JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **31**: 734–755 (2016)
Published online 10 April 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2456

MAINTAINING (LOCUS OF) CONTROL? DATA COMBINATION FOR THE IDENTIFICATION AND INFERENCE OF FACTOR STRUCTURE MODELS

RÉMI PIATEK[a*] AND PIA PINGER[b]
[a] *Department of Economics, University of Copenhagen, Denmark*
[b] *Department of Economics, University of Bonn, Germany*

SUMMARY

Factor structure models are widely used in economics to extract latent variables, such as personality traits, and to measure their impact on outcomes of interest. The identification and inference of these models, however, highly depend on the availability of rich longitudinal data. To overcome the common problem of data scarcity, this paper proposes to combine datasets that each identify some part of the likelihood, thereby recovering the identification of the complete model. The performance of the approach is demonstrated by a Monte Carlo experiment. We apply this technique empirically to study the impact of locus of control on education and wages. Our strategy allows us to elicit the distribution of pre-market locus of control from a sample of young individuals, and to measure its impact on education and wages in a sample of adults. Our findings indicate that the effect of locus of control on wages mainly operates through education. Copyright © 2015 John Wiley & Sons, Ltd.

Received 14 December 2010; Revised 27 January 2015

Supporting information may be found in the online version of this article.

1. INTRODUCTION

Individual abilities and traits are increasingly incorporated into economic models to better explain human behaviors and outcomes (Almlund *et al.*, 2011). To this end, factor models have been successfully applied in the empirical literature to extract the distribution of unobserved traits of interest (such as cognitive skills and personality), and to measure the impact of these traits on schooling decisions, labor market outcomes, and health-related behaviors (Hansen *et al.*, 2004; Aakvik *et al.*, 2005; Cunha *et al.*, 2005; Cunha and Heckman, 2007; Gensowski, 2014; Heckman *et al.*, 2014a; Conti *et al.*, 2014). A particular example of a latent personality trait, which will be the focus of our empirical study, is the psychological concept of *locus of control* (see Lefcourt, 1991, for an introduction). This dimension of personality measures the extent to which individuals believe that what happens to them in life is related to their own actions and decisions or, on the contrary, to fate and luck. Locus of control has received much attention in the literature, and has been shown to explain a wide range of economic outcomes (Coleman and DeLeire, 2003; Cebi, 2007; Caliendo *et al.*, 2015).

However, a recurrent problem inherent to these studies concerns the availability of relevant variables, both to recover the distribution of the latent traits of interest and to measure how these traits influence outcomes. In this respect, the timing of the data collection plays a crucial role. If locus of control is measured at the same time as the outcomes in our example, a reverse causality problem may affect the estimation and lead to an overestimation of its impact on outcomes. To address this issue, many authors use personality measures that were obtained several years prior to labor market outcomes (see, for example, Heckman *et al.*, 2006). In our paper, we refer to the early measured locus

* Correspondence to: Rémi Piatek, Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK-1353, Copenhagen K, Denmark. E-mail: Remi.Piatek@econ.ku.dk

of control as 'pre-market locus of control'. This type of study, however, highly depends on the availability of rich life cycle data that contain both early measurements of unobserved traits and lifetime outcomes. Unfortunately, longitudinal datasets are often truncated in practice, making it impossible to work with complete samples where observations are available for all individuals on all variables.

As an alternative, this paper proposes to use a data combination strategy to remedy a shortage of rich longitudinal data that contain measurements and outcomes at different points in life. The idea, proposed by Cunha *et al.* (2005), represents a promising new avenue to solve this problem. Yet it has not been implemented in the empirical literature so far, nor has its performance been thoroughly tested and investigated.[1] The present paper intends to fill this gap. The approach consists of combining different subsamples, where only a subset of variables is observed in each of them, to fully identify the overall factor model. Identification of the complete model is achieved even if each subsample, taken separately, can only partially identify it. In our example, we extract pre-market locus of control from a subsample of youth individuals, and measure its impact on outcomes observed in a subsample of adults.

Our contribution is twofold. First, we present the technical aspects of the data combination strategy, and thoroughly discuss the identification problems at stake. We conduct a Monte Carlo experiment that investigates how well the procedure works in practice, and that sheds new light on the benefits and on the limitations of the approach. We show that the data combination performs well in most cases, even in relatively small samples, but that it depends on the amount of information available. Importantly, it turns out to be necessary to have some overlap between the subsamples—in terms of observed variables (at least one variable needs to be common to the subsamples, to serve as a bridge) and in terms of numbers of individuals present in the different samples. In addition, our simulations provide evidence of an attenuation bias affecting the coefficients on the factor (*factor loadings*) if the data are not rich enough to identify the overall model empirically, even if identification is achieved theoretically. This result has not been reported in the literature previously, and has important implications for the application of the method.

Second, our empirical study revisits analyses of the impact of locus of control on schooling decisions and labor market outcomes. We use the German Socio-Economic Panel (SOEP), where measures of locus of control are provided for two subsamples of individuals: a sample of youth individuals with pre-market measures of locus of control, and a subsample of adults with contemporaneous measures of locus of control and outcomes. Moreover, a small fraction of the youth individuals enter the job market in the subsequent waves of the panel, so that both early measures of locus of control and labor market outcomes are available for them. Therefore, the structure of the SOEP represents a main advantage of this dataset over other survey data, where measures of personality are usually provided for one specific age only. This particular feature of the data gives us the opportunity not only to apply our data combination strategy, but also to compare it to the more common approach relying on contemporaneous measures. Our results indicate that the use of contemporaneous measures can result in an overestimation of the impact of locus of control. Regarding the effect of pre-market locus of control, we find that an increase of locus of control by one standard deviation results in a 5 percentage point difference in the probability of obtaining an upper secondary school certificate. Increasing locus of control by one standard deviation results in an overall hourly wage increase of about 2% for males and 4% for females. Yet the effect of pre-market locus of control on wages merely operates through its effect on education decisions.

Our methodology is not limited to the study of personality traits and cognitive abilities. More generally, it can be applied to any problem where the data are characterized by imperfect measurements and truncated life cycle information. In this sense, our paper relates to the econometric literature on

[1] Frölich (2008) develops statistical methods for the combination of datasets in the framework of a potential outcome model, i.e. in a different context that does not rely on latent factors.

DOI: 10.1002/jae

measurement error and data combination (see Ridder and Moffitt, 2007, for a survey). Examples of other applications are studies of mental health, life satisfaction and preferences, which also tend to be measured at a single point in time but are affected by past experiences of individuals (Theodossiou, 1998; Kassenboehmer and Haisken-DeNew, 2009; Bowles, 1998). The approach is also particularly relevant to researchers who want to exploit rich register data that contain detailed economic outcomes and combine them with psychological survey on latent (personality) traits.

The paper proceeds as follows. Section 2 discusses conceptual issues in the study of abilities and personality traits, which is the type of application we address with our data combination approach. Section 3 presents our model and explains how factor structure models can be used to solve problems of measurement error and truncated life cycle information. It also explains the identification strategy in the presence of missing information. Section 4 then introduces our Bayesian inferential procedure. Section 5 carries out a Monte Carlo experiment to test the performance of the approach, and Section 6 presents our empirical application. Section 7 concludes.

2. CONCEPTUAL ISSUES IN THE STUDY OF ABILITY AND PERSONALITY TRAITS

From a methodological point of view, there are two major econometric problems at stake in the economic literature on individual abilities and traits: measurement error and endogeneity (Bowles and Gintis, 2002; Borghans *et al.*, 2008). First, measurement error arises because certain traits or characteristics are measured by questions or tests that are imperfect proxies of the true latent construct. Yet, in general, most psychological measures are designed to capture a particular latent trait or skill, such that factor-analytic approaches can be used to distinguish true latent abilities from measurement error (Borghans *et al.*, 2008; Heckman *et al.*, 2006; Hansen *et al.*, 2004). Second, endogeneity bias arises when past or anticipated labor market outcomes affect individual traits. In this case, measures may reflect, rather than cause, the outcomes they are supposed to predict (Borghans *et al.*, 2008).[2]

In the literature, four main strategies have been adopted to address the endogeneity of psychological measures. First, Duncan and Morgan (1981) and Duncan and Dunifon (2012), using the Panel Study of Income Dynamics (PSID), extract measures of personality traits measured 15–25 years prior to earnings. A similar strategy has been adopted by Heckman *et al.* (2006), who use locus of control measurements in the National Longitudinal Survey of Youth (NLSY) taken at age 14–22 to explain later outcomes. Second, Bowles *et al.* (2001), using the National Longitudinal Survey of Young Women (NLSYW), employ contemporary measurements of locus of control, which they purge of past wage influences. Third, Osborne (2000) uses past skills to instrument for contemporaneous skill measures. Last, Cunha and Heckman (2008) explicitly model development and accumulation of skills as a technology of skill formation, in which investments in one period affect the productivity of investments in subsequent periods. Our approach is a fifth strategy that, in its principle, is similar to the first one: it builds on combining measurement information from a sample of youths with outcomes from a sample of adult individuals (Cunha *et al.*, 2005).

