How well can Noncognitive Skills Predict Unemployment? A Machine Learning Approach

Preprint · April 2019

DOI: 10.13140/RG.2.2.33573.96488

2 authors:

Jana Mareckova
University of St. Gallen
9 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Winfried Pohlmeier
University of Konstanz
117 PUBLICATIONS **1,796** CITATIONS
SEE PROFILE

How well can Noncognitive Skills Predict Unemployment? A Machine Learning Approach*

Jana Marečková† Winfried Pohlmeier‡

University of St. Gallen, GSDS University of Konstanz, CASCB, RCEA

This Version: April 14, 2019

Abstract

We study the predictive quality of noncognitive skill measures by means of machine learning techniques. Unlike previous empirical approaches centering around the in-sample explanatory power of noncognitive skills, our approach focuses on the performance of predicting individual unemployment over a long-run horizon of 20 years and more. Our machine learning approach can cope not only with the challenge of selecting the most relevant factors from data with a large number of skill measures but also leads to a sparse set of skill measures which is economically and psychologically interpretable.

Using data from the British Cohort Study (BCS), we compare the predictive power of different noncognitive skill measures and illustrate, how our estimates can be used to optimize the assignment mechanisms for manpower training programs and psychological intervention schemes for youths and young adults.

Keywords: noncognitive skills, unemployment, machine learning, group lasso, prediction

JEL classification: J24, J64, C38 PsycINFO classification: 20 3100, 21 3600

*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. Earlier versions of the paper have been presented at the 22^{nd} SoLE conference, in Raleigh May 2017, and the German Economic Association annual meeting, in Vienna Sept. 2017. For a helpful discussion on an earlier version of the paper we like to thank Anja Achtziger and Stefanie Schurer. All remaining errors are ours.

†Corresponding author: University of St. Gallen, SEW-HSG, Varnbüelstrasse 14 9000 St. Gallen, Switzerland. Phone: +41 71 224 23 42, email: jana.mareckova@unisg.ch.

‡University of Konstanz, Department of Economics, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49 753188 2660, email: winfried.pohlmeier@uni-konstanz.de.

1 Introduction

The role of noncognitive skills in explaining individual differences in educational attainment and labor market performance 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 educational attainment and labor market performance. 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 predicting differences in individual unemployment by means of machine learning techniques. We introduce a multi-step procedure combining unsupervised and supervised machine learning techniques accomplishing: (1) dimensionality reduction of the survey items into noncognitive skill indices in two data-driven steps, (2) a target-oriented index construction of noncognitive skill measures (3) yielding an interpretable predictive model which can be (4) either statistically or economically tuned for classification. Contrary to the standard empirical studies, we follow a strictly predictive modeling approach to data analysis by focusing on the out-of-sample classification qualities of noncognitive skill measures for future unemployment. In the presence of a panoply of competing personality concepts, regularization is particularly attractive in terms of variable and model selection. Our predictive modeling approach contributes to a better understanding of the relevance of noncognitive skills for individual labor market performance for several reasons. 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 items capturing several dimensions of noncognitive skills. Thus, they are not prone to in-sample overfitting, a problem that is likely to occur if many 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. Finally, selecting a sparse model specification with superior predictive performance is attractive when it comes to rank competing predictive factors with respect to their relevance for the design of appropriate intervention strategies or manpower training programs.

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 predictive power of noncognitive skills for individual unemployment. 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 factors out of a large set of potentially relevant factors and eliminate factors of minor relevance in order to obtain a sparse and easy to interpret model. In addition to the simple lasso (Tibshirani, 1996), grouping leads to a further dimension reduction by selecting complete groups of variables for the model specification while suppressing the less relevant groups. In particular, we show, how lassoing and grouping can be used to construct target-oriented 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 based on the British Cohort Study (BCS), we construct target-oriented indices of noncognitive skills for unemployment based on 119 survey items. Noticeably, our approach is not limited 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.

In this study, we propose an interpretable predictive modeling approach. Contrary to a purely predictive modeling approach optimized to find the best predictions based on a large set of covariates, our strategy is to select groups of variables which can serve as the basis

for psychological intervention strategies or manpower training programs. For example, the individually identified groups of factors (e.g. lack of conscientiousness or lack of openness) driving the probability of becoming unemployed may call for an appropriate intervention strategy. Moreover, if the purpose of predictive modeling is an economically optimal design of intervention strategies, tuning the quality of model's predictions with respect to a conventional statistical prediction criterion (e.g. accuracy of classification or informedness) is suboptimal. Therefore we present, in addition to the selection of groups of factors, two approaches to select the threshold parameter in classification models. The first approach is a standard in the literature based on optimizing a statistically motivated criterion ignoring economic costs of the resulting classification. In the alternative second approach we replace the statistical criterion by an economic criterion to tune the threshold parameter so that the resulting classification leads to economic efficiency. Using unemployment classification as an example, with the economic approach the model is tuned to take the overall costs and benefits of an unemployment program into account and classify only those as unemployed so that the whole program is cost efficient.

