income refer to occupational experiences during 2006 (declared income in 2007) and to allow at least two years of possible post graduate work, for the present application we selected the graduates from the University of Milan (awarding a degree during the period starting in 2001 and ending in 2004) residing in this specified area who presented furthermore at least two income declarations from 2002 to 2006 (14,695 records).
Here, the *observed career path* for each graduate starts after graduation and ends at the end of 2006. Depending on the graduation date, the selected graduates may present a different number of (possible) income declarations, ranging from two (incomes in years 2005-2006 for graduation in 2004) to five (incomes in years 2002-2006 for graduation in 2001).

X_1: formative for EduHC	**X_2: formative for JobHC**
e_1: Years of schooling	J_1: Evolution of the career path (observed period)
m_1: Legal duration Time / Time to obtain degree	J_2: Saturation of the career path (observed period)
a_1: Expected Graduation Age /Graduation age	J_3: Time (from degree) to obtain the I^{st} permanent contract
a_2: Graduation score/Mean graduation score	
Y: endogenous indicators	**T: concomitant indicators**
y_1: Income level (last available year)	t_1: Origin's household economic status
y_2: Annual income growth rate (2003-06)	t_2: Size of the origin's household

Table 1 – Indicators and blocks used for the estimation of Human Capital.

Table 1 exhibits the proposed indicators for each block. Specifically, $\mathbf{X_1}$ describes EduHC ($\mathbf{f_1}$) formative indicators involving the stock of formal education (e_1), a motivation indicator (m_1 is the ratio between the months of legal duration and the months to obtain degree) and two ability indicators (a_1 is the ratio of expected age of graduation for a regular University student to the actual age of graduation, whereas a_2 is the ratio of the score at graduation to the mean score, obtained by averaging the scores of graduates that earn the degree in the same academic year from the same Faculty).
As JobHC formative indicators we have considered an indicator called 'evolution', obtained by a longitudinal clustering algorithm (Lovaglio and Mezzanzanica, 2008), that directly elicits the longitudinal evolution of workers' careers in terms of transitions between contractual typologies (Permanent contract, Fixed-term contract, temporary agency work) in the work career during the observed period. Increasing values indicate stability in permanent contract or evolutions towards this category, whereas decreasing values describe persistence/evolutions towards temporary agency work; 2) the saturation of the career, calculated as the ratio of days of full-time work on the total amount of potential days in the observed period and 3) the length of time (in days) to obtain the first permanent contract, after graduation. Moreover, we use two concomitant indicators to describe an individual's personal background: family size and origin's household economic status, represented with an ordinal indicator comprised of ten levels corresponding to increasing wealth (earned and property incomes, real and financial wealth).
As endogenous variables, we use two indicators describing the economic performance of graduates from two simultaneous sides: a cross-sectional (earning levels at a prefixed time) and a longitudinal

(evolution of earning trajectories over time) perspective. Specifically, for the level, we consider for each graduate the most recent declared annual income during the observed career path, whereas for the dynamic, the linear income growth rate (annual raises) during the most recent available triennium, or the most recent biennium for workers with only two income declarations.

Descriptive statistics show that in the selected sample, 90% of graduates present at least three observed income declarations (78% in the last triennium 2004-06) and a large quota of graduates (88%) has observed income in 2006. Thus, available data cover initial labour market vocational experiences reflecting post-graduation performances for recent graduates in the very beginning of their careers.

Using the proposed approach, the model was fitted to the data. The weights, the parameter estimates, their bootstrapped standard errors (obtained with 200 bootstrap samples) and critical ratios of t-statistics (the ratio of parameter estimates on their bootstrapped standard errors) are given in Table 2. The goodness of fit of the model (Ψ=0.312) indicates that about one third of the total variance of the endogenous variables (**Y** and $\mathbf{X_2}$) is accounted for by the two-component model and concomitant indicators. The squared multiple correlations of y_1 (0.56) and y_2 (0.19) indicates that about 50% of the variance of income levels and 20% of the variance of income slopes were explained by the two latent variables and concomitant indicators. Instead, squared multiple correlations of J_i (0.12; 0.07, 0.04 for J_1, J_2 and J_3 respectively) indicate that a small quota of the variance of the JobHC indicators were explained by the education HC.

