Noncognitive Skills and Labor Market Outcomes: A Machine Learning Approach*

Jana Marečková† University of Konstanz

Winfried Pohlmeier‡ University of Konstanz, COFE, RCEA

This Version: May 1, 2017

Abstract

We study the importance of noncognitive skills in explaining differences in the labor market performance of individuals by means of machine learning techniques. Unlike previous empirical approaches centering around the within-sample explanatory power of noncognitive skills our approach focuses on the out-of-sample forecasting and classification qualities of noncognitive skills. Moreover, we show that machine learning techniques can cope with the challenge of selecting the most relevant covariates from big data with a whopping number of covariates on personality traits. This enables us to construct new personality indices with larger predictive power.

In our empirical application we study the role of noncognitive skills for individual earnings and unemployment based on the British Cohort Study (BCS). The longitudinal character of the BCS enables us to analyze predictive power of early childhood environment and early cognitive and noncognitive skills on adult labor market outcomes. The results of the analysis show that there is a potential of a long run influence of early childhood variables on the earnings and unemployment.

*Financial support by the German Research Foundation (DFG) through research unit FOR 1882 Psychoeconomics and the Graduate School of Decision Science (GSDS) is gratefully acknowledged. For a helpful discussion of the first preliminary version of this paper we like to thank the participants of the Psychoeconomics Workshop, Constance June 17^{th}, 2016. All remaining errors are ours.

†Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-5111, fax: -4450, email: Jana.Mareckova@uni-konstanz.de.

‡Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-2660, fax: -4450, email: winfried.pohlmeier@uni-konstanz.de.

Keywords: personality traits, machine learning

JEL classification: J24, J64, C38

PsycINFO classification: 20 3100, 21 3600

1 Introduction

The role of noncognitive skills in explaining individual differences in educational attainment and labor market success has been documented by labor economists and personality psychologists in numerous empirical and experimental studies. There is little doubt that beyond cognitive abilities, individual differences in noncognitive skills explain a large fraction of observed variation in individual labor market outcomes. Existing empirical evidence includes studies investigating the effects of noncognitive skills on earnings (Nyhus and Pons (2005) Mueller and Plug (2006)), job search (Uysal and Pohlmeier (2011), Viinikainen and Kokko (2012), Caliendo et al. (2014a)), occupational choice (John and Thomsen (2014)), self-employment (Caliendo et al. (2014b)) and educational attainment (Duckworth and Seligman (2005), Piatek and Pinger (2016)). Borghans et al. (2008) and Almlund et al. (2011) provide comprehensive overviews over the empirical findings from labor economics and personality psychology.

In this paper, we study the importance of noncognitive skills in explaining difference in the labor market performance of individuals by means of machine learning techniques. Empirical studies based on large-scale observational data usually contain a large number of measures of cognitive and noncognitive skills. Typically, some type of dimension reduction technique is applied in order to reduce the dimensionality problem and to obtain interpretable empirical results. Predominantly, this is done ex-ante via preprocessing the data by principal component analysis (PCA) and related factor modeling strategies or simply by index building. Moreover, some type of dimension reduction is implicitly accomplished by focusing on certain personality concepts (e.g. Big Five, locus of control) and disregarding covariates reflecting more closely alternative (complementary or competing) personality concepts. While conventional statistical approaches are mainly concerned with the within-sample explanatory power of noncognitive skills, the approach pursued in this paper is in the tradition of the machine or statistical learning literature to data analysis by focusing on the out-of-sample forecasting or classification qualities of noncognitive skills.

Our exploratory approach to the data may contribute to a better understanding of the importance of noncognitive skills for individual labor market performance. First, variable selection is strictly based on pseudo-out-of-sample performance, i.e. the selected empirical models have a higher external validity. Second, machine learning techniques can easily cope with the challenge

of selecting the most relevant measures from data sources with a whopping number of covariates on personality traits. Thus, they are not prone to within-sample over-fitting, a problem that is likely to occur if many highly correlated covariates are available. Third, machine learning techniques are particularly suitable for sparse modeling, i.e. they are able to select relevant variables and/or sets of variables among a large number of potential alternative specifications and provide final specifications which are easy to interpret.

Thus far practical experience with machine learning techniques in the context of psychometric or econometric studies is rather limited. It is the aim of this study to investigate to what extent and how machine learning techniques can contribute to a better understanding of the impact of noncognitive skills on the individual labor market performance. In the center of our approach is the group lasso (Yuan and Lin, 2006), as an L_1-norm penalization strategy ("lassoing") is able to select relevant regressors out of a large set of covariates and fixing less relevant covariates to zero. In addition to the simple lasso (Tibshirani (1996)), grouping leads to a further dimension reduction by selecting complete sets of variables for the model specification while suppressing the impact of less relevant groups. In particular, we show how lassoing and grouping can be used to construct context related indices of noncognitive skills. These indices incorporate the most relevant information from a larger set of factors of noncognitive skills where the weights are determined by the predictive relevance for a given outcome variable. In this sense, our approach can be seen as an alternative to the Bayesian exploratory factor approach by Conti et al. (2014) that also produces low-dimensional aggregates from high dimensional psychological measurements.

In our empirical study we construct such context related indices of noncognitive skills for individual wages and for unemployment. But our approach is not restricted to the empirical questions studied here, but has the potential to be a useful tool in similar settings where indices are constructed to reduce dimensionality of the estimation problem and where the focus of interest is external validity. In particular, our approach may have practical implications for pre-employment screening by providing valuable information to what extent a job candidate is likely to perform well in the job he is assigned for.

The fundamental identification problem of the impact of personality on labor market outcome in the presence of the panoply of personality traits and concepts and their corresponding

measurements is discussed at length by Almlund et al. (2011) and Borghans et al. (2011). All psychological measurements of personality are calibrated on measured behavior. In this context, the inability to disentangle behaviors that depend on a single trait or ability gives rise to a fundamental identification problem (Heckman and Kautz (2012)). By incorporating all covariates potentially capturing cognitive and noncognitive skills in several dimensions, our approach is mainly explorative in nature and circumvents any attempts of causal identification. Rather than focusing on one specific dimension of personality (e.g. locus control, self-esteem, Big Five factor et al.) the predictive power of a factor is analyzed in the presence of other competing factors and thereby reducing the omitted variable bias. An obvious example are interpersonal Big Five traits *Extraversion* and *Openness*, which are closely related to the concept of *Locus of Control*. Tables 13 and 14 in Appendix give the correlations between different factors of personality for males and females of the British Cohort Study. Based on the assessment by the mothers at the age ten of the child the correlations between *Agreeableness* and *Emotional Stability* and *Extraversion* and *Emotional Stability* exceed 0.5. Similarly, the correlation between *Self-esteem* and *Locus of Control* amount to .39 and .41 for males and females, respectively. Ignoring one of the factors would seriously overemphasize the role of the included personality traits.

Our argument can be extended to the impossibility to separate properly between cognitive and noncognitive skills. While Deke and Haimson (2006) report rather moderate correlations between maths skills and locus of control based on data of the US National Education Longitudinal Survey, Burks et al. (2009) argue that cognitive skills are related to economic preferences in different choices of domain. For instance, individuals with higher cognitive skills are more likely to be patient in the short and the long-run and to show a greater social awareness by being able to predict the behavior of others more accurately. A high conscientiousness of individual at the work place may simply reflect the ability to plan due to higher cognitive skills.

In the same spirit, Segal (2012) points out that scores obtained in cognitive skill tests may simply reflect the test-takers personality traits. If individuals differ not only in their cognitive abilities but also in their test taking motivation, then in the absence of performance based incentives higher test scores do not necessarily imply higher cognitive ability. Therefore, strong positive correlations between high scores on unincentivized cognitive skills tests and future labor market success observed in many empirical studies maybe due to intrinsic motivation. As

a consequence, failure to appropriately control for personality may overemphasize the impact of cognitive skills on labor market performance.

The paper is organized as follows. In Section 2 we introduce our group lasso approach to select measures of noncognitive skills and compare it with traditional factor analytic approaches. In Section 3 we describe a sample from the British Cohort Study (BCS) which serves as a basis for our empirical application. Section 4 contains the empirical findings for individual wages and for unemployment, while Section 5 concludes and provides an outlook on future research.

2 A Machine Learning Approach to Select Skill Factors

Machine learning techniques are widely used in settings when a researcher wants to learn about a model from a big amount of data. A variety of algorithms was designed to deal with large datasets yielding estimable parsimonious models. In this paper, we want to analyze which noncognitive skills play an important role in predicting labor market outcomes. Depending on the broadness of the survey, the number of questions about noncognitive skills can be arbitrarily large. Unless there is a prior idea which questions are measuring the same concept, a machine learning technique has to be applied to unveil the grouping of the questions. Therefore as a first step, we use clustering techniques to collect many survey questions into groups representing particular noncognitive skills. The questions in a group are then used to create a noncognitive skill index. In the second step, we plug the grouped variables into the model and let group lasso to decide which of these indices are important for predicting labor market outcomes and to estimate the index weights. The results of the group lasso are compared with alternative index constructions and alternative models.

2.1 L_q-Regularization

In order to understand how the group lasso helps to select and estimate index weights, the idea behind L_q regularization is introduced. Consider the following optimization problem

$$\min_{\beta} \|Y - X\beta\|_2^2 \qquad \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j|^q \leq s, \tag{1}$$

where $\|Y - X\beta\|_2^2$ represents sum of squared errors of a linear model and s is a chosen constant for $q \geq 0$. The advantage of the L_1-regularization (i.e. $q = 1$, lassoing) over the L_2-regularization ($q = 2$, ridging) is that under the given restriction there is a high probability that some of the parameters will be set to zero at the minimum. The lasso is therefore able to select relevant explanatory variables in one-step and to choose a sparse specification among a large set of possible specifications. The parameter q can take any positive value leading to different model specifications. However, it can be shown that only for $0 \leq q \leq 1$ there is a non-zero probability to select variables and to shrink the impact of less relevant variables to zero (e.g. Hastie et al. (2009)).

The primal problem (1) is equivalent to

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q. \tag{2}$$

The term $\sum_{j=1}^{p} |\beta_j|^q$ is called a penalty term and λ serves as the regularization (or penalty) parameter. For $0 \leq q \leq 1$, the choice of λ influences the sparsity of the solution (i.e. how many parameters are set to 0). The larger λ, the more sparse is the specification obtained. When $q = 1$, the (2) is called LASSO (Least Absolute Shrinkage and Selection Operator) optimization problem introduced by Tibshirani (1996) and part $\lambda \sum_{j=1}^{p} |\beta_j|$ is called a L_1-norm penalty.

2.2 Group Lasso

Consider now the case where covariates (e.g. facets in the Big Five framework or any set of items describing noncognitive skills) can be divided into J groups with k_j covariates in group j. Further, let N be the number of observations. The group lasso introduced by Yuan and Lin (2006) is used, when it is desirable to select a whole group of variables. This is achieved by solving:

$$\min_{\beta} \|Y - \sum_{j=1}^{J} X_j\beta_j \ - Z\delta\|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{k_j} \|\beta_j\|_2, \tag{3}$$

where Y is an $N \times 1$ vector of the dependent variable, X_j is an $N \times k_j$ matrix of covariates corresponding to the j-th group, β_j is a $k_j \times 1$ parameter vector for group j, Z are covariates which do not belong to any group, δ is a vector of parameters which are not regularized and λ is a regularization parameter. Since $q = 0.5$ in this case, the choice of λ determines how many

groups are selected.[1] The larger λ, the more elements in the penalty term $\sum_{j=1}^{J} \sqrt{k_j}\|\beta_j\|_2$ are forced to zero in order to minimize (3). In this case, the elements are Euclidean norms. A Euclidean norm is equal to zero, when all the components are zero. This means that the whole group is eliminated from the model and sparsity in groups is achieved. Therefore, the group lasso is a natural approach to select the best predicting group of variables related to a certain personality theory among a large set of competing personality theories.