We focus on locus of control as the most widely used personality concepts in economics (see, for example, Heckman *et al.*, 2006; Judge and Bono, 2001; Andrisani, 1977, 1981; Osborne, 2000).[3] Control tendencies have been shown to predict many economic outcomes and measures of this trait are available in well-known datasets such as the NLSY, the PSID or the German Socioeconomic Panel (SOEP).[4] Besides, it may be fair to say that locus of control is 'among the most popular research tools

[2] Arguably, endogeneity bias only arises if an individual's personality is not fixed at birth, but responsive to positive or negative life experiences and individual (labor market) shocks. For a discussion on the stability of locus of control see Appendix A.2 (supporting information).

[3] For additional information and a theoretical framework for how *pre-market* external locus of control may affect labor market returns, see Appendix A.3 (supporting information).

[4] See Cobb-Clark (2015) for an overview of the economics literature on locus of control.

DOI: 10.1002/jae

in psychology' (Furnham and Steele, 1993) and one of the few psychological concepts that are recognized in many subfields of psychology including clinical, developmental, occupational, personality and social psychology (Furnham and Steele, 1993; Lefcourt, 1966; Rotter, 1990; Wang *et al.*, 2010; Ng *et al.*, 2006).[5] Since the early 1980s, the Big Five personality inventory has rivaled locus of control as the most widely accepted personality concept among personality psychologists (Goldberg, 1993).[6] In that taxonomy, locus of control is related to emotional stability (neuroticism) and recent research suggests that locus of control, self-esteem, and emotional stability are part of a common construct termed core self-evaluation (Judge *et al.*, 2002).[7]

Mostly on empirical grounds, many studies agree that locus of control affects a variety of economic choices.[8] This is particularly true for education decisions, which most researchers find to be highly influenced by locus of control. For instance, Coleman and DeLeire (2003) present a model of locus of control and education decisions, where locus of control is viewed as a behavioral trait that affects education decisions, because it has an impact on personal beliefs about the effect of education on expected earnings. Using the National Education Longitudinal Study (NELS), the authors find locus of control to have a high and significant impact on schooling decisions, as well as on *ex ante* expected earnings conditional on schooling. Contrary to this, using the NLSY, Cebi (2007) concludes that locus of control has a productive impact on labor market outcomes and no effect on education choices. Research on the effect of locus of control on labor market returns finds mostly positive effects. For example, Andrisani (1977), using the National Longitudinal Study (NLS), finds that locus of control affects several measures of earnings and occupational attainment of young and middle-aged men. Yet Duncan and Morgan (1981) find mostly non-significant effects of locus of control on the change in hourly earnings of individuals in the PSID. An analysis of the impact of locus of control on wages using German data (and contemporaneous measurements) has been conducted by Heineck and Anger (2010), as well as by Flossmann *et al.* (2007), with both studies finding positive effects.[9]

3. MODEL SPECIFICATION AND IDENTIFICATION THROUGH DATA COMBINATION

This section lays out the theoretical framework we use to measure the impact of locus of control on education and labor market outcomes. We postulate a latent structure with an unobserved factor to explain the correlation between the unobservables of the model. A measurement system is added to the model to recover the distribution of the latent factor (Section 3.1). This type of factor structure model has been extensively used in the empirical literature (e.g. Carneiro *et al.*, 2003; Heckman *et al.*, 2006). Our paper focuses on the special case where some variables of interest—measurements or outcomes—may not be available for all individuals of the sample. This missing data problem results in identification problems that can be tackled through data combination (Section 3.2).

3.1. A Potential Outcome Model with a Measurement System

Individuals with different levels of schooling become active on different segments of the labor market, where their personal characteristics, as well as their level of locus of control, may be valued differently. Depending on education, we model labor market outcomes as a two-stage process: people first select

[5] PsycInfo counts around 1,000 publications on locus of control for the years 2004–2014.

[6] See Section 2 of Almlund *et al.* (2011) for a brief exposition of the history of personality psychology.

[7] Apart from personality psychologists, motivational psychologists also use the concept of locus of control to measure attribution styles that are part of the psychological human capital of an individual that affect motivation and individual behavior (Weiner, 2012).

[8] More recent evidence shows that locus of control is also predictive of health behaviors (Schurer, 2014; Cobb-Clark *et al.*, 2014).

[9] Furthermore, Gallo *et al.* (2003) and Caliendo *et al.* (2015) use German data to investigate the impact of locus of control on transitions from unemployment to employment.

DOI: 10.1002/jae

into the labor market, and a wage equation is then estimated for those actually working. Observed characteristics and unobserved locus of control are allowed to play a role in both stages. Estimating the two equations simultaneously makes it possible to correct for potential sample selection bias that might affect the parameters if only the wage equation for working people were estimated.

3.1.1. Schooling Decision

Given the two possible levels of education considered in our empirical application, the schooling decision is specified as a binary choice model.[10] Each individual achieves education if the utility associated with this decision crosses a given threshold. The utility S^* derived from education is supposed to depend linearly on a vector of personal characteristics X_S and on a latent factor θ capturing locus of control:

$$S = \mathbf{1}\left(S^* > 0\right), \quad S^* = X_S\beta_S + \theta\alpha_S + \varepsilon_S \tag{1}$$

where the indicator function $\mathbf{1}(\cdot)$ is equal to one if the corresponding condition is verified, and to zero otherwise. The vector β_S denotes the vector of parameters related to personal characteristics, α_S is the factor loading associated with θ, and ε_S is an idiosyncratic error term. For identification purposes, the variance of the error term is fixed to one to set the scale of the underlying latent utility, and its mean is fixed to zero to identify the intercept term, i.e. $\mathrm{E}(\varepsilon_S) = 0$ and $\mathrm{V}(\varepsilon_S) = 1$.[11]

3.1.2. Labor Market Participation and Wages

For each level of education $s \in \{0, 1\}$, the decision D_s to be active on the labor market is assumed to be a threshold-crossing model. The latent utility D_s^* of working depends linearly on a set of covariates X_D through a vector of parameters β_D^s, and on the latent factor θ with its associated factor loading α_D^s:

$$D_s = \mathbf{1}\left(D_s^* > 0\right), \quad D_s^* = X_D\beta_D^s + \theta\alpha_D^s + \varepsilon_D^s \tag{2}$$

For wages, a linear specification with education-specific parameters is assumed:

$$Y_s = X_Y\beta_Y^s + \theta\alpha_Y^s + \varepsilon_Y^s \tag{3}$$

where Y_s represents the log hourly wage, for $s \in \{0, 1\}$, X_Y is a set of observed covariates with associated vector of returns β_Y^s, α_Y^s denotes the return to locus of control, and ε_Y^s is an idiosyncratic error term. In compact form, we denote $D^* = (D_0^*, D_1^*)'$, $D = (D_0, D_1)'$, $Y = (Y_0, Y_1)'$, $\beta_W = (\beta_W^{0\prime}, \beta_W^{1\prime})'$ and $\alpha_W = (\alpha_W^0, \alpha_W^1)'$ for $W = D, Y$. Similarly to the schooling equation, we assume that $\mathrm{E}(\varepsilon_D) = 0$ and $\mathrm{V}(\varepsilon_D) = 1$ for identification of the labor market participation equation, while in the wage equation only $\mathrm{E}\left(\varepsilon_Y^s\right) = 0$ is required.

Locus of control, measured by the latent factor θ, can affect labor market outcomes both directly and indirectly. People with a higher θ (called 'internalizers' in the psychological literature) may earn higher wages, because locus of control is directly rewarded on the market, but also because they achieved a level of education that gave them access to a better segment of the job market in the first place. The direct effect is measured by the factor loadings α_D^s and α_Y^s, for $s \in \{0, 1\}$, while the indirect effect operates through the schooling decision.

[10] See Section 6.1 for the definition of the schooling variable S in our empirical application.

[11] These restrictions are analogous to those usually made in the probit model.