The paper is organized as follows. In Section 2 we elaborate on the potential merits of predictive modeling to assess the impact of noncognitive skills on individual labor market performance. Moreover, we introduce our group lasso approach to select measures of noncognitive skills and compare it with traditional factor analytic approaches. In Section 3 we introduce the approach of intervention optimal classification. We describe the BCS sample and discuss further data issues in Section 4. The structure of BCS data allows us to set up an early warning system predicting unemployment over a horizon of 24 years using covariates measured at the age of 10. Our empirical findings are presented in Section 5, while Section 6 concludes and provides an outlook on future research.

2 A Machine Learning Approach to Select Skill Factors

In what follows, we present a general modeling strategy to select noncognitive skill factors to predict individual unemployment. Our empirical application focuses on the selection of factors, which are capable of predicting future unemployment best over a forecasting horizon of 24 years. We argue, that being able to identify the factors with high predictive ability may be helpful to optimize the design for the selection of individuals at risk into intervention strategies and

manpower training programs.

Explaining vs. Predicting

When modeling the impact of noncognitive skills for individual labor market performance, conventional statistical approaches in labor economics and personnel psychology almost exclusively follow an explanatory research strategy. Their goal is to identify and to quantitatively assess at best causal relationships with a high internal validity obtained by maximizing the in-sample explanatory power. In contrast, machine learning techniques generally follow a predictive modeling strategy focusing on high external validity and true out-of-sample predictive performance, i.e. the best model is the one that predicts or classifies the outcome variable best on data which have not been used for the estimation process. In the first place, we regard our empirical strategy, which is a combination of various machine learning techniques (unsupervised learning, group lasso, cross-validation etc.), as being complementary to the explanatory strategies of conventional empirical studies on observational data. Shmueli (2010) summarizes the scientific purpose and added value of predictive modeling in the following four main points:

i. Detection of complex relationships and patterns that are hard to hypothesize in large and rich data sets, especially given theories that exclude newly measurable concepts.

ii. Discovery of new measures as well as comparison of different operationalizations of constructs and different measurement instruments.

iii. Enhancement of existing explanatory models.

iv. Reality check of the strength of causal relationships.

As competing noncognitive skill factors are compared in terms of their predictive performance, our machine learning approach follows a predictive modeling strategy serving to some extent all four purposes listed above. Typically, empirical studies with many potential explanatory factors focusing on the in-sample fit suffer from the risk of over-fitting. Predictive modeling strategies naturally overcome this issue leading to a sparser model specification which separates relevant predictors from less relevant ones.

However, our modeling strategy is not a purely predictive one. It rather can be considered as a hybrid approach in the sense that not single, difficult to interpret variables (e.g. single survey items) are identified among a large set of variables as having a high predictive power like in a kitchen-sink regression. We rather follow a two-stage strategy where groups of survey items are

identified in the first stage by means of cluster analysis unveiling the items that measure the same noncognitive skill factors. In the latter stage, the grouped survey items are then studied with respect to their predictive performance using group-wise L_1-regularization to yield sparse model specifications, which are easy to interpret and reveal a high predictive quality.

Large-scale Skills Measures

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 of the covariate space 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 yielding a low-dimensional index representation. Alternatively, some type of dimension reduction is implicitly accomplished by focusing on certain noncognitive skills (e.g. Big Five, locus of control) and disregarding covariates reflecting more closely alternative (complementary or competing) concepts.

In case of the BCS with its rich information on noncognitive skill measures, the dimensionality of the predictive model grows along two dimensions: (i) number of survey items capturing different noncognitive skills and (ii) number of noncognitive skill indices built from the survey items. As mentioned above, instead of using all the survey items as covariates in the model, the dimensionality of the data set is first reduced by calculating an index from survey items capturing the same noncognitive skill factor. Either the researcher has a prior information on which survey items measure certain noncognitive skills or he applies a statistical dimension reduction technique like PCA or factor analysis to find the index structure in the data set. Apart from PCA, any other unsupervised machine learning technique can be applied that is able to unveil the grouping of the items. Therefore as a first step, we use clustering to collect survey items into groups representing particular noncognitive skills. This leads to the first dimensionality reduction along the dimension (i). The survey items in a group are then used later to calculate an index, which proxies a certain noncognitive skill factor.

In the second step, we plug the grouped survey items into the model and let group lasso to decide which of these groups are important for predicting individual unemployment and to

estimate the group lasso index weights. More details on alternative approaches how to construct the weights are given in Section 2.2. The second step reduces the model dimensionality along the dimension (ii) by reducing the number of noncognitive skill indices in the model. We compare in Section 5 results of the group lasso estimates with more conventional ways of index construction for different model variants.