In a case of highly correlated variables, Zou and Hastie (2005) recommend adding a ridge penalty to (2) with $q = 1$, since LASSO tends to choose only one variable from a group of highly correlated variables. The advantage of additional ridge penalty is its ability to select groups of correlated variables. As it is partially illustrated by Tables 13 and 14 in Appendix, the correlation between indeces can achieve high values. Therefore, we augment the optimization problem (3) as follows to control for this effect

$$\min_{\beta} \|Y - \sum_{j=1}^{J} X_j\beta_j \ - Z\delta\|_2^2 + \lambda \left(\alpha \sum_{j=1}^{J} \sqrt{k_j}\|\beta_j\|_2 + \frac{1-\alpha}{2}\|\beta\|_2^2 \right), \tag{4}$$

where $\alpha \in [0, 1]$ controls the balance between group and ridging penalty and can be chosen in advance.

Lassoing and group lassoing as briefly described are not restricted to the estimation of linear regression models. They can easily be extended to the estimation of nonlinear models. For our application to individual unemployment below, we use a penalized maximum logit approach where the the least squares part of objective function in (2) is replaced by the negative log likelihood of the logit model.

2.3 Index Construction

In the following, we use the properties of L_q-norm regularization for the construction and selection of skill indices. The selection of the indices is based on their ability to predict a given outcome variable. In the sense, the weighting scheme of the lasso-based indices is context related.

Let us now assume that the grouping of the questions/items is given. More details about

[1] One of the options how to select the optimal λ is a K-fold cross-validation. See e.g. Hastie et al. (2009) for more details.

how ex-ante grouping can be achieved by clustering as used in this paper can be found at the end of Subsection 3.2. For example, let $C_{i1}, C_{i2}, \ldots, C_{iK}$ be the responses of individual i on K different questions related to conscientiousness and $E_{i1}, E_{i2}, \ldots, E_{iL}$ be the responses of individual i on L different questions related to extraversion, then the indices take the following form:

$$IC_i^C = \sum_{j=1}^{K} \omega_j^C C_{ij}, \qquad \text{where } \sum_j |\omega_j^C| = 1,$$

$$IC_i^E = \sum_{j=1}^{L} \omega_j^E E_{ij}, \qquad \text{where } \sum_j |\omega_j^E| = 1.$$

The next question is what index weights $\omega_j^A, A = \{C, E\}$ the questions should get. In this paper, we implemented the following three weighting schemes:

1. Imposing equal weights: $\hat{\omega}_j^C = 1/K$, $\hat{\omega}_j^E = 1/L$.

2. Estimate the weights by the PCA: $\hat{\omega}_j^A = \pi_j^{PCA,A} / \sum_j |\pi_j^{PCA,A}|$ by taking loadings of the first component as $\pi_j^{PCA,A}$'s.

3. Estimate the weights using a group lasso penalty on the following model (for illustration we assume a linear model)

$$Y_i = \underbrace{\beta_1^C C_{i1} + \beta_2^C C_{i2} \ldots \beta_K^C C_{iK}}_{\text{total impact of C on Y}} + \underbrace{\beta_1^E E_{i1} + \beta_2^E E_{i2} \ldots \beta_L^E E_{iL}}_{\text{total impact of E on Y}} + \cdots + OtherControls_i'\delta + \varepsilon_i,$$

   yielding situation specific weights: $\hat{\omega}_j^A = \frac{\hat{\beta}_j^A}{\sum_j |\hat{\beta}_j^A|}$. Additionally, the group lasso also selects the most relevant groups (indices) for the model, e.g. the group lasso might assign zero coefficients to all conscientiousness questions, i.e. $\hat{\beta}_j^C = 0, \forall j$. In this case, all the $\hat{\omega}_j^C$ will get zero values and conscientiousness would be considered as irrelevant for predicting Y_i.

Note that only group lasso has a selection property, i.e. it chooses only the most relevant indices in the model. The models using equally weighted indices or PCA indices contain all the indices.

2.4 Model

As in the previous subsection, we assume that the grouping of the questions is known. We want to estimate the model of the following form:

$$Y_i = g(\alpha + Indices_i' \, \gamma + OtherControls_i' \, \delta) + \varepsilon_i, \tag{5}$$

where Y_i is the labor market outcome of interest, e.g log of an hourly gross wage or unemployment. The function $g(\cdot)$ is chosen according to the modeled dependent variable. In a case of continuous dependent variable, $g(\cdot)$ returns a linear regression model. In a case of a binary dependent variable, a conditional mean of a logit model is returned by $g(\cdot)$. Non-cognitive skills are captured in the indices. Models taking equally weighted indices and PCA indices will have all of them. Models based on group lasso results will have only the selected indices. Vector of other controls represents variables which are included to avoid an omitted variable bias. Vector $(\alpha, \gamma', \delta')'$ contains unknown parameters to be estimated. And ε_i is the error term.

To compare different index construction strategies and different models, the following model specifications are analyzed:

A. Restricted models:

   1. Model with control variables only ($\gamma = 0$),
   2. Model with noncognitive skills only ($\delta = 0$).

B. Model with noncognitive skills collected in equally weighted indices, with and without schooling (highest achieved qualification).

C. Model with noncognitive skills collected in PCA indices, with and without schooling (highest achieved qualification).

D. Model with noncognitive skills collected in group lasso indices, with and without schooling (highest achieved qualification).

The whole estimation procedure is summarized below:

1. Estimate the PCA weights with the noncognitive data from the training set (50% of observations).

2. Compute the equally weighted and PCA indices.

3. Estimate the group lasso model on the training set (the same 50% of observations as in point 1 to get reliable weights). Get the optimal lambda by a K-fold cross-validation.

4. Estimate the model (5) on the validation set (25% of the observations) to avoid potential overfitting by using the same data twice for the ω_j and γ estimates. Regarding the estimation of (5), linear models are estimated by OLS with White heteroscedastic errors and logit models are estimated by maximum likelihood.

5. Evaluate the out-of-sample performance on the test set (25% of the observations).

3 Data

Our empirical study is based on data from the British Cohort Study (BCS). The BCS is a wide-ranging data set containing a rich variety of variables of the study members and their families regarding medical, physical, educational, social and economic development as well as several measurements of noncognitive skills. The collection of data began in 1970. Babies born in a particular week in 1970 were tracked during their childhood, youth and adult life roughly every four to five years.

The BCS has been used in several empirical studies on the link between personality traits and labour market outcomes. Notable examples are Prevoo and ter Weel (2015), who analyze the effect of early conscientiousness on a variety of adult outcomes, Blanden et al. (2007) for role of noncognitive skills for intergenerational mobility and Uysal (2015) for the causal effects of education on earnings.

The longitudinal character of the data set enables us to analyze impact of early childhood environment and early cognitive and noncognitive skills on adult labor market outcomes. The adult outcomes were taken from the 2004, 2008 and 2012 waves, i.e. when the study members were 34, 38 and 42 years old. The data on cognitive and noncognitive skills were taken from the 1980 wave, when the study members were 10 years old.

3.1 Outcome Variables

We choose individual wages and unemployment as major outcome variables to analyze the predictive power of noncognitive skills on labor market outcomes. Regarding the unemployment variable, study members who reported that they are: full-time or part-time employed are coded as "Employed"[2]. Those who reported that they are currently unemployed and look for a job, receiving a Jobseeker's Allowance or were unemployed in the last 4 years are coded as "Unemployed" to capture a risk of unemployment as a dependent variable. Study members who reported they are in full-time education, on a government scheme for training, sick, disabled, looking after the family, wholly retired or do not fit in any category are discarded from the sample. Across all the waves, the level of unemployment rate is between 5-6%.

For the wage analysis, an hourly gross wage was computed from the available data. The variables reported are gross pay, period of the reported gross pay (minimal period is one week) and amount of hours worked in a week. A weekly gross pay is constructed based on the gross pay and the period of the reported gross pay. Dividing this result by hours worked in a week yields the hourly gross wage. The lower and upper 1% quantiles of the hourly gross wages were discarded from the sample to eliminate extreme outliers from the analysis. The lower quantile is around 2£ in the 2004 wave and around 4£ in the later waves. The upper quantile in all three waves is around 80-90£.

All model specifications are estimated separately for the samples of males and females to capture gender specific effects on the labor market. The numbers of observations available for the analysis after cleaning and matching the data from 1980 to the adult waves are captured in Table 5 of Appendix A.1. Especially in the case of wages, the later waves exhibit a higher dropout/non-response rate. Since each wave is estimated separately in the analysis and we are interested mainly in the predictive power, the dropout rate is not considered as an issue for the analysis.

3.2 Non-cognitive skills

There are several questionnaires in the BCS measuring behavior and personality of the study members. Answers to the questionnaires are traditionally collected into indices (scales) repre-

[2]Self-employed were discarded from the study.

senting particular noncognitive skills. We follow this approach. For a construction of the index, the answers have to be represented by points and have their corresponding index weights. The point representation of the answers is described below.

At the age of 10 (in the 1980 wave), the study members were asked to complete two questionnaires: the Self-Esteem Scale (LAWSEQ) introduced by Lawrence (1973, 1978) and the Locus of Control Scale (CARALOC) based on Gammage (1975). Self-esteem is a concept capturing a self-evaluation of one's own worthiness. In the BCS study, the Self-Esteem Scale comprises from 12 "Yes/No/I don't know" questions listed in Table 9 (excluding distractor questions). Mimicking the scheme from Lawrence (2006), all "Yes" answers get 1 point and all "No" answers get 2 points. The exception is Question 1, for which a "No" answer gets 1 point and a "Yes" gets 2 points. "I don't know" answers get 1.5 points. The higher score represents higher self-esteem.

The Locus of Control Scale is supposed to capture how much one believes that he is in control of his life (Rotter, 1966). Within this concept, people are then described as internalizers or externalizers. An internalizer thinks that he has a control over the outcomes in his life and that he can influence them by his own actions. An externalizer thinks that the outcomes in his life are determined by higher forces, luck and other external factors out of his control. The Locus of Control Scale in the BCS is covered by 16 "Yes/No/I don't know" questions listed in Table 10 (excluding distractors). Except of question 8, "Yes" gets 1 point, "I don't know" 1.5 point and "No" gets 2 points in a similar logic used for the self-esteem scale. In the case of question 8, "Yes" gets 2 points and "No" 1 point. Higher scores represent an internalizer.

In the 1980 wave, mothers of the study members were asked to answer a set of questions about the behavior of their 10 years old children. In total, there were 38 questions which are listed in Table 11. Mothers answered on a scale from 0 to 100, where 0 indicates "Certainly" and 100 indicates "Does not apply". Based on these questions, two well known instruments - the Rutter Behavior Scale (Rutter, 1967; Rutter et al., 1970) and the Conners Rating Scale (Conners, 1969) - are typically constructed in the literature to capture hyperactivity and anti-social behavior, see e.g. Conti et al. (2014) or Uysal (2015). Instead of applying the two measures mentioned above, we follow a strategy outlined in Butler et al. (1982) where they use a PCA to identify new more detailed scales. A similar approach is applied in Prevoo and ter Weel (2015) who use PCA and cluster analysis to identify new scales. In our study, we focus

on machine learning techniques. Therefore, we decided to apply cluster analysis.