DOI: 10.1002/jae

3.1.3. A Measurement System for Locus of Control

In our empirical application, K ordinal variables obtained from a psychometric test are used to measure locus of control. Each measurement M_k, for $k = 1, \ldots, K$, represents how people rate some statement related to locus of control, on a Likert scale with C_k answers ranging from 'completely disagree' to 'completely agree'. A level of agreement M_k^* is assumed to be underlying the corresponding statement. This unobserved level of agreement is specified to depend linearly on some covariates X_M and on the factor θ, and is discretized by a set of threshold points $\gamma_k = (\gamma_{k,0}, \ldots, \gamma_{k,C_k})'$ to produce the observed measurement M_k:

$$M_k = \sum_{c=1}^{C_k} c\, \mathbf{1}\left(\gamma_{k,c-1} \leq M_k^* < \gamma_{k,c}\right), \quad M_k^* = X_M \beta_{M_k} + \theta \alpha_{M_k} + \varepsilon_{M_k} \tag{4}$$

for $k = 1, \ldots, K$, where β_{M_k} and α_{M_k} denote, respectively, the vector of parameters associated with X_M and the factor loading, and the idiosyncratic term ε_{M_k} captures the remaining unobserved heterogeneity. The cutoff values are ordered such that $\gamma_{k,0} = -\infty < \gamma_{k,1} < \ldots < \gamma_{k,C_k} = +\infty$. To simplify notation, we denote $M = (M_1, \ldots, M_K)'$, $M^* = (M_1^*, \ldots, M_K^*)'$, and $\gamma = (\gamma_1', \ldots, \gamma_K')'$. As in the previous discrete equations (1) and (2), we assume that $\mathrm{E}(\varepsilon_{M_k}) = 0$ and $\mathrm{V}(\varepsilon_{M_k}) = 1$ for identification purposes. These restrictions are, however, not sufficient, as it is still possible to shift both the latent variables M_k^* (through the intercept term) and the cutoff points in the same direction without affecting the likelihood function. Constraining the first cutoff point to be equal to zero removes this indeterminacy, i.e. $\gamma_{k,1} = 0$, for $k = 1, \ldots, K$.[12]

3.1.4. Unobservables of the Model

The overall model consists of a system of $Q = K + 5$ equations that all have an underlying linear structure. To complete the specification of this simultaneous equation model, some independence and distributional assumptions are required. First, the latent factor is assumed to be centered, have finite variance, and be independent of the covariates:

$$\mathrm{E}(\theta) = 0, \quad \mathrm{V}(\theta) = \sigma_\theta^2 \ll \infty, \quad \theta \perp\!\!\!\perp X$$

while the error terms $\varepsilon = (\varepsilon_S, \{\varepsilon_D^s, \varepsilon_Y^s\}_{s=0,1}, \{\varepsilon_{M_k}\}_{k=1,\ldots,K})'$ are assumed to be mutually independent, and independent of the covariates and of the latent factor:

$$\varepsilon \perp\!\!\!\perp X, \quad \varepsilon \perp\!\!\!\perp \theta, \quad \varepsilon_j \perp\!\!\!\perp \varepsilon_{j'} \tag{5}$$

where $X = (X_S, X_D, X_Y, X_M)$, and for all $\varepsilon_j, \varepsilon_{j'} \in \varepsilon$ and $\varepsilon_j \neq \varepsilon_{j'}$. These independence assumptions, required for identification, guarantee that the factor θ captures the unobserved trait of interest (locus of control) and thereby represents the only source of dependence between the outcomes and the measurements, conditional on the observed covariates X. For inference purposes, we rely on a parametric approach and assume normality of the unobservables:

$$\theta \sim \mathcal{N}\left(0;\, \sigma_\theta^2\right), \quad (\varepsilon_S, \varepsilon_D^0, \varepsilon_D^1, \varepsilon_{M_1}, \ldots, \varepsilon_{M_K})' \sim \mathcal{N}(0;\, \Sigma) \tag{6}$$

[12] Alternative normalizations have been proposed. For example, Song and Lee (2014) advocate relaxing the restriction on the variance of the error term and fixing *two* cutoff points to predetermined values to achieve identification.

DOI: 10.1002/jae

where $\sigma_\theta^2 = 1$ and $\Sigma = I_{K+3}$. For the error terms of the wage equation, we relax normality by specifying a mixture of H_s normal distributions with zero mean:

$$\varepsilon_Y^s \sim \sum_{h=1}^{H_s} \pi_h^s \mathcal{N}\left(\mu_h^s; (\omega_h^s)^2\right), \quad \mathrm{E}\left(\varepsilon_Y^s\right) = \sum_{h=1}^{H_s} \pi_h^s \mu_h^s = 0 \tag{7}$$

for $s = 0, 1$, where $\vartheta_h^s = (\pi_h^s, \mu_h^s, \omega_h^s)$ denotes, respectively, the weight, mean and standard deviation of mixture component h for outcome Y_s, and $\vartheta = (\vartheta^0, \vartheta^1)$ with $\vartheta^s = (\vartheta_1^s, \ldots, \vartheta_{H_s}^s)$. Mixtures of normals are widely used as a flexible semiparametric approach to density estimation (Ferguson, 1983; Escobar and West, 1995).

Since θ is not observed but latent, two nontrivial problems need to be addressed to identify the model. The first one concerns the general identification of the factor model with continuous and discrete variables, and can be achieved by imposing appropriate restrictions. The identification of factor models is well documented in the literature (Anderson and Rubin, 1956) and many different strategies have been applied in practice to achieve it (see, for instance, Carneiro *et al.*, 2003). We provide a detailed account of the identification issues in Appendix B (supporting information). The second challenge is related to the particular structure of our dataset, where the model cannot be fully identified in each subsample of the data. Instead, a piecewise identification strategy needs to be implemented to recover all parameters. Section 3.2 explains how to combine data sets for this purpose.

3.2. Data Combination Strategy to Overcome the Missing Data Problem

If all variables are observed for all individuals, the likelihood of the model can be expressed as

$$\begin{aligned} \mathcal{L}(\Psi \mid S, D, Y, M, X) = \int_\Theta & \Pr(S = s \mid \Psi, X_S, \theta) \\ & \times \prod_{s=0}^{1} \left[p(D_s \mid \Psi, X_D, \theta) p(Y_s \mid \Psi, X_Y, \theta)\right]^{\mathbb{1}(S=s)} \\ & \times \prod_{k=1}^{K} p(M_k \mid \Psi, X_M, \theta) p(\theta \mid \Psi) \, \mathrm{d}\theta \end{aligned} \tag{8}$$

where Ψ represents the vector containing all model parameters, $p(\cdot)$ invariantly denotes a density function, and Θ is the support of the distribution of the factor. Standard methods can be applied to estimate the model, after identification has been secured using appropriate restrictions (see Appendix B, supporting information).

Unfortunately, in many applications only subsamples with a subset of the variables of interest can be observed. This is the case, for instance, with life cycle data, where it may be difficult to collect information on all individuals for all time periods. Another example is provided by administrative datasets, which often contain very detailed information about individual labor market behavior, but lack good personality measures, which can only be obtained from survey data. To deal with such cases where missing data problems plague the analysis, we rely on an idea proposed by Cunha *et al.* (2005). It consists of identifying one part of the likelihood in each subsample, getting rid of the unobserved variables by integrating them out of the likelihood.

DOI: 10.1002/jae

Before introducing the mechanisms of the data combination strategy, it is important to lay out the main assumptions we rely on. We assume in this section that two datasets are available to identify and estimate our model. The first one contains observations on education and locus of control measurements, whereas in the second one only education and labor market outcomes are available.[13] Under this assumption, the model can be identified from the covariances of the observed variables, or from their polychoric correlations in the case of the discrete measurements (see Appendix B, supporting information, for details). Moreover, the combined likelihood can be used to recover the parameters by means of maximum likelihood estimation or by using Bayesian methods. In the latter case, the likelihood enters the derivation of the posteriors (see Appendix C.3, supporting information).

Main assumption. *Any observation that contributes to the likelihood of the model (or to different model parts) and that is used for inference is assumed to be generated by the structural model defined by Eqs. 1–4, and to verify the independence assumptions specified in Eq. 5. This should hold for the subset of observed variables among* $S_i, D_{S_i,i}, Y_{S_i,i}, \{M_{k,i}\}_{k=1,\dots,K}$, *for all individuals* $i = 1, \dots, N$.

This assumption can be read as follows. The implementation of the data combination strategy requires that each observation, whether it belongs to the youth sample or to the adult sample, is generated by the same underlying structural model. This needs to hold for the observed variables that are available for inference only, not for those that are missing. The schooling equation—the only one that is common to both subsamples and that allows us to elicit the distribution of the latent factor from the youth sample—therefore plays a central role. Its parameters should be the same across the two subsamples. We come back to this assumption in the online Appendix to this paper (see Appendix J, supporting information). As for the covariates, their distribution does not need to be the same across the two subsamples. It is crucial, however, that the distribution of the latent factor is the same in the youth and in the adult samples, as it is linking the two subsamples through the schooling equation. These assumptions imply that differences in levels are allowed in the observed variables, as long as they can be controlled for by age or cohort effects.[14]

With these assumptions in hand, we can now provide the intuition behind the data combination. First, derive the contribution to the likelihood of a person with higher education. Since her future labor market participation and wage cannot be observed, they are integrated out to provide

$$\begin{aligned}
&\int_{\Theta} \Pr(S = 1 \mid \Psi, X_S, \theta) \left\{ \iint p(D_1 \mid \Psi, X_D, \theta) p(Y_1 \mid \Psi, X_Y, \theta) \mathrm{d}D_1 \mathrm{d}Y_1 \right\} \\
&\times \prod_{k=1}^{K} p(M_k \mid \Psi, X_M, \theta) p(\theta \mid \Psi) \, \mathrm{d}\theta \\
&= \int_{\Theta} \Pr(S = 1 \mid \Psi, X_S, \theta) \prod_{k=1}^{K} p(M_k \mid \Psi, X_M, \theta) p(\theta \mid \Psi) \, \mathrm{d}\theta
\end{aligned}$$

Consequently, only the parameters of the measurement system and of the schooling equation can be identified from the youth sample, using the identification strategy presented in Appendix B (supporting information). This subsample, however, does not provide any information to identify the parameters of the labor market participation and wage equations.

[13] This is the configuration we have in our empirical application, where only education is common to the two subsamples.
[14] This would result in a difference in the intercept terms across the two subsamples only, not in the regression coefficients.