2.1 L_q-Regularization

In order to understand how the group lasso helps to select and estimate index weights, let us briefly introduce the general idea behind L_q-regularization first. Consider the following least squares minimization problem:

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

where $\|Y - X\beta\|_2^2$ denotes sum of squared errors of a linear model, Y is an $N \times 1$ vector of the dependent variable, N represents the number of observations, X is an $N \times k$ matrix of covariates, β is an unknown $k \times 1$ parameter vector and s is a chosen non-negative constant for $q \geq 0$. When $q = 1$, the (1) is called LASSO (Least Absolute Shrinkage and Selection Operator) optimization problem introduced by Tibshirani (1996). The advantage of the L_1-regularization (i.e. $q = 1$, LASSO) over the L_2-regularization ($q = 2$, ridge) 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 predictors yielding a sparse model specification among a large set of possible specifications. The parameter q can take any positive value. However, it can be shown that only for $0 \leq q \leq 1$ there is a non-zero probability to select covariates and to shrink the impact of less relevant covariates to zero (e.g. Hastie et al. (2015)).

The primal problem (1) can be reformulated in a Lagrangian representation:

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

where $\sum_{j=1}^{k} |\beta_j|^q$ is the penalty function and λ the regularization (or penalty) parameter. For $0 \leq q \leq 1$, the choice of λ determines the sparsity of the solution, i.e. the size of λ determines how many elements in the estimated β are set to 0. The larger the λ is, the sparser is the

specification obtained and the fewer covariates are included in the regression equation. However, applying the standard lasso as a selection tool to choose the most relevant survey items as proxies for noncognitive skills is somewhat opaque as some of the items in a questionnaire are likely to be highly correlated and in such a setting, the standard lasso technique randomly chooses one predictor from each group of highly correlated predictors not caring which one is selected, therefore making the model estimates hard to interpret (Zou and Hastie, 2005).

Group Lasso

The group lasso introduced by Yuan and Lin (2006) overcomes this problem and accounts for a given natural group structure in the data. It guarantees that all coefficients within a group become nonzero or zero simultaneously. Consider now the case where covariates (e.g. facets in the Big Five framework or a set of survey items measuring noncognitive skills) can be divided into J groups with k_j covariates in group j. The selection of groups is achieved by solving:[1]

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

where Y is an $N \times 1$ vector of the dependent variable, N represents the number of observations, X_j is an $N \times k_j$ matrix of covariates corresponding to the j-th group of skill factors, β_j is a $k_j \times 1$ parameter vector for group j. The method allows for additional covariates contained in Z with a corresponding parameter vector δ which is not subject to regularization, i.e. the covariates Z are always in the model.

Since $q = 0.5$ for every parameter vector β_j, the choice of the regularization parameter λ determines how many groups are selected. Typically, the optimal λ is chosen by K-fold cross-validation, see e.g. Hastie et al. (2015) for more details. The larger the λ, the more elements in the penalty term $\sum_{j=1}^{J} \sqrt{k_j} \|\beta_j\|_2 = \sum_{j=1}^{J} \sqrt{k_j} \sqrt{\beta_{j,1}^2 + \cdots + \beta_{j,k_j}^2}$ are forced to zero in order to minimize (3). In case of group lasso, the elements in the penalty term 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. In addition, the penalty for a group j receives more weight, the larger the number of group items k_j is. This balances out the effect that bigger groups would be more likely to be selected without this

[1] See Hastie et al. (2015) for different variants of the group lasso.

correction. 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.

Lassoing and group lassoing as briefly described above 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, we use a regularized maximum likelihood logit approach where the least squares part of objective function in (2) is replaced by the negative log likelihood of the logit model. Under sparsity the group lasso is consistent in a sense that a difference between the estimated and true logit link function is bounded with high probability even when the number of predictors exceeds the number of observations (Meier et al. (2008)).

2.2 Index Construction

In the following, we use the properties of L_q-norm regularization for the construction and selection of skill indices along with the conventional approaches. Assume that the grouping of the survey items with respect to different noncognitive skill factors is given. Detailed information about how ex-ante grouping can be achieved by clustering as used in this paper is provided in Section 4. For example, let $C_{i1}, C_{i2}, \ldots, C_{iK_C}$ be the responses of individual i on K_C different items related to conscientiousness and $E_{i1}, E_{i2}, \ldots, E_{iK_E}$ be the responses of individual i on K_E different items related to extraversion, then in general the indices for the two personality traits take the form of a weighted average across the corresponding items:

$$IC_i^C = \sum_{l=1}^{K_C} \omega_l^C C_{il}, \qquad IC_i^E = \sum_{l=1}^{K_E} \omega_l^E E_{il},$$

with $\sum_l \omega_l^C = \sum_l \omega_l^E = 1$. In the following, we consider three different strategies to determine the index weights ω_l^I, where I is a set collecting all available personality traits. In our example: $I = \{C, E\}$.