The results of the cluster analysis are captured in Figure 1. We use a hierarchical clustering based on Ward's method, which starts with single clusters and in each step decides which pair to merge such that the within-cluster variance minimally increases (Ward Jr, 1963). The outcome of the hierarchical clustering is a dendogram which plots the whole path of the merging steps. With the help of Figure 1, the procedure can be illustrated as follows. The algorithm starts with 38 clusters, i.e. each question is a cluster. In the next step, the algorithm finds 2 clusters which are the most similar to each other and merges them. In this case, it merges first `clumsy` and `trips` yielding 37 clusters. Then the algorithm finds again the two most similar clusters and merges them until there is only 1 cluster with all the questions. The order of merging steps is represented in Figure 1 by the height level of the junctions. The lower the level, the earlier in the algorithm the clusters were merged.

The next step is how to choose the optimal level of clustering. For the hierarchical modeling, one of the recommended methods is to plot the number of clusters against a measure of similarity (Mooi and Sarstedt, 2011, Ch. 9). Based on this plot, one looks for a break (jump) which represents that the algorithm put together clusters which are too dissimilar and the merging should stop. In our case we plot the number of clusters against the increase in the within sum of squares after merging, sometimes called as a "merging cost" (Pragarauskaite and Dzemyda, 2012). When the merging cost is relatively too high, the merging should stop as the newly created cluster is too heterogeneous. The analysis of merging costs suggests 8 clusters since going from 8 to 7 clusters is relatively costly, see Figure 2. These 8 clusters are represented by red boxes in Figure 1 and got following labels:

- MC = Motor Control,
- HEC = Hand-Eye-Coordination,
- T = Behavioral Trauma,
- E = Extraversion,
- C = Conscientiousness,
- H = Hyperactivity,
- A = Agreeableness,
- ES = Emotional Stability.

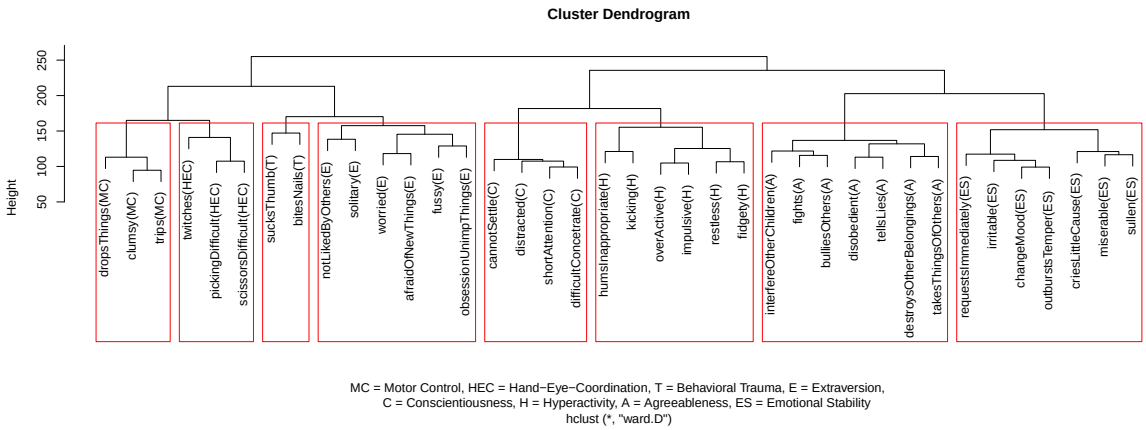
In total, we have 10 indices for the analysis. The 8 clusters from the mother-rated items, self-esteem index and the locus of control index. For a preliminary analysis, Tables 13 and 14 contain correlation between all 10 equally weighted indices and *Ability* measured by a Friendly Math Test. The results show that there is a slightly positive correlation between *Locus of Control* and *Ability* indicating a potential that there might be a channel between noncognitive and cognitive skills.

The 4 personality traits coming from Big Five (E, C, A and ES) are very similar to the clusters obtained in Prevoo and ter Weel (2015). Blanden et al. (2007) point out that the relevant variables in the BCS70 are rather close to the variables of Five Factor model.

3.3 Background Variables

As controls in the regressions, we use variables listed in Table 8. Social class of the family captures home environment, cognitive skills are captured in the ability variable based on a Friendly Math Test and effect of schooling is captured in the level of the highest achieved qualification. The last 3 variables are capturing the effects of being non-british, having children and being married.

Figure 1: Clusters of Mother-rated Items



4 Empirical Results

4.1 Wage Predictions

Table 1 contains the results of the wage equation augmented by cognitive and noncognitive skill factors for the 34 year old workers. Compared to conventional cross-sectional estimates of wage equations the coefficient estimates are not well determined, due to a rather small variation in the dependent variable at the beginning of the life-cycle, the cohort structure of the data and the comparatively small size of the samples ($N = 425, 414$).[3] Only coefficients on the ability variable (math score), home environment (social class), higher education and the children variable for the sample of the females are statistically significant different from zero at the conventional levels. The out-of-sample performance is measured by the out-of-sample R^2, which is based on the out-of-sample residuals obtained from test set. Note, that the out-of-sample R^2 is not bounded to the 0-1 range. In fact, it may take on negative values which would indicate that the model under consideration has lower predictive power than an intercept only model. This case, however, never occurred for any of our specifications.

A number of empirical studies show that noncognitive skills effect labor market performance directly, but also indirectly through higher education attainment. Therefore, we present a reduced form wage equation without the education variables as well as wage equation conditional on education such that the effects of cognitive and non-cognitive skills on wages have to be interpreted as conditional effects given education. The additional explanatory power of education variables turns out to be moderate within and out-of-sample. By including the education variables the predictive power increases only by 2 percent for the males and around 5 percent for the females, where the latter result for the females is driven by the contribution of the 2 highest education groups (HighQual4 and HighQual5) to the overall fit.

The prediction exercise reported in Table 1 was repeated for the samples of the 38 year old and the 42 year old, see Tables 15 and 16 in Appendix A.1. Most interestingly, the predictive power remains fairly constant for the wages earned at 38 and 42. This holds for the wage equation with and without the education variables (the only exception is a regression for male at 42). The high stability of the prediction power of the non-cognitive skill variables measured

[3] See Table 7 in the Appendix A.1.

at age 10 for wages measured at ages 34, 38 and 42 maybe seen as an additional support of the hypothesis that non-cognitive skills remain fairly constant over the life-cycle.

The predictive performance of the wage equation is also robust with respect to the way how noncognitive skills are measured. Constructing more sophisticated measures of noncognitive skills by means of principle components does not improve the estimates within and out-of-sample. Comparing the findings based on the equally weighted index (Table 1 and Tables 15 and 16 in Appendix A.1) with ones based on the PCA-based indices (Tables 17, 18 and 19 in Appendix A.1) shows that no measurement strategy dominates the other.

Table 2 contains the results for the wage equation of the 34 year old based on the group lasso. Here the indices for the non-cognitive skills are computed using a group lasso, so that the situation specific weights for the facets are taken into account. The optimal shrinkage parameter is evaluated by 5-fold cross validation and α is set to 0.5. The out-of sample prediction performance in terms of the R^2 improves slightly over the estimates based on the EW- and the PCA indices. But note that the admittedly small prediction improvement of the group lasso is obtained by incorporating fewer factors. For the group lasso specifications the number of explanatory variables are reduced to 15 for the males and 11 for the females compared to 21 parameters for the models using conventionally constructed indices. Most interesting, the group lasso selects the indices for *Hand-Eye Coordination* (X.HEC), *Behavioral Trauma* (X.TR), *Extraversion* (X.E) and *Locus of Control* (X.LC) for the wage equation of the males as valuable predictors, but excludes them for the wage equation of the females (see also Figures 3 and 4 in Appendix A.2). We interpret this as an indication that these non-cognitive factors are picked-up completely by the education variables for the case of the females. Non-cognitive skills for the wage equation of the females does not play a role even for the samples at later stages of the life-cycle.

Our results are confirmed for the wage predictions at ages 38 and 42 (see Tables 20 and 21 in Appendix A.1). Again the non-cognitive skill factors reveal some predictive power for the males but not for the females conditional on education and cognitive ability. In fact, the prediction quality increases over the life-cycle. For both, the samples for the males and the females at age 42, the predictive R^2 increases by around 1 percent for males and by around 3-5 percent for females compared to the values obtained for age 34.

Table 1: Wage Equation Estimates with EW Index, Sample 34Y

	Male	Female	Male	Female
Intercept	2.1868***	2.0932***	2.0522***	1.9867***
	(0.0935)	(0.0787)	(0.1091)	(0.0921)
MC	0.0201	0.0463	0.0205	0.0533*
	(0.0254)	(0.0256)	(0.0251)	(0.0239)
HEC	−0.0071	−0.0014	−0.0077	0.0104
	(0.0224)	(0.0265)	(0.0213)	(0.0253)
C	0.0317	0.0079	0.0310	0.0183
	(0.0315)	(0.0272)	(0.0330)	(0.0274)
H	0.0005	0.0033	−0.0078	0.0162
	(0.0262)	(0.0343)	(0.0271)	(0.0344)
A	−0.0321	−0.0153	−0.0156	−0.0159
	(0.0272)	(0.0307)	(0.0279)	(0.0303)
ES	−0.0061	−0.0439	−0.0085	−0.0330
	(0.0332)	(0.0287)	(0.0344)	(0.0267)
TR	0.0228	0.0282	0.0332	0.0161
	(0.0224)	(0.0233)	(0.0230)	(0.0223)
E	−0.0078	0.0000	−0.0180	−0.0090
	(0.0241)	(0.0246)	(0.0234)	(0.0239)
SE	0.0398	0.0357	0.0434	0.0242
	(0.0265)	(0.0265)	(0.0264)	(0.0264)
LC	0.0776**	0.0607*	0.0619*	0.0485
	(0.0267)	(0.0294)	(0.0269)	(0.0286)
Ability10Y	0.1097***	0.1643***	0.0872**	0.1244***
	(0.0325)	(0.0325)	(0.0333)	(0.0329)
SocialClass10Y	0.0674***	0.0688***	0.0566**	0.0430*
	(0.0194)	(0.0199)	(0.0203)	(0.0198)
NonBritish	0.1446	0.2404*	0.1089	0.1240
	(0.1224)	(0.0986)	(0.1288)	(0.1199)
Children	0.0218	−0.1097***	0.0275	−0.1003***
	(0.0278)	(0.0233)	(0.0270)	(0.0236)
Married	0.1055	0.0561	0.0975	0.0544
	(0.0684)	(0.0518)	(0.0689)	(0.0514)
HighQual1			0.0170	0.0017
			(0.1122)	(0.1126)
HighQual2			0.1185	0.0821
			(0.0933)	(0.0674)
HighQual3			0.2259*	0.1048
			(0.0959)	(0.0862)
HighQual4			0.2420*	0.3565***
			(0.0948)	(0.0735)
HighQual5			0.3094*	0.4258***
			(0.1264)	(0.1195)
R^2 (In sample)	0.2064	0.2847	0.2316	0.3518
MSE (In sample)	0.1981	0.1878	0.1918	0.1701
R^2 (Out of sample)	0.1312	0.1459	0.1536	0.2001
MSE (Out of sample)	0.1957	0.2010	0.1907	0.1882

Least squares estimates of the wage equation for the sample of the 34 year old, augmented by skill factors based on equally weighted indices (EW). Dep. var.: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 2: Group Lasso Wage Equation Estimates: Sample 34Y

	Male	Female
Intercept	2.0611***	1.9545***
	(0.1043)	(0.0829)
Ability10Y	0.1071***	0.1528***
	(0.0288)	(0.0301)
SocialClass10Y	0.0585**	0.0462*
	(0.0201)	(0.0186)
NonBritish	0.0981	0.0947
	(0.1256)	(0.1314)
Children	0.0252	−0.1027***
	(0.0267)	(0.0233)
Married	0.1037	0.0677
	(0.0679)	(0.0509)
HighQual1	−0.0064	0.0105
	(0.1047)	(0.1097)
HighQual2	0.1025	0.0938
	(0.0903)	(0.0613)
HighQual3	0.2068*	0.1103
	(0.0932)	(0.0788)
HighQual4	0.2180*	0.3776***
	(0.0950)	(0.0686)
HighQual5	0.2822*	0.4511***
	(0.1254)	(0.1151)
X_HEC	−0.0065	
	(0.0214)	
X_TR	0.0310	
	(0.0283)	
X_E	0.0477	
	(0.0640)	
X_LC	0.1618**	
	(0.0540)	
R^2 (In sample)	0.2186	0.3253
MSE (In sample)	0.1951	0.1771
R^2 (Out of sample)	0.1554	0.1975
MSE (Out of sample)	0.1903	0.1888

Group Lasso estimates of the wage equation for the sample of the 34 year old, augmented by skill factors based on group lasso weighted indices. Dependent variable: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, post-lasso standard errors.