DOI: 10.1002/jae

Similarly, consider a person from the adult sample who did not achieve higher education. Her contribution to the likelihood is

$$\begin{aligned}&\int_{\Theta} \Pr(S=0 \mid \Psi, X_S, \theta)\, p(D_0 \mid \Psi, X_D, \theta) p(Y_0 \mid \Psi, X_Y, \theta) \\ &\times \left\{ \prod_{k=1}^{K} \int p(M_k \mid \Psi, X_M, \theta) dM_k \right\} p(\theta \mid \Psi)\, \mathrm{d}\theta \\ &= \int_{\Theta} \Pr(S=0 \mid \Psi, X_S, \theta)\, p(D_0 \mid \Psi, X_D, \theta) p(Y_0 \mid \Psi, X_Y, \theta) p(\theta \mid \Psi)\, \mathrm{d}\theta\end{aligned}$$

and is obtained by integrating out the measures for pre-market locus of control, as these are not observed for this individual. Identification of the full model is infeasible in this subsample, since no observations on pre-market locus of control are available for the adults. Only the parameters of the outcome equations can be identified from the adult subsample.

Two main observations can be made. First, each subsample allows us to identify a subset of the parameters of the overall model, Ψ^{youth} for the youth sample and Ψ^{adult} for the adult sample:

$$\begin{aligned}\Psi^{\text{youth}} &= \{\beta_S, \alpha_S, \beta_M, \alpha_M, \gamma\} \\ \Psi^{\text{adult}} &= \{\beta_S, \alpha_S, \beta_D, \alpha_D, \beta_Y, \alpha_Y, \vartheta\}\end{aligned}$$

In the end, all model parameters can be recovered by the data combination ($\Psi = \Psi^{\text{youth}} \cup \Psi^{\text{adult}}$). Second, the parameters of the schooling equation are the only ones that are common to the two subsamples ($\Psi^{\text{youth}} \cap \Psi^{\text{adult}} = \{\beta_S, \alpha_S\}$). The parameters can all be identified using the restrictions described in Section 3.1, using the (polychoric) correlations as explained in Appendix B (supporting information) to retrieve the factor loadings.

To fully understand why the data combination strategy works and is a valid approach, it is important to realize that the model could be split into two separate models, and estimated separately with each subsample: youth sample for the measurement system, adult sample for the outcome system. However, doing so would provide no guarantee that the factor extracted from the adult subsample would capture locus of control, as it would not be related to any measurements of this trait. Estimating the two systems of equations simultaneously solves this problem: the schooling equation is common to both systems, thus forcing the latent factor to capture the same trait across the two subsamples and to have the same impact on the schooling decision. Therefore, the schooling equation serves as a bridge between the two subsamples.

4. BAYESIAN INFERENCE

We use a Bayesian approach and rely on Markov chain Monte Carlo (MCMC) methods for the inference of our model (Chib and Greenberg, 1996). These methods consist of drawing the parameters of interest sequentially from their respective conditional distributions to produce a posterior sample that can be used to summarize the results. MCMC methods appear particularly suited to the estimation of factor models, as they allow us to simulate the unobservable components of the model (latent factor and latent utilities of the discrete variables) through data augmentation methods (Tanner and Wong, 1987),[15] thus bypassing the potentially high-dimensional integration problem inherent to these models. Bayesian methods have been widely used in the empirical literature to extract latent traits and

[15] Data augmentation procedures are increasingly used in applied labor market and education research (for recent examples see Horny *et al.*, 2012; Koop and Tobias, 2004; Li, 2006).

DOI: 10.1002/jae

measure their impact on outcomes, with applications in labor economics, economics of education, health economics (Carneiro *et al.*, 2003; Hansen *et al.*, 2004; Cunha *et al.*, 2005; Heckman *et al.*, 2006; Conti and Heckman, 2010). To carry out our inference, we need to specify a prior distribution on the parameters of the model. The priors we use are standard, and are parametrized such that no strong prior information is incorporated into the model. Full details on the prior specification, as well as on the MCMC sampler, can be found in Appendix C (supporting information).

5. MONTE CARLO STUDY

To investigate how the data combination strategy helps in making inference on the model in the case of missing information, we run several experiments using synthetic data generated from the model described in Section 3.1.

5.1. Experimental Design

The data we simulate mimic the real data used in our empirical application, where a subsample of youth individuals allows us to identify the measurement system and the schooling equation, while a subsample of adults identifies the outcome system. In this setup, the schooling equation links the two subsamples. Full details on the setup of the experiments are provided in Appendix D (supporting information).

We vary in different ways the amount of information that is available from the simulated data to identify the distribution of the latent factor. First, we allow the two subsamples to overlap, and look at four scenarios: full overlap (i.e. no missing data problem), no overlap (no individuals with M, S and Y observed simultaneously), and partial overlap (all variables available for 40% or 13% of the individuals, where the latter case corresponds to our empirical application). Intuitively, the more overlap between the two samples, or the larger the number of observations, the easier it should be to extract information to proxy the factor, and thus the more precisely we should be able to measure its impact on the outcomes. Second, the schooling equation is specified either as binary or as continuous,[16] so as to better apprehend the role played by this equation in linking the two data sets. The inference should be facilitated when the schooling equation is continuous rather than binary, as continuous variables bring more information to the table.

5.2. Simulation Results

Table I reports the results of the Monte Carlo study for the factor loadings of the schooling and potential outcome equations. These parameters are the most relevant and the most likely to be affected by the missing data problem. The posterior results are summarized by the means, standard deviations (SDs) and root mean squared errors (RMSEs) of the parameters over the 100 Monte Carlo replications.

The first row of the table shows the results for the optimal scenario where there is no missing data problem (100% sample overlap). This case can be used as a baseline to gauge the performance of our inference procedure in the other cases where data are partly missing. Overall, it is worth noting that our data combination strategy works well in recovering the factor loadings, even when the sample size is small and when there is little overlap between the two subsamples. Particularly interesting is the third case with 13% overlap, which is similar to our empirical application when the schooling variable is binary and $N = 1,500$. In this example, full information (measurements, schooling and outcomes) is available for only 195 individuals, whereas only schooling and outcomes can be observed for 585

[16] In the continuous schooling case, the potential outcomes are observed based on the sign of the schooling variable.

DOI: 10.1002/jae

Table I. Monte Carlo simulation results for 100 replications

		Continuous schooling						Binary schooling					
		$N = 1,500$			$N = 10,000$			$N = 1,500$			$N = 10,000$		
	True	Mean	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE
Sample overlap: **100%**													
α_S	0.4	0.400	0.032	0.032	0.398	0.012	0.013	0.413	0.066	0.067	0.401	0.027	0.027
α_Y^1	0.3	0.300	0.020	0.020	0.300	0.008	0.008	0.300	0.020	0.020	0.300	0.008	0.008
α_Y^0	−0.3	−0.299	0.023	0.023	−0.299	0.010	0.010	−0.298	0.024	0.024	−0.299	0.010	0.010
Sample overlap: **40%**													
α_S	0.4	0.399	0.036	0.036	0.397	0.014	0.014	0.415	0.070	0.072	0.400	0.030	0.029
α_Y^1	0.3	0.295	0.029	0.029	0.300	0.012	0.011	0.294	0.028	0.029	0.300	0.012	0.012
α_Y^0	−0.3	−0.294	0.034	0.034	−0.298	0.013	0.013	−0.293	0.034	0.035	−0.298	0.014	0.014
Sample overlap: **13%**													
α_S	0.4	0.399	0.036	0.036	0.397	0.014	0.015	0.417	0.076	0.077	0.400	0.031	0.031
α_Y^1	0.3	0.287	0.041	0.043	0.300	0.016	0.016	0.282	0.044	0.047	0.299	0.019	0.019
α_Y^0	−0.3	−0.286	0.045	0.047	−0.298	0.018	0.018	−0.279	0.050	0.054	−0.299	0.021	0.021
Sample overlap: **0%**													
α_S	0.4	0.408	0.042	0.042	0.399	0.015	0.015	0.428	0.098	0.101	0.403	0.036	0.036
α_Y^1	0.3	0.252	0.068	0.083	0.291	0.031	0.032	0.080	0.096	0.240	0.227	0.064	0.097
α_Y^0	−0.3	−0.249	0.063	0.081	−0.285	0.031	0.034	−0.091	0.085	0.225	−0.218	0.064	0.104

Note: Monte Carlo mean, standard deviation (SD) and root mean squared error (RMSE) computed over 100 replications.

other individuals. Still, the procedure succeeds in drawing satisfactory inference on the factor loadings, with reasonable levels of bias and RMSE.