Measurement model: weights (standard errors)	**Structural model: parameters (t-statistic)**	**Concomitant Indicators parameters (t-statistic)**
Formative for $\mathbf{f_1}$ (EduHC)	$\mathbf{f_1 \rightarrow X_2}$	$\mathbf{t_1 \rightarrow Y}$
$e_1 \rightarrow \mathbf{f_1}$ 0.791 (0.06) $m_1 \rightarrow \mathbf{f_1}$ 0.090 (0.14)	$\mathbf{f_1} \rightarrow J_1$ 0.093 (5.74) $\mathbf{f_1} \rightarrow J_2$ 0.089 (4.69)	$\mathbf{t_1} \rightarrow y_1$ 0.016 (0.46)
$a_1 \rightarrow \mathbf{f_1}$ 0.555 (0.07) $a_2 \rightarrow \mathbf{f_1}$ 0.239 (0.09)	$\mathbf{f_1} \rightarrow J_3$ -0.021 (-0.56)	$\mathbf{t_1} \rightarrow y_2$ 0.035 (1.36)
Formative for $\mathbf{f_2}$ (JobHC)	$\mathbf{f_1\, f_2 \rightarrow Y}$	$\mathbf{t_2 \rightarrow Y}$
$J_1 \rightarrow \mathbf{f_2}$ 0.696 (0.11) $J_2 \rightarrow \mathbf{f_2}$ 0.646 (0.12)	$\mathbf{f_1} \rightarrow y_1$ 0.137 (5.39) $\mathbf{f_1} \rightarrow y_2$ 0.039 (1.48)	$\mathbf{t_2} \rightarrow y_1$ -0.008 (-0.33)
$J_3 \rightarrow \mathbf{f_2}$ -0.613 (0.13)	$\mathbf{f_2} \rightarrow y_1$ 0.220 (8.81) $\mathbf{f_2} \rightarrow y_2$ 0.122 (4.77)	$\mathbf{t_2} \rightarrow y_2$ -0.022 (-0.87)

Table 2 –Weights and parameter estimates, obtained from the GERA model.

About the measurement models, excluding m_1 (critical t-ratio=0.64), the component weights associated with EduHC were all significant and positive: EduHC is largely attributed to years of formal schooling (e_1) and to age of graduation (a_1), showing the largest weights and in a lesser extent to graduation score (a_2). For the other dimension, all selected (standardized) indicators are largely significant and equally contribute to the formation of JobHC scores.

As far as causal relations between LC and endogenous variables are concerned, except the slight effect of EduHC on the income slope, both HC dimensions have significantly effects on income levels variability (y_1) and on earnings longitudinal dynamics (y_2). Concerning the structural parameters among HC dimensions, EduHC has a significant impact on both contractual stability (J_1) and saturation (J_2) of the career path, but has no effect on the speed to find a permanent job (J_3).

Further, concomitant indicators have not significant impact on **Y**. Surprisingly, the parents' economic status (t_1) does not have a direct impact on the workers' economic performance. This may be interpreted by the consideration that, typically, parents' economic status largely contribute to the formation of Education HC, supporting the notion that much of the influence of family background is indirect and nested in educational choices. This expected relationship could be specified by adding a causal path between **T** and formative indicators of EduHC. Unfortunately, the proposed model does not allow specifying this kind of relations, meaning that concomitant indicators affect both endogenous variables and formative indicators of some LCs. An extension of this framework is under investigation.

7. Conclusion

The evidence on the shortcomings of classical methods to assess the value of HC stock grounded on national accounting schemes (both the retrospective approach based on the production cost of HC as a good measure of its value, and the prospective one based on the present value of labour income streams) has opened up new areas of research. HC is increasingly recognized as having several sources that are linked not only to formal education and training but also to culture, family background, social context and – to a significant extent – innate and non-cognitive abilities and skills (Heckman *et al.*, 2005).