4.2 Unemployment Classifications

Due to the rather low unemployment rates of around 5-6 percent in our different samples, the precision of the coefficient estimates is rather low. Moreover, the uneven distribution of the binary outcome variable leads to a high prediction accuracy (share of correct predictions) within- and out-of-sample despite the low precision of the coefficient estimates. In the following, we therefore discuss out-of-sample predictions in terms of sensitivity and specificity. All results reported below are based on cut-off thresholds chosen to maximize Youden's J-statistic. Geometrically, the maximum Youden's J-statistic maximizes the vertical distance of the Receiver Operating Characteristic (ROC) curve from a 45 degree or chance line. The sensitivity value is of the most interest, if the prediction exercise is to select a model that is most successful in predicting individual unemployment within the group of the unemployed (share of true positives among the positives), e.g. in order to use the model to select candidates for an intervention.

All models are estimated by maximum likelihood logit using the equally weighted index, the PCA index and the group lasso index. To get the group lasso index a penalized maximum likelihood method was implemented with $\alpha = 0.5$ and λ chosen by a 30-fold cross validation. Table 3 reports on the results when only the noncognitive skill factors are used. The predictive power of this equation is rather high compared to the augmented specifications which include in addition conventional regressors such as socio-economic controls (married, children, social class), and cognitive ability educational attainment (see Tables 22 and 23 in Appendix A.1). Even without conventional economic predictors the classification is rather high. The sensitivity improves somewhat (but not in all cases) when additional covariates are included. Interestingly, contrary to the wage predictions, the quality of the unemployment classifications (within- and out-of-sample) turn out to be sensitive to the construction of the index and varies across gender and across the different samples.[4]

Table 4 contains the results based on the group lasso estimates for the samples of the 34 year old and the 42 year old. *Agreeableness* turns out to be the most important predictor at the age of 34 for both males and females. Otherwise, group lasso chooses different predictors for each gender. The out-of-sample sensitivity is .55 and .56 for the 34 year old males and females but decreases for the classifications at age 42 with values of .39 and .28.

[4]Estimates for the sample 38Y and 42Y can be obtained from the authors on request.

Table 3: Unemployment Equation Estimates, Skill Factors only, Sample 34Y

	EWI - M	EWI - F	PCA - M	PCA - F
Intercept	-2.5844^{***}	-3.2238^{***}	-2.5871^{***}	-3.2095^{***}
	(0.1858)	(0.2528)	(0.1861)	(0.2505)
MC	0.0514	−0.0285	0.0522	−0.0163
	(0.1887)	(0.2457)	(0.1953)	(0.2544)
HEC	−0.2291	−0.0284	−0.2259	−0.0796
	(0.2155)	(0.2700)	(0.2172)	(0.2860)
C	−0.2084	−0.1362	−0.2264	−0.1714
	(0.2164)	(0.2839)	(0.2162)	(0.2897)
H	0.1748	0.6266	0.1320	0.5954
	(0.2204)	(0.3683)	(0.2201)	(0.3703)
A	0.1978	0.6424^{**}	0.1639	0.6125^{*}
	(0.1701)	(0.2463)	(0.1679)	(0.2448)
ES	0.3160	−0.4769	0.3318	−0.4100
	(0.2112)	(0.3834)	(0.2112)	(0.3812)
TR	0.1544	−0.0930	0.1787	−0.0705
	(0.1807)	(0.2380)	(0.1831)	(0.2380)
E	−0.2959	0.7399^{**}	−0.2783	0.7259^{**}
	(0.1901)	(0.2603)	(0.1911)	(0.2630)
SE	0.0334	0.0370	0.0927	0.0117
	(0.1882)	(0.2483)	(0.1942)	(0.2482)
LC	-0.4228^{*}	0.0226	-0.4763^{*}	−0.0003
	(0.1826)	(0.2250)	(0.1860)	(0.2297)
Accuracy (In sample)	0.7559	0.8919	0.6543	0.7623
Specificity (In sample)	0.7702	0.9196	0.6532	0.7773
Sensitivity (In sample)	0.5952	0.3333	0.6667	0.4583
Accuracy (Out of sample)	0.7333	0.8406	0.5961	0.6752
Specificity (Out of sample)	0.7633	0.8781	0.5949	0.6921
Sensitivity (Out of sample)	0.3902	0.0833	0.6098	0.3333
Cutoff	0.1014	0.0951	0.0795	0.0539

ML logit estimates of the unemployment equation for males and females based on skill factors only. Dependent variable: unemployed = 1, employed = 0,
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Unemployment Equation Estimates, Lassoed Factors, Sample 34Y and 42Y

	34Y		42Y	
	Male	Female	Male	Female
Intercept	-1.4378^{*}	-1.6454^{*}	-1.2484	-0.5356
	(0.6137)	(0.7424)	(0.8033)	(0.7770)
Ability10Y	-0.1601	0.3450	-0.0342	-0.3247
	(0.1849)	(0.2829)	(0.2454)	(0.2522)
SocialClass10Y	0.0075	0.0436	-0.2252	-0.2386
	(0.1433)	(0.1918)	(0.1874)	(0.1921)
Children	-0.1915	-0.1883	-0.0920	-0.3506
	(0.2061)	(0.2474)	(0.2105)	(0.2514)
Married	-0.9535^{*}	-1.3833^{**}	-1.1824^{*}	-0.0235
	(0.4074)	(0.4590)	(0.4866)	(0.4765)
HighQual1	-0.2551	-0.7152	0.2060	-0.9612
	(0.4929)	(0.5853)	(0.6798)	(0.5679)
HighQual2	-0.2210	-0.9252	-0.0527	-1.6569^{*}
	(0.5207)	(0.6589)	(0.7390)	(0.6791)
X_A	-0.0416	0.4453		
	(0.3452)	(0.3872)		
X_TR	0.0182			
	(0.2272)			
X_E	-0.4589			0.4749
	(0.3888)			(0.5729)
X_SE	0.9913			
	(0.5361)			
X_LC	0.9199			
	(0.7105)			
X_C		-1.3142^{*}		
		(0.5490)		
X_ES		0.2209		
		(0.4729)		
Accuracy (In sample)	0.6172	0.5442	0.7248	0.8365
Specificity (In sample)	0.6106	0.5299	0.7368	0.8522
Sensitivity (In sample)	0.6905	0.8333	0.5600	0.5909
Accuracy (Out of sample)	0.5608	0.5669	0.7253	0.7699
Specificity (Out of sample)	0.5458	0.5620	0.7478	0.7994
Sensitivity (Out of sample)	0.7317	0.6667	0.3913	0.2857
Cutoff	0.0677	0.0295	0.0832	0.0868

Group Lasso logit estimates of the unemployment equation for males and females, lassoed skill factors, sample Y34. Dependent variable: unemployed = 1, employed = 0, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, post-lasso standard errors.

5 Conclusion

This paper takes a closer look on the predictive power of non-cognitive skills for labor market outcomes by means of machine learning techniques. This existence of considerable predictive power of noncognitive skills has been claimed in many empirical studies which led to the claim that public policies should pay more attention to programs to enhance these skills (Heckman and Kautz (2012)). This paper is trying to provide further evidence on this hypothesis based on real out-of-sample forecast.

The group lasso approach proposed here accounts for several desirable features. It guarantees a parsimonious selection of factors and avoids over-fitting in the presence of data sets containing a large number of skill factors. Moreover lassoing helps to construct context related skill indices and select only those indices which are most relevant in predicting a certain outcome variable. We show that our group lasso approach cannot generally be outperformed by standard approaches using equally weighted indices or PCA based indices.

Our empirical findings are based on data from the BCS containing life cycle information of individuals born in 1970. Admittedly, expecting non-cognitive skill factors which were surveyed at the age of 10 to be predictors of individual labor market outcomes 24, 28 and 32 years later is very ambitious. Nevertheless, these factors seem to have some explanatory power for wages and unemployment many years later.

The BCS is a very rich, but also very specific data source in terms of its design and the definition of non-cognitive skill factors. Therefore, in future work our findings should be confronted with results based on different data sources with alternative definitions of the skill factors and different forecasting horizons. Moreover, the choice of the shrinkage parameter of the group lasso is rather conventionally chosen by means of 5-fold or 30-fold cross-validation. Here we see room for further improvement, since cross-validation is known to yield rather unstable estimates of the optimal shrinkage parameter. Future work should consider stability selection strategy based on subsampling to create more stable solutions as proposed by Meinshausen and Bühlmann (2010).

References

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011): "Chapter 1 - Personality Psychology and Economics," in *Handbook of The Economics of Education*, ed. by S. M. Eric A. Hanushek and L. Woessmann, Elsevier, vol. 4 of *Handbook of the Economics of Education*, 1 – 181.

Blanden, J., P. Gregg, and L. Macmillan (2007): "Accounting for Intergenerational Income Persistence: Non-Cognitive Skills, Ability and Education," *Economic Journal*, 117, C43 – C60.

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008): "The Economics and Psychology of Personality Traits," *Journal Human Resources*, 43, 972–1059.

Borghans, L., B. H. Golsteyn, J. Heckman, and J. E. Humphries (2011): "Identification problems in personality psychology," *Personality and Individual Differences*, 51, 315 – 320, special Issue on Personality and Economics.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009): "Cognitive skills affect economic preferences, strategic behavior, and job attachment," *Proceedings of the National Academy of Sciences*, 106, 7745–7750.

Butler, N., M. Haslum, W. Barker, and A. Morris (1982): "Child health and education study: First report to the Department of Education and Science on the 10-year follow-up," *Bristol: Department of Child Health, University of Bristol.*

Bynner, J., N. Butler, E. Ferri, P. Shepherd, and K. Smith (2002): "The Design and Conduct of the 1999-2000 Surveys of the National Child Development Study and the 1970 British Cohort Study," Tech. rep., Centre for Longitudinal Studies. Institute of Education University of London.