Comparison of the different cases reveals that the Monte Carlo SDs and RMSEs of the loadings always increase faster for the parameters of the outcome equations compared to the schooling equation when the sample overlap decreases. This, however, does not indicate a worse impact of the missing data problem on α_Y^1 and α_Y^0 compared to α_S. It only reflects a mechanical effect, as the number of observations is constant in the schooling equation, whereas for the outcomes it decreases by construction when the overlap decreases, resulting in a loss of precision.[17]

Most importantly, these simulations allow us to understand how much information can be extracted from the data to identify the model. The size of the overlap sample and the type of the schooling variable play an important role: the larger the overlap sample, the more information is available to identify to full likelihood function, and therefore the more precise the results. Similarly, more information is available when the schooling variable is continuous compared to the binary case, which improves the inference. Conversely, when less information is available (smaller overlap and/or binary schooling equation), the results become less precise and a bias emerges. This is particularly striking in the case with no sample overlap, displayed in the last row of Table I. This extreme case with no overlap is more problematic. There is clear evidence of an attenuation bias, particularly visible in small samples and in the binary schooling case. Even with as many as 10,000 observations, this bias remains substantial in the binary case. This result is, however, not surprising: in this particular setup of the model, the outcome system is not identified in the adult sample. Therefore, the only source of identification is the schooling equation—identified from the youth sample—when there is no sample overlap. In this context, the inference becomes a difficult task when the available information is very limited.

Overall, these results are very encouraging and have important implications. They show that the data combination strategy can successfully be implemented when only subsamples that partially identify

[17] Note that Y_1 and Y_0 are potential outcomes, such that only one of them is observed for each individual.

DOI: 10.1002/jae

the likelihood are available. The more information, the better the inference, but even challenging cases where sample size or sample overlap are limited can be addressed in a satisfactory way. Nevertheless, the analyst interested in applying this approach should bear in mind its limitations, and first thoroughly probe the amount of information that can be used. Even if the model is theoretically identified by the data combination strategy, the empirical identification can be more difficult. Our simulations provide evidence of a potential attenuation bias affecting the loadings in the outcome system. This can be a concern when the posterior distributions of the parameters are used for posterior inference, for instance to compute treatment effects (Heckman *et al.*, 2014b).

6. EMPIRICAL APPLICATION

In this section, we investigate how *pre-market* locus of control affects schooling decisions and wages in Germany. Although the SOEP is a longitudinal study, youth are surveyed since 2000 only, and only a small fraction of our sample has entered the labor market in 2011. To conduct our analysis, we therefore face a dilemma that is common to many datasets: on the one hand, we have a large dataset of working-age people ('adult sample'). Schooling decision and labor market outcomes are observed for these people, but no information is available on their locus of control at the time of schooling. On the other hand, a sample of 17-year-olds is available ('youth sample'), including pre-market locus of control measurements and schooling decision. Labor market outcomes are only available for a very small group among the youths (mostly of low-educated individuals, constituting the 'overlap sample'). We show that the overlapping adult and youth samples can be combined to overcome the missing data problem to elicit *pre-market* locus of control.

We also assess the robustness of our approach by comparing our results to those obtained from more traditional estimation methods. More specifically, we take advantage of the fact that the SOEP contains measures of locus of control administered at the same time as the labor market outcomes, i.e. after the individuals made their schooling decisions. We use these measures to re-estimate the model and compare our results to what researchers usually obtain using these contemporaneous measurements (results available in Appendix I, supporting information).

6.1. Data and Sample Construction

We draw a combined sample of 1901 youths (age 17–27) and 1606 adults (age 25–35). Information for the youth sample was collected in the years 2001–2011, when the subjects were 17 years of age.[18] For the adults, we take most information from the 2011 wave, but replace missing data points with information from waves 2004–2010.[19] Note that we choose a narrow age window for the adult sample and exclude individuals from East Germany. This way we ensure that the individuals in our data are very similar and that the common coefficient assumption across the two samples is justified. Note that education is measured for both samples, but that labor market outcomes are available only for the adult sample and for a small part of the youth sample. The sample sizes of the youth, adult and overlap samples (i.e. those for whom we dispose of youth measurements and adult labor market information) are displayed in Table II.[20] Our Monte Carlo Study in the previous section, tailored to the sample size of our real dataset, showed that our approach provides good results, despite the relatively small sample size.

[18] This is an advantage of the SOEP compared to for example the NLSY, where personality is administered when many individuals have already entered the labor market.

[19] We convert all wage information to 2009 price levels.

[20] The middle part of Table II shows that, among the youth sample, there is a slightly higher fraction of highly educated individuals. As part of the robustness section we show that this is a trend effect which can be accommodated by including age and cohort dummies as covariates in the education, employment and wage equations.

DOI: 10.1002/jae

Table II. Sample sizes and descriptive statistics

Sample size

	Youths	Adults	Total
Females	949	583	1,532
Males	962	622	1,584

Overlap (youth LOC and adult wages)

		percent
Females	197	12.86%
Males	208	13.13%

Higher education

	Sample	Mean	N
Females	Youths	0.53	949
Males		0.46	962
Females	Adults	0.48	794
Males		0.44	812

Wages (adult sample)

	Low education		High education		
	Mean	N	Mean	N	p-value
Females	13.05	334	10.59	404	0.00
Males	15.63	282	11.39	502	0.00

Source: SOEP youth and adult samples 2000–2011.
Note: Wages are obtained using the most recent wage information from waves 2004–2011. p-values of a two-sided t-test for differences in means are reported.

We define higher education as having obtained at least an academic high school degree (German *Abitur* or *Fachabitur*), which entitles students to attend German higher education institutions.[21] Summary statistics of the education variable in the two samples are presented in the middle part of Table II. Summary statistics on hourly wages of the combined sample can be found in the bottom panel of Table II. The table displays that males earn higher wages than females, and that the observed wage gap between high- and low-educated individuals is higher for males than for females.

6.1.1. Locus of Control Measurements

Much of the more recent psychological literature on locus of control centers on the correct way to measure locus of control and its dimensionality. The original locus of control scale, as developed by Rotter (1966), is a single unidimensional concept with control tendencies ranging from internal (individuals who believe what happens to them in life is related to their own actions and decisions) to

[21] We use the final education outcome whenever possible. Only for those individuals who were still in school in 2011, we replace this with the most likely education outcome using current school track and the aspired education level. For a detailed exposition of the coding of all variables, see Appendix E (supporting information).

DOI: 10.1002/jae

external (individuals who believe events in their life derive primarily from fate or luck). Since the early 1970's, however, there exists an ongoing debate about the true underlying factor structure and today most psychometric scales measure two underlying traits, called *internal* locus of control and *external* locus of control, respectively (Furnham and Steele, 1993). Importantly, psychometric measurement scales have to be set up differently depending on the intended number of dimensions.

The locus of control scale in the SOEP youth questionnaire, as developed by Nolte *et al.* (1997), was designed to reflect two dimensions of locus of control: fairness and social involvement. Overall, the questionnaire comprises 10 items and each question is answered on a Likert scale ranging from 1 ('disagree completely') to 4 ('agree completely'). Items Q3 ('Success is a matter of fate or luck'), Q4 ('Others decide about my life'), Q6 ('In case of difficulties, doubts about own abilities'), Q7 ('Possibilities in life depend on social conditions') and Q9 ('I have little control over what happens to me') describe external locus of control tendencies. Three other items (Q1, Q5 and Q8) describe internal locus of control tendencies (Weinhardt and Schupp, 2011). Moreover, two items (Q2 and Q10) are part of the questionnaire that describe an individual's sense of fairness and her degree of social involvement (for a summary table of all locus of control questions and items see Table F.5, supporting information). These measures are related to locus of control, but not part of the original concept as developed by Rotter (1966).

An exploratory factor analysis shows that locus of control is indeed best represented by two factors. A principal component analysis yields two eigenvalues larger than 1. Similarly, a scree plot analysis suggests two underlying factors.[22] The results of the exploratory factor analysis show that researchers who use an index (i.e. forcing each of the measurement items to enter the index with an equal weight) construct a locus of control measure that is potentially flawed by measurement error, and therefore to coefficients that suffer from attenuation bias (Heckman *et al.*, 2013).

Given the results of the exploratory factor analysis, we decide to follow the original psychometric design of the locus of control scale in the SOEP youth questionnaire, as well as the recommendations in Weinhardt and Schupp (2011), and extract a single external locus of control factor. This has two advantages. First, the reliability of the internal locus of control scale, which we do not retain, is problematically low in the SOEP.[23] Second, external locus of control relates to adjustment and achievement levels, while internal locus of control is more related to self-esteem and self-acceptance (Furnham and Steele, 1993). Not surprisingly, external locus of control is therefore also more predictive of economic outcomes (see results in Caliendo *et al.*, 2015; Heineck and Anger, 2010). For ease of interpretation, we normalize the model such that lower scores of the latent factor are associated with a more external locus of control, and higher scores with a less external locus of control.[24]

6.1.2. Covariates

To make the samples comparable and to account for family background, socioeconomic status and labor market conditions, we include cohort dummies, a large range of background variables and local unemployment rates in our schooling and labor market equations. With respect to the measurement system for locus of control, a potential limitation of our approach may be that all locus of control measures were administered at the age of 17. They may therefore capture *pre-market* locus of control, but not necessarily *pre-compulsory-school-track* locus of control. One way to tackle this problem is to purge locus of control of certain schooling variables in the measurement system. This can be done by adding these variables as covariates in the schooling equation.

[22] See Appendix F (supporting information) for more detailed results of the exploratory factor analysis.