The definition of HC as a 'latent effect' of formative variables regarding its investment and, at the same time, as a 'latent cause' of manifest indicators, measured on individuals or families, obtained by means of adequate structural equations models (when reliable information is available) may address all these problems. To this end, we have to specify a measurement model consistent with realistic process and to utilize a proper LV technique for the estimation of its scores. Although Covariance-based SEM, PLSPM and CA models overlap in analysis objectives, there are distinct differences among these approaches that make each more or less appropriate for certain types of analysis.

The proposed estimation method is simple yet versatile enough to fit various complex relationships among variables, including direct effects of manifest variables (concomitant variables). It estimates model parameters by minimizing an overall model fit, such as the sum of squares of discrepancies between the manifest endogenous variables and their predicted counterparts from the exogenous and concomitant variables without any explicit distributional assumptions.

Concerning the empirical application focused on Graduates HC, a consistent model is proposed in coherence with well-established economic theory that informs and supports the specification of relationships in the structural model. Specifically, the estimation procedure really estimates both latent variables respecting their economic definitions (and the causal role in the structural model), as that stock of investment in education and personal characteristics (EduHC) and performance in the labour market (JobHC) that are relevant to economic activity, removing the effect of concomitant indicators.

However, the presented methodology has two noticeable limitations. Firstly, although endogenous LCs are allowed in the model, it cannot assume any latent variables for the manifest endogenous variables. Specifically, since the method accommodate only formative variables for estimating latent composites, a direct structural link between an exogenous and an endogenous LC is impossible to assume, but it has to be mediated by the formative indicators of the endogenous LC. Secondly, as emerged in the empirical application, one might also be interested in measuring the effect of concomitant indicators on the latent constructs identified from the model. Specifically, concomitant indicators may have a causal impact on manifest endogenous variables and onto latent composites, too. Future studies are needed to investigate the feasibility of these extensions.

To conclude, for the measurement of HC many problems remain: reliable estimation of the total costs of investing in HC, quality of education and several dimensions (such as intelligence, ability, hard work and on-the-job learning) are difficult to obtain (both at individual or aggregate level). Moreover, returns (reflective indicators) may be 'hidden' in a firm's results and consequently not measured by earnings (Oxley *et al.* 2008). Finally, other classical topics deal with the problems of endogeneity and causality. To this end, the integration of official sample surveys (such as AlmaLaurea), capturing typically unmeasured dimensions such as the degree's efficacy and consistence with job, work's satisfaction, and amount of training on the job periods, etc.) with administrative archives, collecting official information on earnings, appear as a promising strategy for HC estimation.