Caliendo, M., D. A. Cobb-Clark, and A. Uhlendorff (2014a): "Locus of Control and Job Search Strategies," *Review of Economics and Statistics*, 97, 88–103.

Caliendo, M., F. Fossen, and A. S. Kritikos (2014b): "Personality characteristics and the decisions to become and stay self-employed," *Small Business Economics*, 42, 787–814.

Conners, C. K. (1969): "A teacher rating scale for use in drug studies with children," *Amer-*

ican journal of Psychiatry, 126, 884–888.

Conti, G., S. Frühwirth-Schnatter, J. J. Heckman, and R. Piatek (2014): "Bayesian exploratory factor analysis," *Journal of Econometrics*, 183, 31 – 57.

Deke, J. and J. Haimson (2006): "Valuing Student Competencies: Which Ones Predict Postsecondary Educational Attainment and Earnings, and for Whom?" Tech. rep., Mathematica Policy Research Inc.

Duckworth, A. L. and M. E. Seligman (2005): "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents," *Psychological Science*, 16, 939–944.

Gammage, P. (1975): "Socialisation, schooling and locus of control," Ph.D. thesis, Bristol University, Bristol, England.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer New York.

Heckman, J. J. and T. Kautz (2012): "Hard evidence on soft skills," *Labour Economics*, 19, 451 – 464.

John, K. and S. L. Thomsen (2014): "Heterogeneous returns to personality: the role of occupational choice," *Empirical Economics*, 47, 553–592.

Lawrence, D. (1973): *Improved Reading through Counselling*, Ward Lock Educational: London.

——— (1978): *Counselling students with reading difficulties: a handbook for tutors and organizers*, Good Reading: London.

——— (2006): *Enhancing Self-esteem in the Classroom*, SAGE Publications.

Meinshausen, N. and P. Bühlmann (2010): "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473.

Mooi, E. and M. Sarstedt (2011): *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*, Springer-Verlag Berlin Heidelberg.

Mueller, G. and E. Plug (2006): "Estimating the Effect of Personality on Male-Female Earnings," *Industrial and Labor Relations Review*, 60, 3–22.

Nyhus, E. and E. Pons (2005): "The Effects of Personality on Earnings," *Journal of Economic Psychology*, 26, 363–384.

Piatek, R. and P. Pinger (2016): "Maintaining (Locus of) Control? Data Combination for the Identification and Inference of Factor Structure Models," *Journal of Applied Econometrics*, 31, 734–755.

Pragarauskaite, J. and G. Dzemyda (2012): "Visual decisions in the analysis of customers online shopping behavior," *Nonlinear Analysis: Modelling and Control*, 17, 355–368.

Prevoo, T. and B. ter Weel (2015): "The importance of early conscientiousness for socio-economic outcomes: evidence from the British Cohort Study," *Oxford Economic Papers*, 67, 918–948.

Rotter, J. (1966): "Generalized Expectancies for Internal versus External Control of Reinforcement," *Psychological Monographs*, 80(609), 1–28.

Rutter, M. (1967): "A children's behaviour questionnaire for completion by teachers: preliminary findings," *Journal of child Psychology and Psychiatry*, 8, 1–11.

Rutter, M., J. Tizard, and K. Whitmore (1970): *Education, health and behaviour*, Longman.

Segal, C. (2012): "Working When No One Is Watching: Motivation, Test Scores, and Economic Success," *Management Science*, 58, 1438–1457.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Uysal, S. D. (2015): "Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling," *Journal of Applied Econometrics*, 30, 763–786.

Uysal, S. D. and W. Pohlmeier (2011): "Unemployment duration and personality," *Journal of Economic Psychology*, 32, 980–992.

Viinikainen, J. and K. Kokko (2012): "Personality traits and unemployment: Evidence from longitudinal data," *Journal of Economic Psychology*, 33, 1204 – 1222.

Ward Jr, J. H. (1963): "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, 58, 236–244.

Yuan, M. and Y. Lin (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68, 49–67.

Zou, H. and T. Hastie (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 301–320.

A Appendix

A.1 Tables

A.1.1 Variable Definitions Summary Statistics

Table 5: Number of Observations

	Unemployment		Wage	
	Male	Female	Male	Female
34Y	2044	2034	1698	1654
38Y	1514	1524	1153	1298
42Y	1461	1464	853	836

Table 6: Number of Observations, Unemployment Equation

		Training Set	Validation Set	Test Set	Total
34Y	Male	1022	512	510	2044
	Female	1017	509	508	2034
	Pooled	2039	1020	1019	4078
38Y	Male	756	380	378	1514
	Female	761	382	381	1524
	Pooled	1519	760	759	3038
42Y	Male	730	367	364	1461
	Female	732	367	365	1464
	Pooled	1462	732	731	2925

Table 7: Number of Observations, Wage Equation

		Training Set	Validation Set	Test Set	Total
34Y	Male	849	425	424	1698
	Female	827	414	413	1654
38Y	Male	576	289	288	1153
	Female	649	325	324	1298
42Y	Male	426	214	213	853
	Female	418	209	209	836

Table 8: Description Control Variables

Variable	Description
SocialClass10Y	social class of the family at the age of 10, the better occupation of the parents taken. 6 categories: = 6 if professional, = 5 if managerial-technical, = 4 if skilled non-manual, = 3 if skilled manual, = 2 if partly skilled, = 1 if unskilled
Ability10Y	standardized score of a "Friendly Math Test" taken at the age of 10
HighQual•	• = Level of the Highest Achieved Qualification (see Table 12 for the definitions)
NonBritish	= 1 if non-British and non-Irish background, = 0 if British background
Children	number of children in the household
Married	= 1 if married/cohabiting, = 0 if not married

Table 9: Non-cognitive Skills Measures: Self-Esteem

Self-Esteem Scale Questions	
1	Do you think that your parents like to hear about your ideas?
2	Do you often feel lonely at school?
3	Do other children often break friends or fall out with you?
4	Do you think that other children often say nasty things about you?
5	When you have to say things in front of teachers, do you usually feel shy?
6	Do you often feel sad because you have nobody to play with at school?
7	Are there lots of things about yourself you would like to change?
8	When you have to say things in front of other children, do you usually feel foolish?
9	When you want to tell a teacher something, do you usually feel foolish?
10	Do you often have to find new friends because your old friends are playing with somebody else?
11	Do you usually feel foolish when you talk to your parents?
12	Do other people often think that you tell lies?

Answers coded as: "Yes" = 1 point, "No" = 2 points and "I don't know" = 1.5 point, except Question 1 for which the points for "Yes" and "No" are switched. Higher score indicates higher self-esteem.

Table 10: Non-cognitive Skills Measures: Locus of Control

Locus of Control Scale Questions	
1	Do you feel that most of the time it's not worth trying hard because things never turn out right anyway?
2	Do you feel that wishing can make good things happen?
3	Are people good to you no matter how you act towards them?
4	Do you usually feel that it's almost useless to try in school because most children are cleverer than you?
5	Is a high mark just a matter of "luck" for you?
6	Are tests lust a lot of guess work for you?
7	Are you often blamed for things which just aren't your fault?
8	Are you the kind of person who believes that planning ahead makes things turn out better?
9	When bad things happen to you, is it usually someone else's fault?
10	When someone is very angry with you, is it impossible to make him your friend again?
11	When nice things happen to you is it only good luck?
12	Do you feel sad when it's time to leave school each day?
13	When you get into an argument is it usually the other person's fault?
14	Are you surprised when your teacher says you've done well?
15	Do you usually get low marks, even when you study hard?
16	Do you think studying for tests is a waste of time?

Answers coded as: "Yes" = 1 point, "No" = 2 points and "I don't know" = 1.5 point, except Question 8 for which the points for "Yes" and "No" are switched. Higher score indicates an internalizer.

Table 11: Non-cognitive Skills Measures: Mother-rated Items (after clustering)

Motor Control (MC) Items	
1	Child drops things which are being carried
2	Child is noticeably clumsy
3	Child trips or falls easily into objects or other people
Hand Eye Coordination (HEC) Items	
4	Child has twitches mannerisms or tics of the face and body
5	Child has difficulty picking up small objects
6	Child has difficulty in using scissors
Conscientiousness (C) Items	
7	Child cannot settle to anything for more than a few moments
8	Child is inattentive, easily distracted
9	Child fails to finish things he/she starts, short attention span
10	Child has difficulty concentrating on any particular task though may return to it frequently
Hyperactivity (H) Items	
11	Child is squirmy or fidgety
12	Child is very restless, i.e. running often, jumping up and down
13	Child shows restless or over-active behavior
14	Child hums or makes other odd noises at inappropriate times
15	Child is given to rhythmic tapping or kicking
Agreeableness (A) Items	
16	Child frequently fights with other children
17	Child bullies other children
18	Child interferes with the activity of others
19	Child is often disobedient
20	Child often tells lies
21	Child often destroys own or others' belongings
22	Child sometimes takes things belonging to others
Emotional Stability (ES) Items	
23	Child cries for little cause
24	Child often appears miserable, unhappy
25	Child is sullen or sulky
26	Child is irritable
27	Child changes mood quickly and drastically
28	Child displays outbursts of temper, explosive or unpredictable behavior
29	Child's requests must be met immediately, easily frustrated
30	Child is impulsive, excitable
Behavioral Trauma (TR) Items	
31	Child frequently sucks thumb or fingers
32	Child frequently bites nails or fingers
Extraversion (E) Items	
33	Child is not much liked by other children
34	Child tends to do things on his own
35	Child is fussy or over particular
36	Child is often worried
37	Child tends to be fearful or afraid of new things or new situations
38	Child becomes obsessional about unimportant things

Answers coded on a scale from 0 to 100 where 0 = "certainly" and 100 = "does not apply".

Table 12: Mapping of Educational Qualifications into Levels

Merged Level	Level	General (Academic)	Vocationally-related (Applied)	Occupational (Vocational)
2	5	Higher Degree		NVQ level 5 PGCE
2	4	Degree HE Diploma	BTEC Higher Certificate/Diploma HNC/HND	NVQ level 4 Professional degree level qualifications Nursing/paramedic Other teaching training qualification City & Guilds Part 4/Career Ext/Full Tech RSA Higher Diploma
1	3	A level AS levels Scottish Highers Scottish Cert of Sixth-Year Studies	Advanced GNVQ BTEC National Diploma ONC/OND	NVQ level 3 City & Guild Part 3/Final/Advanced Craft RSA Advanced Diploma Pitmans level 3
1	2	GCSE grade A*-C O levels grade A-C O levels grade D-E CSE grade 1 Scottish standard grades 1-3 Scottish lower or ordinary grades	Intermediate GNVQ BTEC First Certificate BTEC First Diploma Other BTEC	NVQ level 2 Apprenticeships City & Guilds Part 2/Craft/Intermediate City & Guilds Part 1/Other RSA First Diploma Pitmans level 2
0	1	GCSE grade D-G CSEs grades 2-5 Scottish standard grades 4-5 Other Scottish school qualification	Foundation GNVQ Other GNVQ	NVQ level 1 Other NVQ Units towards NVQ RSA Cert/Other Pitmans level 1 Other vocational qualifications HGV
0	0	None	None	None

Source: Bynner et al. (2002)

Table 13: Correlation Matrix for Equally Weighted Indices for Male - 34Y (Wage Data)