[23] We tried to re-estimate our model with internal locus of control, but the model fails to converge, presumably because the very low intra-item correlation (Cronbach's $\alpha = 0.32$ in our sample).

[24] We later conduct robustness checks using all locus of control items in the data. We find that the use of all items simultaneously does not have a significant impact on the results, because the factor loadings on the internal items converge to zero.

DOI: 10.1002/jae

A particularity of the German education system is that student tracking already takes place after fourth grade. At that time, teachers have to evaluate the cognitive ability of each student and have to recommend a school track depending on what they think is the type of curriculum (i.e. secondary school type) a student is able to master. The track recommendation is therefore perceived as a very strong signal of a child's cognitive ability and in many of the federal states (Länder) this recommendation is binding.[25] Hence, to proxy cognitive skills, and to account for the fact that cognition might affect the items revealing pre-market locus of control, we include the primary school teacher track recommendation as a control variable in the measurement system. Moreover, in the measurement system and the schooling equation, we control for parental education, a large set of other background variables, region fixed effects (that capture school quality) and home investments.[26] We differentiate between three West German regions to account for the north–south divide in Germany and exclude all individuals who went to school in East Germany (the former German Democratic Republic) from the sample, to make sure that our results are not flawed by post-1991 schooling and labor market adjustments (Lammers, 2003).[27] We also exclude all individuals from our sample with missing locus of control measures, missing schooling information, or missing information on the covariates. A detailed description of the coding of all variables and summary can be found in Appendix E (supporting information).

6.2. Empirical Results

We run our sampler on the model separately for males and females, using a total of 22,000 MCMC iterations, where the first 2,000 iterations are discarded as burn-in period. To ensure convergence, we generate trace plots and compute Geweke convergence diagnostics separately for each parameter (Geweke, 1992). The results are presented and discussed in two stages. We first provide a description of the main findings in Section 6.2.1, with an emphasis on the impact of locus of control when the model is estimated with pre-market and contemporaneous measures, respectively. In Section 6.2.2 we discuss the empirical implications of our model results.

6.2.1. Factor Loadings

The factor loadings, displayed in Table III, express how the different measurements and outcomes are driven by the latent factor.[28] The larger the magnitude of the loadings, the higher is the contribution of the corresponding measurements to the distribution of the latent factor. In the education, employment and wage equations, the loadings measure the impact of the factor on the respective outcomes. The table compares results when using pre-market and contemporaneous measures. The variance of the factor is fixed to one for identification, such that cross-model comparisons of loadings and their credible intervals are straightforward.

In the measurement system, pre-market and contemporaneous items load about equally on the underlying external locus of control factor.[29] In the outcome system of equations, the factor loading of the education equation is always positive and far from zero across specifications that use youth (pre-market) and adult (contemporaneous) measures of locus of control. For highly educated

[25] Unfortunately, we do not observe the actual secondary school track a student attends. However, track recommendation tends to be a very close proxy of the actual track type chosen, because it is binding in many federal states and closely followed in others.

[26] We later also present robustness checks where, in addition to track recommendation, we use school grades in the measurement system.

[27] The regions are defined as follows. North comprises the federal states Berlin, Bremen, Hamburg, Lower Saxony and Schleswig-Holstein. South includes Bavaria and Baden-Württemberg and West includes Hessen, North Rhine-Westphalia, Rhineland-Palatinate and Saarland.

[28] An analysis of model fit can be found in Appendix H (supporting information).

[29] For the regression coefficients, posterior results can be found in Tables G.6 and G.7 in Appendix G (supporting information).

DOI: 10.1002/jae

Table III. Posterior results for the factor loadings using youth (pre-market) measures and adult (contemporaneous) measures respectively

	Males				Females			
	Pre-market measures		Contemporaneous measures		Pre-market measures		Contemporaneous measures	
Measurement system: locus of control items								
Q3	−0.580***	(0.062)	−0.504***	(0.041)	−0.452***	(0.054)	−0.504***	(0.046)
Q4	−0.880***	(0.096)	−0.868***	(0.057)	−0.736***	(0.071)	−0.718***	(0.059)
Q6	−0.512***	(0.056)	−0.705***	(0.049)	−0.557***	(0.060)	−0.504***	(0.045)
Q7	−0.430***	(0.053)	−0.404***	(0.039)	−0.396***	(0.053)	−0.352***	(0.041)
Q9	−0.771***	(0.083)	−1.206***	(0.096)	−1.231***	(0.178)	−1.032***	(0.095)
Education choice								
S	0.191***	(0.070)	0.115**	(0.045)	0.215***	(0.068)	0.192***	(0.050)
Labor market participation								
D_0	0.144	(0.100)	0.334***	(0.085)	0.192*	(0.109)	0.184**	(0.090)
D_1	−0.068	(0.134)	−0.071	(0.114)	−0.005	(0.114)	0.201*	(0.111)
Log wages								
Y_0	0.061*	(0.036)	0.072***	(0.020)	0.080**	(0.037)	0.042*	(0.024)
Y_1	−0.066	(0.062)	0.013	(0.034)	0.012	(0.051)	0.069**	(0.029)

Note: Standard deviations of the posterior distributions in brackets. Posterior check: */**/*** if zero lies outside the 90%/95%/99% confidence interval of the posterior distribution of the corresponding parameter.

individuals, we find that the direct impact of pre-market locus of control on employment decisions and wages is mostly close to zero or even negative (columns 1 and 3). For highly educated males and females, we find that contemporaneous locus of control effects on wages tend to be larger than pre-market locus of control effects and more often distinct from zero (columns 2 and 4). For low-educated individuals we also find that the impact of contemporaneous locus of control on wages tends to be bigger, with a more concentrated posterior (i.e. more precise), but this difference seems to be larger for males. Note that the difference in posterior means can be large, as is the case for highly educated individuals, but for most cases the posterior distributions do not differ dramatically across pre-market and contemporaneous measures. Yet differences in the size and precision of the estimates would likely induce researchers to conclude that locus of control is important for wages when contemporaneous measures are used and unimportant when pre-market measures are used.

Several points can explain the differences between the pre-market and contemporaneous locus of control results. First, our pre-market locus of control results are somewhat conservative due to potential attenuation bias and a loss in precision we found for small samples in Section 5. Given the results of our Monte Carlo experiment, however, this effect seems negligible. A second concern, raised by Cobb-Clark and Schurer (2013), is an attenuation bias which results from an errors-in-variables problem. The idea is that if locus of control changes randomly over time, pre-market locus of control becomes a noisy measure of contemporaneous locus of control, and effect sizes are attenuated due to classical measurement error bias (for a discussion of this effect see Section 2.1 in Cobb-Clark and Schurer, 2013). Third, the results that use contemporaneous measures might be explained by endogeneity bias, i.e. because locus of control is influenced by past labor market shocks. This last effect motivated our approach and is particularly relevant in our sample of young labor market entrants (evidence of personality changes in response to initial labor market experiences is provided in Roberts *et al.*, 2003). Concluding, we can say that as long as personality traits cannot be randomly assigned, the question of whether pre-market or contemporaneous measures should be used is a decision that has to be made by the researcher. It depends on whether the latent personality construct is likely to be subject to endogeneity bias and on the extent to which researchers want to be conservative in the results

DOI: 10.1002/jae

they report. In the following, we decide to concentrate on the effect of *pre-market* locus of control on schooling decisions and wages, which is in line with the empirical contribution of the paper.

6.2.2. Simulation of the Effect of Pre-market Locus of Control

To shed more light on the implications of our model, we need to go beyond the mere interpretation of the factor loadings. Their posterior distribution reveals an impact of locus of control on the outcomes, but is quite uninformative regarding the magnitude of this impact.[30] Since the effects of pre-market locus of control are intertwined and potentially operate through different channels on wages, the best way to understand our model is to simulate it.

To gain more insight into the impact of pre-market locus of control on later outcomes, we can investigate how education outcomes and wages of a given individual would be affected if she were exogenously moved along the distribution of the latent factor, for a given set of observed characteristics X (Heckman *et al.*, 2006).[31] The top part of Figure 1 illustrates that moving the mean individual from the first to the last decile of the distribution results in a 0.2 point increase in the probability of achieving higher education, for both males and females. This is in line with the current psychological literature, which shows that neuroticism is a strong predictor of education and test outcomes (Almlund *et al.*, 2011). At first sight, the effect of locus of control on the education choice seems large. For instance, the average male individual would be 20 percentage points more likely to attend higher education in the highest versus the lowest decile. However, it is unrealistic to see an individual move all the way across the distribution. Cobb-Clark and Schurer (2013), for example, find for a sample of older adults that a worsening of finances reduces locus of control by 20% of a standard deviation and Gottschalk (2005) reports that a welfare-to-work program reduced the probability of agreeing with external answers by 15–39 percentage points. Also, the top part of Figure 1 shows that in the middle of the distribution the locus of control effect is much smaller. Concerning wages, the middle part of Figure 1 shows that for a given level of schooling the posterior distribution of the factor loading for low-educated individuals is diffuse and the effect is nonexistent for the highly educated. Moreover, the bottom part of Figure 1 shows that the combined effect of locus of control through education and labor market returns of moving the mean individual from the first to the last decile amounts to a maximum wage return of 6% for males and of 18% for females. However, the confidence bands do not exclude the possibility that the true effect of such a move is zero. From this we can conclude that pre-market locus of control increases wages only slightly, and that the effect operates merely through higher education.