References

Abraham, B. and Merola, G. (2001): Dimensionality reduction approach to multivariate prediction. In Esposito Vinzi V. et al. (Eds.): *PLS and related methods*, CISIA, 3-17.
Becker, G. S. (1975): *Human Capital*, 2nd ed. Columbia University Press, New York.
Bentler, P.M. (1992): On the fit of models to covariances and methodology to the Bulletin. In: *Psychological Bulletin,* 112(3): 400–404.
Bougeard S., Hanafi M., Lupo C. and Qannari E.M. (2008): From multiblock Partial Least Squares to multiblock redundancy analysis. A continuum approach. International Conference on Computational Statistics, Porto (Portugal).
Croon, M. and Bolck, A. (1997): On the use of factor scores in structural equations models (Tech. Rep. No. 97.10.102/7). Work and Organization Research Centre, Tilburg University. (WORC Paper). The Netherlands.
Dagum C., Vittadini G., Lovaglio P.G. (2007) Formative Indicators and Effects of a Causal Model for Household Human Capital with Application. In: Econometric Reviews, 26(5), pp. 579–596 .
Dagum, C. (1994): Human capital, income and wealth distribution models and their applications to the USA. *Proceedings of the 154th Meeting of the American Statistical Association* (pp. 253–258).
Dagum, C. and Slottje, D.J. (2000): A new method to estimate the level and distribution of the household human capital with application. In: *Structural Change and Economic Dynamics* 11: 67–94.
Dagum, C., Vittadini, G. and Lovaglio, P.G. (2007): Formative indicators and effects of a causal model for household human capital with applications. In: *Econometric Review* 26(5): 579–596.
Heckman, J.J., Stixrud, J. and Urzua, S. (2005): The effects of cognitive and non cognitive abilities on labor market outcomes and social behaviour. Paper presented at the Second IZA Prize Conference, Berlin.
Izenman, A.J., (1975): Reduced-rank regression for the multivariate linear model. In: Journal of Multivariate Analysis 5, 248–264.
Jôreskog, K. G. (1970): A general method for analysis of covariance structures. In: *Biometrika,* 57, 409-426.
Jôreskog, K.G., and Goldberger, A.S. (1975): Estimation of a model with multiple indicators and multiple causes of a single latent variable. In: *Journal of the American Statistical Association,* 10, 63 1-639.
Kiers, H. A. L., Takane, Y., and ten Berge, J.M.F (1996): The analysis of multitrait-multimethod matrices via constrained components analysis. In: *Psychometrika,* 61, 601-628.
Lovaglio P.G. (2008) Process of Accumulation of Italian Human Capital. In: Structural Change and Economic Dynamics, 19, 342-356
Lovaglio P.G., Mezzanzanica M. (2008): Modelling longitudinal sequences by optimal scaling techniques. In: Italian journal of applied statistics, 20(3), pp. 251-272.
McDonald, R. P. (1996): Path analysis with composite variables. In: *Multivariate Behavioral Research,* 31, 239-270.
Millsap, R. E., and Meredith, W. (1988): Component analysis in cross-sectional and longitudinal data. In: *Psychometrika,* 53, 123-134
Mincer, J. (1970): The distribution of labor incomes: a survey. In: *Journal of Economic Literature*, 8(1), 1–26.
OECD (1998): *Human capital investment: an international comparison,* Centre for International Research and Innovation, Paris.
Oxley, L., Le, T. and Gibson, J. (2008): Measuring human capital: alternative methods and international evidence. In: *Korean Economic Review* 24(2): 283–344.
Petty, W. (1690): *Political Arithmetik*, *or a Discourse Conceirning the Extent and Value of Lands, People, Buildings*. Reprinted in C.H. Hull (1899): *The Economic Writings of Sir William Petty*. Cambridge: Cambridge University Press.
Rao. C. R. (1980): Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 3-22). Amsterdam: North Holland.
Reinsel., G. C.., and Velu, R. P. (1998): *Multivariate Reduced-Rank Regression.* New York: Springer-Verlag.
Schônemann, P. H., and Steiger, J. H. (1976): Regression component analysis. In: *British Journal of Mathematical and Statistical Psychology,* 29, 175-189.
Schultz, T. W. (1961): Investment in human capital. *American Economic Review*, 51(1), 1–17.
Takane Y. and Hwang, H. (2005): An extended redundancy analysis and its applications to two practical examples. In: Computational Statistics & Data Analysis, 49(3), 785-808
ten Berge, J. M. F. (I993). Least squares optimization in multivariate analysis. Leiden: DSWO Press.
Tenenhaus, M., Vinzi, V., Esposito, V., Chatelin, Y.M. and Lauro, C. (2005): PLS path modelling. In: *Computational Statistics & Data Analysis* 48(1): 159–205.
Tso, M. K.-S. (1981): Reduced-rank regression and canonical analysis. In: *Journal of the Royal Statistical Society B,* 43, 183-189.
United Nations (2002): *Human Development Report 2000: Human Rights and Human Development.* New York: Oxford University Press.
van den Wollenberg, A.L. (1977): Redundancy analysis: an alternative for canonical analysis. In: Psychometrika 42, 207–219.
Velicer, W. F., and Jackson, D.N. (1990): Component analysis vs. common factor analysis: Some further observations. In: *Multivariate Behavioral Research, 25*, 95-112
Vittadini G. (1989): Indeterminacy problems in the Lisrel model. In: *Multivariate Behavioral Research*, 24(4), pp.397-414.

Wold, H. (1982): Soft modeling: The basic design and some extensions. In K. G Jôreskog and H. Wold (Eds.), *Systems Under Indirect Observations: Causality, Structure, Prediction, Part* 2 (pp. 1-54). Amsterdam: North-Rolland.
World Bank (1995): Workers in an integrating world. *World Development Report*. Washington, DC.
Wossmann, L. (2003): Specifying human capital. In: *Journal of Economic Surveys* 17(3):239–270.