	MC	HEC	C	H	A	ES	TR	E	SE	LC
HEC	0.45									
C	0.35	0.29								
H	-0.35	-0.24	-0.56							
A	0.37	0.37	0.48	-0.49						
ES	0.34	0.29	0.41	-0.53	0.57					
TR	-0.14	-0.17	-0.14	0.16	-0.20	-0.16				
E	0.29	0.30	0.26	-0.33	0.30	0.53	-0.17			
SE	-0.10	-0.01	-0.21	0.11	-0.10	-0.13	0.05	-0.09		
LC	-0.05	-0.04	-0.25	0.11	-0.14	-0.10	0.07	-0.06	0.39	
Abil	-0.06	-0.07	-0.33	0.12	-0.18	-0.14	0.05	-0.06	0.25	0.45

Table 14: Correlation Matrix for Equally Weighted Indices for Female - 34Y (Wage Data)

	MC	HEC	C	H	A	ES	TR	E	SE	LC
HEC	0.44									
C	0.42	0.33								
H	-0.42	-0.32	-0.57							
A	0.42	0.50	0.50	-0.51						
ES	0.35	0.31	0.45	-0.56	0.60					
TR	-0.16	-0.19	-0.18	0.21	-0.23	-0.20				
E	0.26	0.31	0.33	-0.39	0.40	0.53	-0.17			
SE	-0.09	-0.05	-0.13	0.10	-0.08	-0.11	0.08	-0.08		
LC	-0.05	-0.04	-0.12	0.09	-0.09	-0.07	0.08	-0.03	0.41	
Abil	-0.06	-0.05	-0.25	0.14	-0.16	-0.10	0.05	-0.07	0.12	0.26

A.1.2 Further Estimation Results

Table 15: Wage Equation Estimates with EW Index, Sample 38Y

	Male	Female	Male	Female
Intercept	2.2919***	2.2638***	2.2884***	2.1161***
	(0.0729)	(0.0784)	(0.1029)	(0.1298)
MC	−0.0093	0.0056	−0.0128	0.0006
	(0.0260)	(0.0305)	(0.0269)	(0.0290)
HEC	0.0056	0.0202	0.0046	0.0148
	(0.0243)	(0.0314)	(0.0252)	(0.0293)
C	−0.1019**	−0.0331	−0.0943**	−0.0286
	(0.0351)	(0.0352)	(0.0352)	(0.0351)
H	−0.0142	−0.0221	−0.0144	−0.0194
	(0.0319)	(0.0371)	(0.0326)	(0.0391)
A	0.0314	−0.0249	0.0366	−0.0216
	(0.0304)	(0.0372)	(0.0307)	(0.0322)
ES	−0.0171	0.0144	−0.0197	0.0317
	(0.0339)	(0.0352)	(0.0347)	(0.0341)
TR	0.0308	0.0212	0.0325	0.0234
	(0.0258)	(0.0273)	(0.0261)	(0.0270)
E	0.0133	0.0027	0.0133	−0.0074
	(0.0249)	(0.0366)	(0.0254)	(0.0350)
SE	−0.0454	−0.0003	−0.0394	0.0037
	(0.0272)	(0.0283)	(0.0273)	(0.0288)
LC	0.0376	0.0778**	0.0296	0.0516
	(0.0300)	(0.0283)	(0.0316)	(0.0272)
Ability10Y	0.1084***	0.1334***	0.1010**	0.0878*
	(0.0318)	(0.0387)	(0.0342)	(0.0405)
SocialClass10Y	0.0749***	0.0491*	0.0695***	0.0415*
	(0.0193)	(0.0198)	(0.0204)	(0.0201)
NonBritish	0.0861	0.1935	0.0268	0.1706
	(0.2822)	(0.1720)	(0.2898)	(0.1911)
Children	0.0330	−0.0373	0.0307	−0.0216
	(0.0316)	(0.0272)	(0.0309)	(0.0259)
Married	0.1743*	0.0213	0.1790**	0.0250
	(0.0708)	(0.0521)	(0.0682)	(0.0501)
HighQual1			0.1506	−0.0203
			(0.2120)	(0.1443)
HighQual2			−0.0743	0.0121
			(0.0908)	(0.1210)
HighQual3			0.0527	−0.0414
			(0.1126)	(0.1406)
HighQual4			0.0799	0.3137*
			(0.0964)	(0.1217)
HighQual5			0.0049	0.3681*
			(0.1167)	(0.1680)
R^2 (In sample)	0.2592	0.1936	0.2814	0.2913
MSE (In sample)	0.1535	0.1945	0.1489	0.1709
R^2 (Out of sample)	0.1253	0.1827	0.1341	0.2370
MSE (Out of sample)	0.1650	0.1896	0.1633	0.1770

Least squares estimates of the wage equation for the sample of the 38 year old, augmented by skill factors based on equally weighted indices (EW). Dependent var.: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 16: Wage Equation Estimates with EW Index, Sample 42Y

	Male	Female	Male	Female
Intercept	2.5015***	2.5670***	2.6517***	2.2017***
	(0.1102)	(0.1156)	(0.2107)	(0.1676)
MC	−0.0025	0.0179	−0.0330	0.0082
	(0.0399)	(0.0400)	(0.0410)	(0.0419)
HEC	−0.0744	0.0622	−0.0537	0.0699
	(0.0494)	(0.0436)	(0.0549)	(0.0421)
C	−0.0722	0.0372	−0.0322	0.0587
	(0.0502)	(0.0433)	(0.0446)	(0.0447)
H	−0.0009	0.0298	−0.0087	0.0399
	(0.0509)	(0.0481)	(0.0459)	(0.0474)
A	−0.0472	0.0257	−0.0369	0.0333
	(0.0513)	(0.0407)	(0.0442)	(0.0403)
ES	−0.0043	−0.0424	0.0065	−0.0258
	(0.0525)	(0.0425)	(0.0504)	(0.0385)
TR	−0.0341	0.0678	−0.0172	0.0779*
	(0.0343)	(0.0365)	(0.0321)	(0.0334)
E	0.0471	−0.0046	0.0223	−0.0195
	(0.0546)	(0.0291)	(0.0527)	(0.0261)
SE	0.0110	0.0649	0.0435	0.0531
	(0.0328)	(0.0398)	(0.0316)	(0.0374)
LC	0.0544	0.0791*	0.0387	0.0463
	(0.0369)	(0.0386)	(0.0367)	(0.0363)
Ability10Y	0.0671	0.1636**	0.0327	0.1229**
	(0.0350)	(0.0508)	(0.0334)	(0.0456)
SocialClass10Y	0.0706*	0.0180	0.0418	−0.0264
	(0.0280)	(0.0308)	(0.0290)	(0.0294)
NonBritish	0.0875	0.6748***	0.0145	0.4404***
	(0.1961)	(0.1562)	(0.2275)	(0.1304)
Children	0.0216	−0.0713*	0.0181	−0.0572
	(0.0315)	(0.0340)	(0.0284)	(0.0326)
Married	0.1790*	0.0221	0.1431*	0.0573
	(0.0756)	(0.0721)	(0.0694)	(0.0690)
HighQual1			−0.1358	0.1942
			(0.2651)	(0.1655)
HighQual2			−0.3006	0.3221*
			(0.1876)	(0.1530)
HighQual3			−0.0291	0.2798
			(0.1920)	(0.1611)
HighQual4			0.1116	0.6808***
			(0.1922)	(0.1486)
HighQual5			0.2835	0.7376***
			(0.2032)	(0.1722)
R^2 (In sample)	0.2298	0.2653	0.3372	0.4195
MSE (In sample)	0.2010	0.1994	0.1730	0.1575
R^2 (Out of sample)	0.1830	0.0375	0.2063	0.1687
MSE (Out of sample)	0.1883	0.2690	0.1829	0.2323

Least squares estimates of the wage equation for the sample of the 42 year old, augmented by skill factors based on equally weighted indices (EW). Dependent var.: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 17: Wage Equation Estimates with PCA Index, Sample 34Y

	Male	Female	Male	Female
Intercept	2.1907***	2.0917***	2.0578***	1.9831***
	(0.0931)	(0.0788)	(0.1093)	(0.0922)
MC	0.0180	0.0476	0.0192	0.0550*
	(0.0264)	(0.0254)	(0.0260)	(0.0238)
HEC	−0.0028	−0.0037	−0.0056	0.0061
	(0.0230)	(0.0270)	(0.0223)	(0.0258)
C	0.0338	0.0072	0.0330	0.0180
	(0.0317)	(0.0274)	(0.0333)	(0.0278)
H	0.0039	0.0057	−0.0049	0.0181
	(0.0259)	(0.0339)	(0.0267)	(0.0339)
A	−0.0309	−0.0121	−0.0137	−0.0129
	(0.0269)	(0.0302)	(0.0276)	(0.0298)
ES	−0.0042	−0.0423	−0.0074	−0.0328
	(0.0329)	(0.0294)	(0.0340)	(0.0273)
TR	0.0252	0.0296	0.0350	0.0172
	(0.0218)	(0.0234)	(0.0223)	(0.0224)
E	−0.0115	−0.0037	−0.0219	−0.0107
	(0.0246)	(0.0249)	(0.0240)	(0.0245)
SE	0.0304	0.0319	0.0346	0.0223
	(0.0262)	(0.0266)	(0.0262)	(0.0264)
LC	0.0973***	0.0657*	0.0816**	0.0515
	(0.0259)	(0.0296)	(0.0261)	(0.0286)
Ability10Y	0.0986**	0.1625***	0.0776*	0.1237***
	(0.0333)	(0.0328)	(0.0338)	(0.0332)
SocialClass10Y	0.0698***	0.0690***	0.0589**	0.0431*
	(0.0193)	(0.0199)	(0.0202)	(0.0199)
NonBritish	0.1293	0.2432*	0.0943	0.1250
	(0.1228)	(0.0994)	(0.1288)	(0.1204)
Children	0.0244	−0.1138***	0.0288	−0.1032***
	(0.0274)	(0.0236)	(0.0267)	(0.0237)
Married	0.0968	0.0611	0.0909	0.0576
	(0.0681)	(0.0522)	(0.0686)	(0.0518)
HighQual1			0.0137	0.0094
			(0.1133)	(0.1129)
HighQual2			0.1234	0.0864
			(0.0933)	(0.0674)
HighQual3			0.2240*	0.1080
			(0.0960)	(0.0871)
HighQual4			0.2343*	0.3596***
			(0.0952)	(0.0738)
HighQual5			0.3030*	0.4226***
			(0.1268)	(0.1187)
R^2 (In sample)	0.2160	0.2852	0.2395	0.3518
MSE (In sample)	0.1957	0.1876	0.1899	0.1702
R^2 (Out of sample)	0.1294	0.1402	0.1511	0.1959
MSE (Out of sample)	0.1962	0.2023	0.1913	0.1892

Least squares estimates of the wage equation for the sample of the 42 year old, augmented by skill factors based on PCA indices. Dependent variable: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 18: Wage Equation Estimates with PCA Index, Sample 38Y