Our results seem somewhat contrary to the more direct link between locus of control and wages that has been found in some of the literature (Heckman *et al.*, 2006; Heineck and Anger, 2010). Three different answers can be put forward to address this apparent contradiction. First, the term 'noncognitive skills' is very often used as a generic expression encompassing many different personal abilities and traits, sometimes leading to confusion. A comparison of results is possible only if the same concept is used. For instance, Heckman *et al.* (2006) find a significant effect of noncognitive skills on wages. However, they use a single underlying factor for noncognitive skills constructed from two psychometric tests, namely the Rosenberg self-esteem scale and the Rotter scale. This composite factor thus captures a different dimension from our factor, especially since it loads more on the self-esteem scale than on the locus of control scale in their empirical study. Second, we only look at a sample of young labor market entrants. At this stage, wage setting is likely to be merely a function of formal qualifications. Hence only after individuals have entered the labor market does a complex dynamic interaction process begin. While working on-the-job, individuals learn about their abilities, and at the same time employers adapt their knowledge about an individual's locus of control. Third, and most

[30] In frequentist terms, we would look at the statistical significance, which is different from the economic significance; see, for example, McCloskey and Ziliak (1996).

[31] For details on how we simulate outcomes see Appendix H.1 (supporting information).

DOI: 10.1002/jae

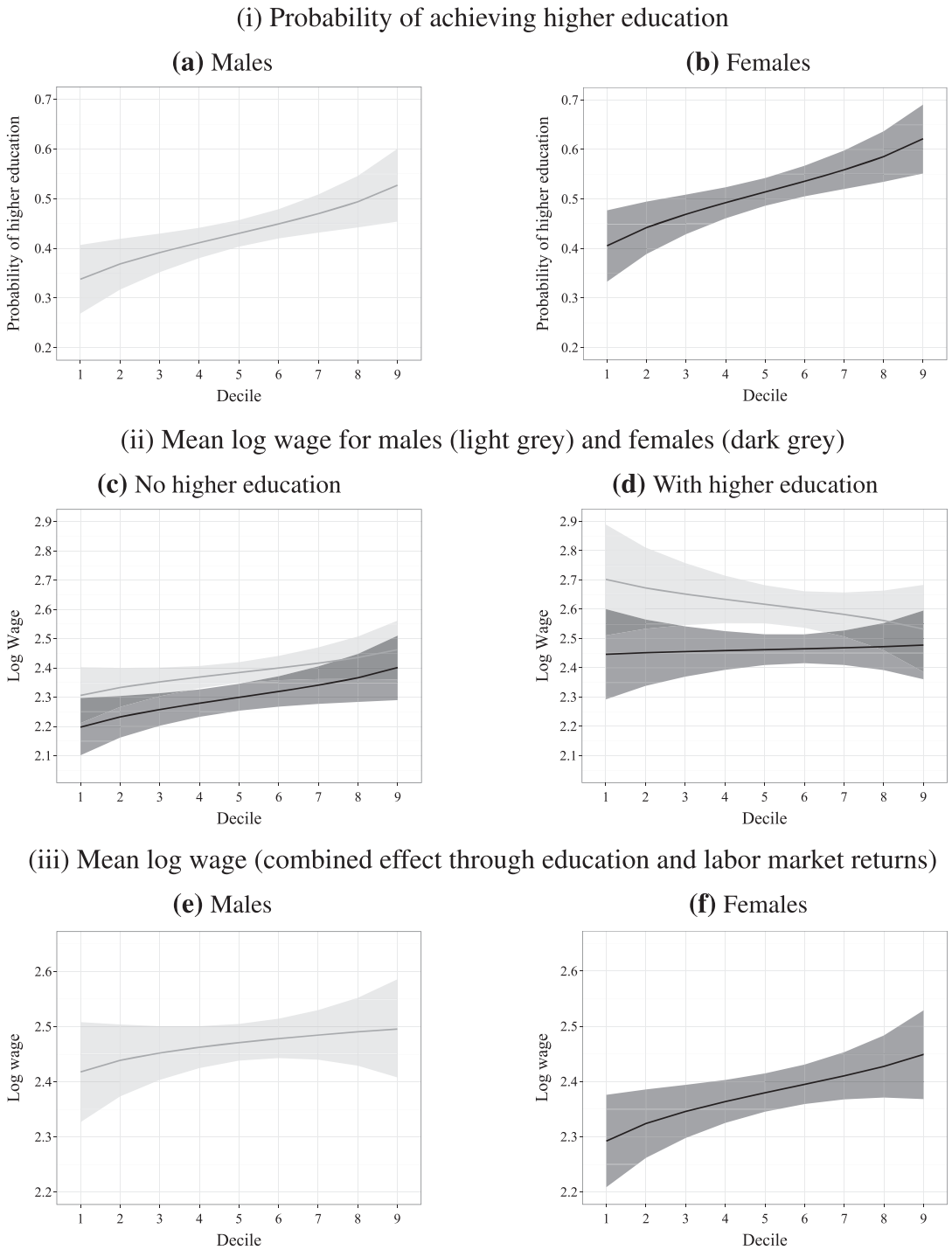

Figure 1. Education and wages by decile of the factor distribution. Simulation from the posterior draws of the model using 20,000 replications, for the mean individual of the sample. Shaded areas show 95% confidence bands.

DOI: 10.1002/jae

importantly, we focus on *pre-market* locus of control as a measure of locus of control that is independent of labor market experience. As a consequence, our findings differ from the results presented by Heineck and Anger (2010), who find a strong and significant impact of locus of control on wages, even after controlling for education.[32]

7. CONCLUSIONS

This paper applies and investigates the performance of a new data combination strategy that relies on factor structure models. It makes it possible to extract information on personality measures from one sample, and to measure the impact of the extracted traits on outcomes observed in another sample. Our approach can be a solution to measurement error and endogeneity problems, in situations where researchers are confronted with truncated life cycle data or outcome data with missing personality measurements. These are well-known and challenging problems for researchers who work at the intersection of psychology and economics.

We contribute to the literature by thoroughly explaining and illustrating the identification issues at stake. We do this both theoretically and empirically through a Monte Carlo study and a real data application. Our results indicate that the method works well in practice, even in relatively small samples, as long as enough information can be extracted and combined from the subsamples to identify the complete likelihood of the model. In extreme situations (e.g. too small sample size, insufficient data overlap), we provide evidence of a potential attenuation bias on the factor loadings. The empirical analysis establishes that an individual's *pre-market* locus of control substantially raises the probability of choosing higher education—a one standard deviation change in locus of control results in a 5% point difference in the probability of obtaining an upper secondary school certificate. The effect on wages, however, merely operates through schooling. Increasing locus of control by one standard deviation results in an overall hourly wage increase of about 2% for males and 4% for females. After controlling for education, this effect becomes much smaller. Using contemporaneous measurements (as is common practice in the literature) tends to result in larger and more precisely estimated impacts. The difference is large enough for researchers to conclude that locus of control is important for wages when contemporaneous measures are used, and unimportant when pre-market measures are used.

Data combination methods for factor structure models open up new horizons for applied researchers. They allow us to tackle a persistent problem in empirical research: the scarcity of rich datasets required to address important issues. The procedure can, for instance, be used to deal with simultaneity bias and reverse causality problems by combining early-life personality measures with later-life outcomes. Importantly, such analyses are not limited to locus of control but may focus on many other human capital measures such as cognitive abilities, mental health or measures of well-being. In addition, our method could be used to combine information on labor market outcomes from large administrative datasets, which often do not contain good individual-level data, with high-quality personality measures, possibly administered by psychologists. The present study may therefore motivate the use of more and richer human capital measures in empirical economic models.

ACKNOWLEDGEMENTS

We thank the co-editor Edward Vytlacil and four anonymous referees for their valuable comments and suggestions, which helped us improve considerably the original manuscript. We are also grateful to James Heckman and the members of his research group at the University of Chicago, as well as Gerard van den Berg, François Laisney, Winfried Pohlmeier, Friedhelm Pfeiffer, Arne Uhlendorff, Tim

[32] Another reason could be that the authors do not estimate separate models by education level. Yet Table III shows that even if wage equations are estimated separately by education levels the use contemporaneous measurements yields higher and more significant estimates.

DOI: 10.1002/jae

Kautz, Miriam Gensowski, Verena Niepel and Ina Drepper for very helpful comments and stimulating discussions. Pia Pinger's research was supported by the Network 'Noncognitive Skills: Acquisition and Economic Consequences' funded by the Leibniz Association. A first version of this paper was written while Rémi Piatek was working at the Chair of Economics and Econometrics at the University of Konstanz. Financial support by the State of Baden-Württemberg (LGFG-scholarship) and by the German Academic Exchange Service (DAAD) is gratefully acknowledged by Rémi Piatek.

REFERENCES

Aakvik A, Heckman JJ, Vytlacil EJ. 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* **125**: 15–51.