	Male	Female	Male	Female
Intercept	2.2939***	2.2553***	2.2951***	2.1111***
	(0.0731)	(0.0773)	(0.1033)	(0.1285)
MC	−0.0096	0.0132	−0.0123	0.0065
	(0.0273)	(0.0316)	(0.0283)	(0.0299)
HEC	0.0053	0.0208	0.0012	0.0145
	(0.0261)	(0.0292)	(0.0269)	(0.0268)
C	−0.1010**	−0.0331	−0.0940**	−0.0292
	(0.0348)	(0.0348)	(0.0349)	(0.0351)
H	−0.0133	−0.0252	−0.0141	−0.0215
	(0.0319)	(0.0366)	(0.0323)	(0.0390)
A	0.0272	−0.0369	0.0335	−0.0287
	(0.0286)	(0.0365)	(0.0289)	(0.0319)
ES	−0.0205	0.0177	−0.0237	0.0308
	(0.0347)	(0.0356)	(0.0355)	(0.0348)
TR	0.0282	0.0234	0.0291	0.0253
	(0.0263)	(0.0272)	(0.0266)	(0.0269)
E	0.0222	0.0020	0.0234	−0.0055
	(0.0252)	(0.0384)	(0.0257)	(0.0374)
SE	−0.0445	−0.0122	−0.0380	−0.0054
	(0.0279)	(0.0286)	(0.0280)	(0.0300)
LC	0.0383	0.0931***	0.0284	0.0632*
	(0.0318)	(0.0275)	(0.0338)	(0.0271)
Ability10Y	0.1065**	0.1279***	0.1001**	0.0847*
	(0.0327)	(0.0378)	(0.0351)	(0.0399)
SocialClass10Y	0.0750***	0.0505*	0.0699***	0.0424*
	(0.0193)	(0.0197)	(0.0203)	(0.0202)
NonBritish	0.0942	0.1879	0.0345	0.1690
	(0.2837)	(0.1751)	(0.2909)	(0.1940)
Children	0.0325	−0.0371	0.0304	−0.0220
	(0.0314)	(0.0270)	(0.0307)	(0.0258)
Married	0.1735*	0.0271	0.1784**	0.0288
	(0.0699)	(0.0519)	(0.0674)	(0.0503)
HighQual1			0.1460	−0.0236
			(0.2103)	(0.1462)
HighQual2			−0.0809	0.0198
			(0.0906)	(0.1211)
HighQual3			0.0471	−0.0409
			(0.1115)	(0.1409)
HighQual4			0.0729	0.3095*
			(0.0954)	(0.1211)
HighQual5			−0.0043	0.3661*
			(0.1151)	(0.1672)
R^2 (In sample)	0.2587	0.2018	0.2808	0.2949
MSE (In sample)	0.1536	0.1925	0.1490	0.1701
R^2 (Out of sample)	0.1304	0.1713	0.1358	0.2341
MSE (Out of sample)	0.1640	0.1923	0.1630	0.1777

Least squares estimates of the wage equation for the sample of the 38 year old, augmented by skill factors based on PCA based indices. Dependent variable: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 19: Wage Equation Estimates with PCA Index, Sample 42Y

	Male	Female	Male	Female
Intercept	2.4892***	2.5604***	2.6434***	2.2015***
	(0.1078)	(0.1148)	(0.2106)	(0.1665)
MC	−0.0009	0.0156	−0.0308	0.0081
	(0.0399)	(0.0404)	(0.0413)	(0.0417)
HEC	−0.0698	0.0582	−0.0515	0.0623
	(0.0462)	(0.0379)	(0.0510)	(0.0383)
C	−0.0643	0.0429	−0.0253	0.0625
	(0.0528)	(0.0424)	(0.0466)	(0.0426)
H	0.0113	0.0395	−0.0007	0.0436
	(0.0514)	(0.0472)	(0.0465)	(0.0471)
A	−0.0396	0.0229	−0.0279	0.0341
	(0.0479)	(0.0420)	(0.0411)	(0.0426)
ES	0.0084	−0.0363	0.0178	−0.0243
	(0.0535)	(0.0409)	(0.0520)	(0.0365)
TR	−0.0377	0.0734*	−0.0194	0.0820*
	(0.0324)	(0.0370)	(0.0304)	(0.0335)
E	0.0322	−0.0001	0.0070	−0.0183
	(0.0543)	(0.0289)	(0.0526)	(0.0256)
SE	0.0122	0.0577	0.0471	0.0479
	(0.0329)	(0.0410)	(0.0315)	(0.0387)
LC	0.0693	0.0905*	0.0515	0.0527
	(0.0392)	(0.0402)	(0.0394)	(0.0380)
Ability10Y	0.0624	0.1585**	0.0282	0.1214**
	(0.0351)	(0.0513)	(0.0338)	(0.0466)
SocialClass10Y	0.0727**	0.0209	0.0434	−0.0246
	(0.0276)	(0.0306)	(0.0287)	(0.0294)
NonBritish	0.0926	0.7111***	0.0216	0.4615***
	(0.1989)	(0.1629)	(0.2285)	(0.1347)
Children	0.0221	−0.0672*	0.0183	−0.0549
	(0.0313)	(0.0339)	(0.0281)	(0.0325)
Married	0.1797*	0.0140	0.1441*	0.0535
	(0.0754)	(0.0731)	(0.0690)	(0.0701)
HighQual1			−0.1471	0.1804
			(0.2648)	(0.1673)
HighQual2			−0.3061	0.3223*
			(0.1899)	(0.1533)
HighQual3			−0.0263	0.2753
			(0.1937)	(0.1608)
HighQual4			0.1115	0.6746***
			(0.1948)	(0.1492)
HighQual5			0.2861	0.7295***
			(0.2043)	(0.1726)
R^2 (In sample)	0.2279	0.2692	0.3380	0.4202
MSE (In sample)	0.2015	0.1983	0.1727	0.1573
R^2 (Out of sample)	0.1902	0.0272	0.2068	0.1561
MSE (Out of sample)	0.1867	0.2719	0.1828	0.2358

Least squares estimates of the wage equation for the sample of the 42 year old, augmented by skill factors based on PCA based indices. Dependent variable: log hourly gross wage, White s.e., $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

Table 20: Group Lasso Wage Equation Estimates: Sample 38Y

	Male	Female
Intercept	2.2828***	2.0786***
	(0.1015)	(0.1196)
Ability10Y	0.1205***	0.1214***
	(0.0306)	(0.0357)
SocialClass10Y	0.0681***	0.0435*
	(0.0202)	(0.0192)
NonBritish	0.0471	0.1721
	(0.2808)	(0.2009)
Children	0.0319	−0.0222
	(0.0300)	(0.0253)
Married	0.1682**	0.0323
	(0.0647)	(0.0491)
HighQual1	0.1498	−0.0172
	(0.1937)	(0.1376)
HighQual2	−0.0858	0.0153
	(0.0906)	(0.1112)
HighQual3	0.0826	−0.0206
	(0.1049)	(0.1323)
HighQual4	0.0873	0.3358**
	(0.0924)	(0.1136)
HighQual5	0.0382	0.4086**
	(0.1091)	(0.1559)
X_MC	0.0374	
	(0.0411)	
R^2 (In sample)	0.2456	0.2744
MSE (In sample)	0.1563	0.1750
R^2 (Out of sample)	0.1607	0.2405
MSE (Out of sample)	0.1583	0.1762

Group Lasso estimates of the wage equation for the sample of the 38 year old, augmented by skill factors based on group lasso weighted indices. Dep. var.: log hourly gross wage, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, post-lasso s.e.

Table 21: Group Lasso Wage Equation Estimates: Sample 42Y

	Male	Female
Intercept	2.6862***	2.1653***
	(0.1847)	(0.1317)
Ability10Y	0.0602	0.1612***
	(0.0336)	(0.0417)
SocialClass10Y	0.0479	−0.0293
	(0.0273)	(0.0292)
NonBritish	0.0375	0.3903**
	(0.2175)	(0.1398)
Children	0.0213	−0.0491
	(0.0285)	(0.0324)
Married	0.1417*	0.0691
	(0.0673)	(0.0656)
HighQual1	−0.2026	0.1978
	(0.2301)	(0.1506)
HighQual2	−0.4064*	0.3586**
	(0.1605)	(0.1174)
HighQual3	−0.1259	0.3123*
	(0.1651)	(0.1259)
HighQual4	0.0561	0.6989***
	(0.1617)	(0.1130)
HighQual5	0.1700	0.8091***
	(0.1756)	(0.1332)
X_H	0.1405	
	(0.0810)	
R^2 (In sample)	0.2948	0.3547
MSE (In sample)	0.1840	0.1751
R^2 (Out of sample)	0.1619	0.2170
MSE (Out of sample)	0.1932	0.2188

Group Lasso estimates of the wage equation for the sample of the 38 year old, augmented by skill factors based on group lasso weighted indices. Dep. var.: log hourly gross wage, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, post lasso s.e.

Table 22: Unemployment Equation Estimates, Augmented Specifications, EW, Sample 34Y

	Male	Female	Male	Female
Intercept	-2.1178^{***}	-1.9896^{**}	-1.8876^{**}	-1.5621^{*}
	(0.5272)	(0.7113)	(0.6414)	(0.7911)
MC	0.0315	−0.0266	0.0227	−0.0180
	(0.1959)	(0.2787)	(0.1958)	(0.2898)
HEC	−0.2180	−0.0682	−0.2137	−0.0518
	(0.2212)	(0.2910)	(0.2217)	(0.2985)
C	−0.2629	0.0215	−0.2511	0.0213
	(0.2225)	(0.2926)	(0.2230)	(0.2926)
H	0.1103	0.5922	0.1051	0.5806
	(0.2349)	(0.3906)	(0.2341)	(0.3902)
A	0.2221	0.6433^{*}	0.2357	0.6304^{*}
	(0.1770)	(0.2572)	(0.1797)	(0.2588)
ES	0.2761	−0.6023	0.2700	−0.6153
	(0.2247)	(0.4107)	(0.2249)	(0.4198)
TR	0.1868	−0.0611	0.1997	−0.0377
	(0.1843)	(0.2526)	(0.1857)	(0.2567)
E	−0.3411	0.8061^{**}	−0.3475	0.8134^{**}
	(0.2029)	(0.2801)	(0.2026)	(0.2846)
SE	0.0611	0.0981	0.0565	0.1226
	(0.2015)	(0.2595)	(0.2010)	(0.2635)
LC	-0.4468^{*}	−0.1341	-0.4475^{*}	−0.1197
	(0.2081)	(0.2539)	(0.2091)	(0.2555)
Ability10Y	−0.1225	0.5344	−0.1185	0.6256
	(0.1994)	(0.3296)	(0.2043)	(0.3372)
SocialClass10Y	0.1181	−0.1514	0.1036	−0.1096
	(0.1503)	(0.2015)	(0.1535)	(0.2066)
Children	−0.2890	−0.1048	−0.2847	−0.1359
	(0.2093)	(0.2546)	(0.2106)	(0.2619)
Married	-0.8768^{*}	-1.2506^{*}	-0.8605^{*}	-1.2309^{*}
	(0.4122)	(0.4859)	(0.4149)	(0.4869)
HighQual1			−0.3774	−0.5920
			(0.4965)	(0.6140)
HighQual2			−0.1241	−0.8000
			(0.5352)	(0.6827)
Accuracy (In sample)	0.8066	0.8232	0.8125	0.7701
Specificity (In sample)	0.8298	0.8330	0.8383	0.7773
Sensitivity (In sample)	0.5476	0.6250	0.5238	0.6250
Accuracy (Out of sample)	0.7725	0.7736	0.7882	0.6909
Specificity (Out of sample)	0.8017	0.7934	0.8188	0.7045
Sensitivity (Out of sample)	0.4390	0.3750	0.4390	0.4167
Cutoff	0.1179	0.0704	0.1255	0.0552

ML logit estimates of the unemployment equation for males and females augmented specification with EW indices . Dependent variable: unemployed = 1, employed = 0, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 23: Unemployment Equation Estimates, Augmented Specifications, PCA, Sample 34Y