Almlund M, Duckworth AL, Heckman J, Kautz T. 2011. Personality psychology and economics. In *Handbook of the Economics of Education*, Hanushek EA, Machin S, Woessmann L (eds), Vol. 4. Elsevier: Amsterdam; 1–181.

Anderson T, Rubin H. 1956. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Neyman J (ed), Vol. 5. University of California Press: Berkeley, CA; 111–150.

Andrisani PJ. 1977. Internal–external attitudes, personal initiative, and the labor market experience of black and white men. *Journal of Human Resources* **12**: 308–328.

Andrisani PJ. 1981. Internal–external attitudes, sense of efficacy, and labor market experience: a reply to Duncan and Morgan. *Journal of Human Resources* **16**: 658–666.

Borghans L, Duckworth AL, Heckman JJ, ter Weel B. 2008. The economics and psychology of personality traits. *Journal of Human Resources* **43**: 972–1059.

Bowles S. 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. *Journal of Economic Literature* **36**: 75–111.

Bowles S, Gintis H. 2002. Schooling in capitalist america revisited. *Sociology of Education* **75**: 1–18.

Bowles S, Gintis H, Osborne MA. 2001. The determinants of earnings: a behavioral approach. *Journal of Economic Literature* **39**: 1137–1176.

Caliendo M, Cobb-Clark DA, Uhlendorff A. 2015. Locus of control and job search strategies. *Review of Economics and Statistics* **97**(1): 88–103.

Carneiro P, Hansen K, Heckman JJ. 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review* **44**: 361–422.

Cebi M. 2007. Locus of control and human capital investment revisited. *Journal of Human Resources* **42**: 919–932.

Chib S, Greenberg E. 1996. Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory* **12**: 409–431.

Cobb-Clark DA. 2015. Locus of control and the labor market. *IZA Journal of Labor Economics* **4**(3): 1–19.

Cobb-Clark DA, Schurer S. 2013. Two economists' musings on the stability of locus of control. *Economic Journal* **123**: 358–400.

Cobb-Clark DA, Kassenboehmer SC, Schurer S. 2014. Healthy habits: the connection between diet, exercise, and locus of control. *Journal of Economic Behavior and Organization* **98**: 1–28.

Coleman M, DeLeire T. 2003. An economic model of locus of control and the human capital investment decision. *Journal of Human Resources* **38**: 701–721.

Conti G, Heckman JJ. 2010. Understanding the early origins of the education–health gradient: a framework that can also be applied to analyze gene–environment interactions. *Perspectives on Psychological Science* **5**: 585–605.

Conti G, Frühwirth-Schnatter S, Heckman JJ, Piatek R. 2014. Bayesian Exploratory Factor Analysis. *Journal of Econometrics* **183**: 31–57.

Cunha F, Heckman JJ. 2007. Identifying and estimating the distributions of ex post and ex ante returns to schooling: A survey of recent developments. *Labour Economics* **14**: 870–893.

Cunha F, Heckman JJ. 2008. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* **43**: 738–782.

Cunha F, Heckman JJ, Navarro S. 2005. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers* **57**: 191–261.

DOI: 10.1002/jae

Duncan G, Morgan J. 1981. Sense of efficacy and subsequent change in earnings: a replication. *Journal of Human Resources* **16**: 649–657.
Duncan GJ, Dunifon R. 2012. 'Soft-skills' and long-run labor market success. *Research in Labor Economics* **35**: 313–339.
Escobar MD, West M. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**: 577–588.
Ferguson TS. 1983. Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, Chernoff H, Rizvi M, Rustagi J, Siegmund D (eds). Academic Press: New York; 287–302.
Flossmann AL, Piatek R, Wichert L. 2007. Going beyond returns to education: the effect of noncognitive skills on wages in Germany, *Paper presented at the European Meeting of the Econometric Society*: Budapest.
Frölich M. 2008. Statistical treatment choice: an application to active labor market programs. *Journal of the American Statistical Association* **103**: 547–558.
Furnham A, Steele H. 1993. Measuring locus of control: a critique of general, children's, health-and work-related locus of control questionnaires. *British Journal of Psychology* **84**: 443–479.
Gallo W, Endrass J, Bradley E, Hell D, Kasl S. 2003. The influence of internal control on the employment status of German workers. *Schmollers Jahrbuch* **123**: 71–81.
Gensowski M. 2014. Personality, IQ, and lifetime earnings. Discussion paper 8235, Institute for the Study of Labor (IZA).
Geweke J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds), Vol. 4. Oxford University Press: Oxford; 169–193.
Goldberg LR. 1993. The structure of phenotypic personality traits. *American Psychologist* **48**: 26–34.
Gottschalk P. 2005. Can work alter welfare recipients' beliefs? *Journal of Policy Analysis and Management* **24**: 485–498.
Hansen KT, Heckman JJ, Mullen KJ. 2004. The effect of schooling and ability on achievement test scores. *Journal of Econometrics* **121**: 39–98.
Heckman JJ, Stixrud J, Urzua S. 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24**: 411–482.
Heckman J, Pinto R, Savelyev P. 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* **103**: 2052–2086.
Heckman JJ, Humphries JE, Urzúa S, Veramendi GF. 2014a. Education, health and wages. NBER working Paper 19971.
Heckman JJ, Lopes HF, Piatek R. 2014b. Treatment effects: a Bayesian perspective. *Econometric Reviews* **33**: 36–67.
Heineck G, Anger S. 2010. The returns to cognitive abilities and personality traits in Germany. *Labour Economics* **17**: 535–546.
Horny G, Mendes R, Van den Berg GJ. 2012. Job durations with worker-and firm-specific effects: MCMC estimation with longitudinal employer–employee data. *Journal of Business and Economic Statistics* **30**: 468–480.
Judge TA, Bono JE. 2001. Relationship of core self-evaluations traits-self-esteem, generalized self-efficacy, locus of control, and emotional stability-with job satisfaction and job performance: a meta-analysis. *Journal of Applied Psychology* **86**: 80–92.
Judge TA, Erez A, Bono JE, Thoresen CJ. 2002. Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct. *Journal of Personality and Social Psychology* **83**: 693–710.
Kassenboehmer SC, Haisken-DeNew JP. 2009. You're fired! The causal negative effect of entry unemployment on life satisfaction. *Economic Journal* **119**: 448–462.
Koop G, Tobias J. 2004. Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics* **19**: 827–849.
Lammers K. 2003. Süd-Nord-Gefälle in West- und Ostdeutschland. *Wirtschaftsdienst* **83**: 736–739.
Lefcourt HM. 1966. Internal versus external control of reinforcement: a review. *Psychological Bulletin* **65**: 206–220.
Lefcourt HM. 1991. Locus of control. In *Measures of personality and social psychological attitudes*, Robinson JP, Shaver PR, Wrightsman LS (eds), Vol. 1. Academic Press: San Diego, CA; 413–499.
Li M. 2006. High school completion and future youth unemployment: new evidence from high school and beyond. *Journal of Applied Econometrics* **21**: 23–53.
McCloskey DN, Ziliak ST. 1996. The standard error of regressions. *Journal of Economic Literature* **34**: 97–114.

DOI: 10.1002/jae

Ng TW, Sorensen KL, Eby LT. 2006. Locus of control at work: a meta-analysis. *Journal of Organizational Behavior* **27**: 1057–1087.
Nolte H, Weischer C, Wilkesmann U, Maetzel J, Tegethoff H. 1997. Kontrolleinstellungen zum Leben und zur Zukunft: Auswertung eines neuen sozialpsychologischen Itemblocks im sozio-ökonomischen Panel. Working paper, Faculty of Social Sciences, Ruhr University.
Osborne MA. 2000. The power of personality: labor market rewards and the transmission of earnings. PhD thesis, University of Massachusetts.
Ridder G, Moffitt R. 2007. The econometrics of data combination. In *Handbook of Econometrics*, Heckman J, Leamer E (eds), Vol. 6B. Elsevier: Amsterdam; 5469–5547.
Roberts BW, Caspi A, Moffitt TE. 2003. Work experiences and personality development in young adulthood. *Journal of personality and social psychology* **84**: 582–593.
Rotter JB. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* **80**: 1–28.
Rotter JB. 1990. Internal versus external control of reinforcement: a case history of a variable. *American Psychologist* **45**: 489–493.
Schurer S. 2014. Bouncing back from health shocks: locus of control, labor supply, and mortality. IZA discussion paper 8203.
Song XY, Lee SY. 2014. Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology* **57**: 29–52.
Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**: 528–540.
Theodossiou I. 1998. The effects of low-pay and unemployment on psychological well-being: a logistic regression approach. *Journal of Health Economics* **17**: 85–104.
Wang Q, Bowling NA, Eschleman KJ. 2010. A meta-analytic examination of work and general locus of control. *Journal of Applied Psychology* **95**: 761–768.
Weiner B. 2012. An attribution theory of motivation. In *Handbook of Theories of Social Psychology*, Vol. 1. Sage: London; 135–155.
Weinhardt M, Schupp J. 2011. Multi-Itemskalen im SOEP Jugendfragebogen. Data documentation 60. German Institute for Economic Research, DIW Berlin.

DOI: 10.1002/jae