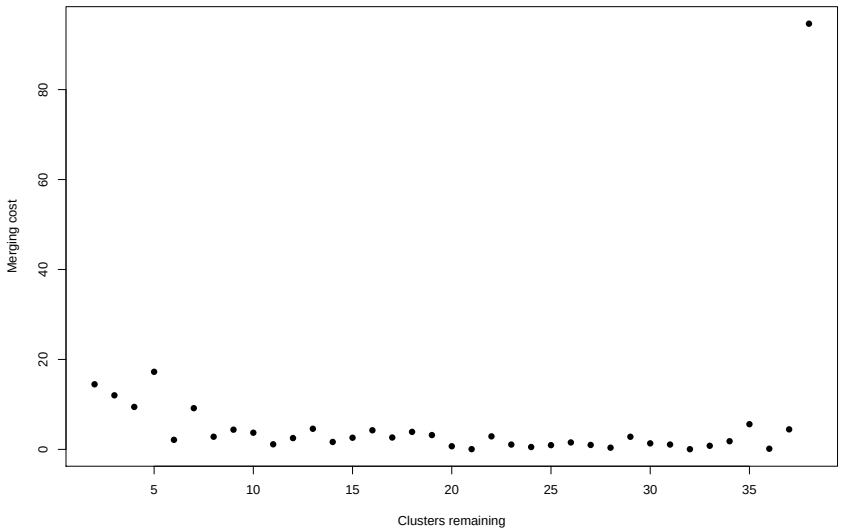
	Male	Female	Male	Female
Intercept	-2.1285^{***}	-1.9741^{**}	-1.8655^{**}	-1.5801^{*}
	(0.5264)	(0.7061)	(0.6367)	(0.7846)
MC	0.0355	−0.0016	0.0279	0.0091
	(0.2040)	(0.2896)	(0.2039)	(0.3011)
HEC	−0.2128	−0.1418	−0.2113	−0.1251
	(0.2230)	(0.3127)	(0.2223)	(0.3208)
C	−0.2684	−0.0076	−0.2537	−0.0142
	(0.2229)	(0.2971)	(0.2233)	(0.2985)
H	0.0761	0.5898	0.0698	0.5808
	(0.2350)	(0.3933)	(0.2344)	(0.3926)
A	0.1866	0.6196^{*}	0.2018	0.6036^{*}
	(0.1764)	(0.2589)	(0.1784)	(0.2615)
ES	0.2799	−0.5328	0.2725	−0.5395
	(0.2242)	(0.4098)	(0.2246)	(0.4177)
TR	0.2177	−0.0444	0.2316	−0.0274
	(0.1861)	(0.2530)	(0.1876)	(0.2573)
E	−0.3140	0.7994^{**}	−0.3221	0.8032^{**}
	(0.2042)	(0.2846)	(0.2037)	(0.2885)
SE	0.1053	0.0703	0.1025	0.0931
	(0.2063)	(0.2600)	(0.2055)	(0.2627)
LC	-0.4783^{*}	−0.1566	-0.4897^{*}	−0.1494
	(0.2095)	(0.2535)	(0.2112)	(0.2570)
Ability10Y	−0.1263	0.5364	−0.1173	0.6282
	(0.1986)	(0.3274)	(0.2037)	(0.3362)
SocialClass10Y	0.1153	−0.1501	0.1002	−0.1035
	(0.1487)	(0.2004)	(0.1521)	(0.2064)
Children	−0.2869	−0.0993	−0.2850	−0.1356
	(0.2086)	(0.2541)	(0.2101)	(0.2607)
Married	-0.8453^{*}	-1.2698^{**}	-0.8257^{*}	-1.2466^{*}
	(0.4131)	(0.4837)	(0.4159)	(0.4843)
HighQual1			−0.4354	−0.5432
			(0.4965)	(0.6125)
HighQual2			−0.1503	−0.7919
			(0.5365)	(0.6831)
Accuracy (In sample)	0.8438	0.7780	0.8438	0.6365
Specificity (In sample)	0.8766	0.7856	0.8745	0.6309
Sensitivity (In sample)	0.4762	0.6250	0.5000	0.7500
Accuracy (Out of sample)	0.7980	0.7185	0.8020	0.5630
Specificity (Out of sample)	0.8294	0.7355	0.8337	0.5640
Sensitivity (Out of sample)	0.4390	0.3750	0.4390	0.5417
Cutoff	0.1374	0.0592	0.1374	0.0360

ML logit estimates of the unemployment equation for males and females augmented specification with PCA indices . Dependent variable: unemployed = 1, employed = 0, $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

A.2 Figures

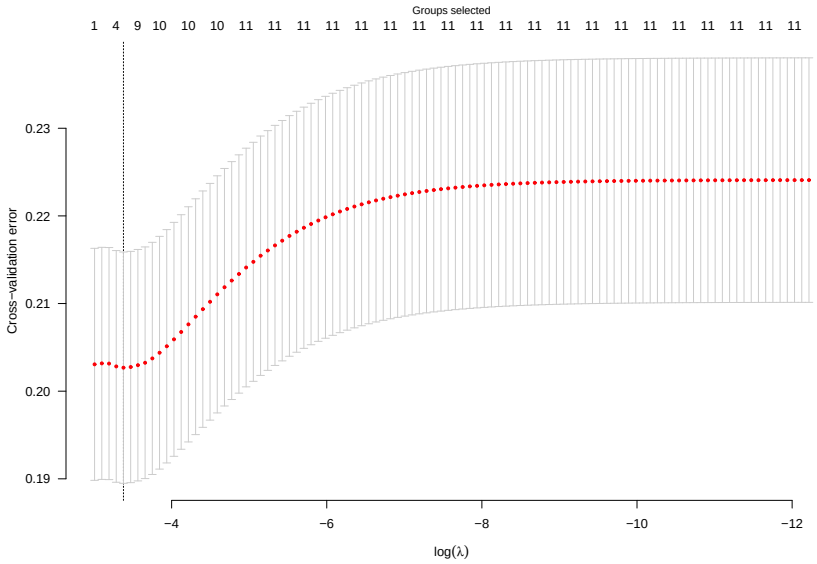
A.2.1 Cluster Analysis

Figure 2: Merging Costs of K clusters



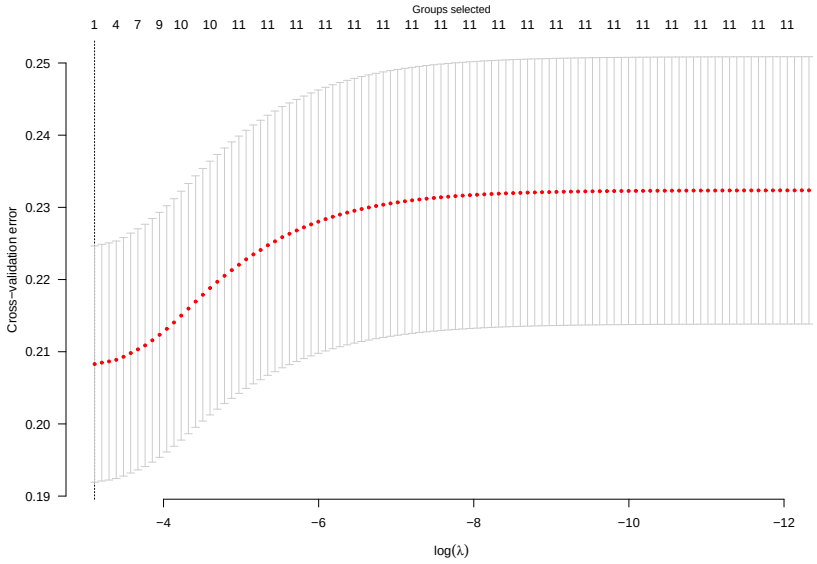
A.2.2 Cross-Validation of the Wage Equation

Figure 3: 5-fold CV λ, Group Lasso, Male, 34Y



The group lasso chooses 4 groups for males: Hand-Eye Coordination, Behavioral Trauma, Extraversion and Locus of Control.

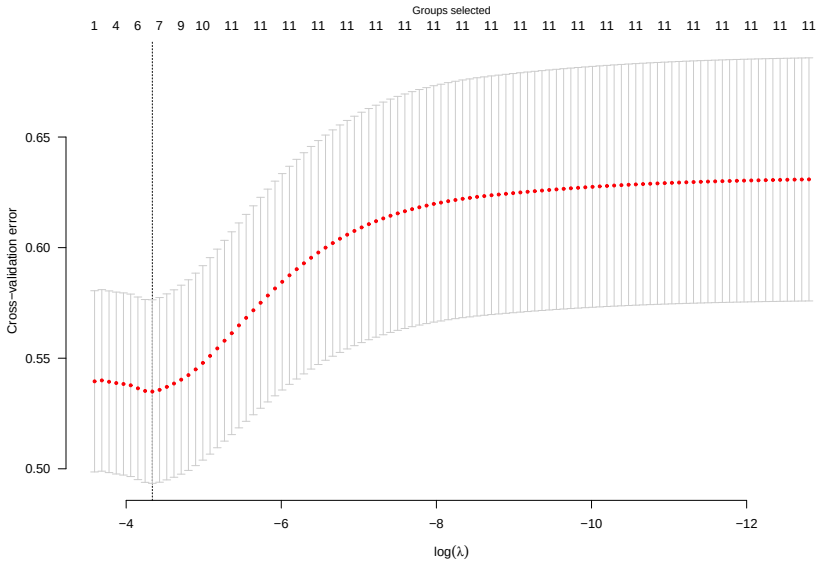
Figure 4: 5-fold CV λ, Group Lasso, Females, 34Y



The group lasso chooses 0 groups for females.

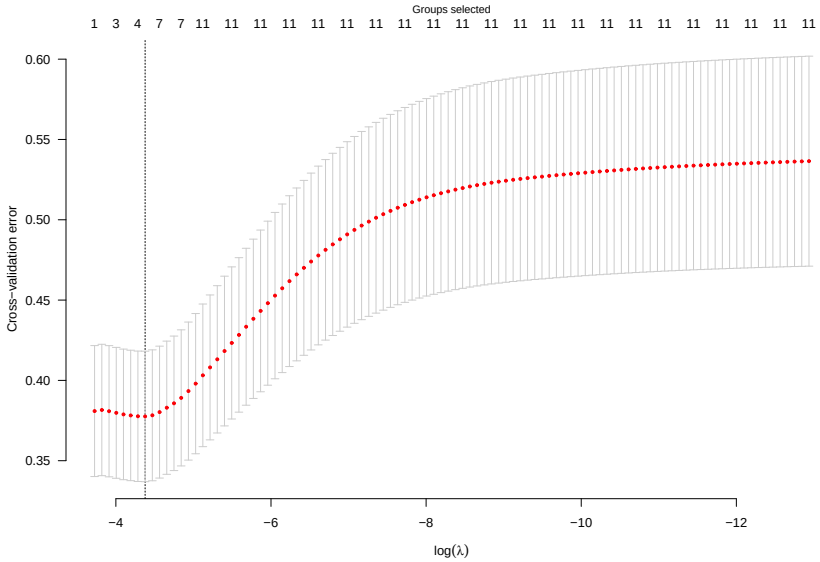
A.2.3 Cross-Validation of the Unemployment Equation

Figure 5: 30-fold CV λ, Group Lasso, Males, 34Y



The group lasso chooses 5 groups for males: Agreeableness, Behavioral Trauma, Extraversion, Self-Esteem and Locus of Control.

Figure 6: 30-fold CV λ, Group Lasso, Females, 34Y



The group lasso chooses 3 groups for females: Agreeableness, Conscientiousness and Emotional Stability.