1. *Equal weights*: $\hat{\omega}_l^I = 1/K_I$

   This is the most simple and most often used way of constructing a skill index. It is not data driven and ignores that some items maybe superior to others in approximating the underlying skill factor.

2. *PCA based weights*: $\hat{\omega}_l^I = \pi_l^{PCA,I} / \sum_l \pi_l^{PCA,I}$.

   The weights are constructed by taking loadings of the first component as $\pi_l^{PCA,I}$'s. This index is data driven and assigns weights to items in a sense that the variance of the obtained scores is maximal.

3. *Group lasso based weights*: Here we use for illustrative purposes a linear model

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

   and estimate it by a group lasso. The index weights are constructed from the estimated regression coefficients: $\hat{\omega}_l^I = \frac{\hat{\beta}_l^I}{\sum_l \hat{\beta}_l^I}$ (see Appendix B). Note that the size of the weights for a certain item now depends on how strongly this item can predict the outcome variable. The size of the weights is now data driven and context specific, i.e. the relative importance of the items can differ and depends on the outcome variable under consideration. This may be important for items from surveys which were not designed for a specific purpose and items that are only loosely connected to a specific theory of noncognitive skills. Moreover, note that the use of group lasso also allows to select the most relevant groups (indices) for the model, e.g. the group lasso might assign zero coefficients to all conscientiousness items, i.e. $\hat{\beta}_l^C = 0, \forall l$. In this case, all the $\hat{\omega}_l^C$ will get zero values and conscientiousness would be considered as irrelevant for predicting Y_i. In contrast to the other two methods of index construction, the group lasso has a selection property while the (i) equally weighted and (ii) the PCA based approaches always include the entire set of indices in the predictive model. The resulting index with context specific weights is called target-oriented as the outcome variable of interest determines the index weights.

2.3 The Predictive Model

As in the previous subsection, assume that the grouping of items is given. The goal is to find the best predictive model based on the following form:

$$Y_i = g_i(\theta) = g(\alpha + Indices_i' \, \gamma + OtherControls_i' \, \delta) + \varepsilon_i, \tag{4}$$

where Y_i is the individual labor market outcome to be predicted and $\theta = (\alpha, \gamma', \delta')'$ the parameter vector to be estimated. The functional form $g(\cdot)$ depends on the type of dependent variable to be predicted. In our empirical application Y_i is the binary indicator for unemployment, so that we simply choose $g(\cdot)$ to be the cdf of the logistic distribution. The different noncognitive skill factors are captured by the $Indices$. In order to study the predictive performance of noncognitive skills and to check how the different methods of index construction affect the quality of the predictions we compare the following 9 model variants:

A. Baseline model

   1. Model with control variables only (M_0): $\gamma = 0$

B. Restricted models with noncognitive skill factors without control variables: $\delta = 0$

   2. Model with equally weighted indices (*onlyEW*),
   3. Model with PCA indices (*onlyPCA*),
   4. Model with group lasso indices (*onlyGL*),
   5. Model with post group lasso weighted indices (*post only GL*).

C. Unrestricted models with noncognitive skill factors and control variables: $\delta \neq 0$ and $\gamma \neq 0$

   6. Model with noncognitive skills with equally weighted indices and controls (*EW*),
   7. Model with noncognitive skills with PCA indices and controls (*PCA*),
   8. Model with noncognitive skills with group lasso indices and controls (*GL*)[2],
   9. Model with noncognitive skills with post group lasso indices and controls (*post GL*).

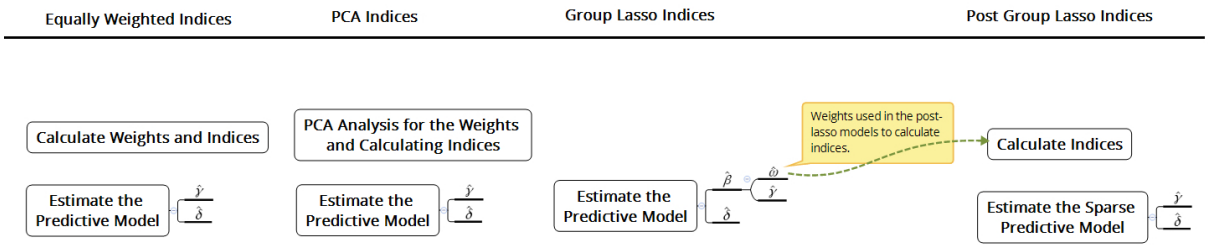
Model A.1 serves as the benchmark model. A comparison with models C.6-C.9 yields insights to what extent noncognitive skill factors contain any additional predictive power. Comparing models B.2-B.5 with models C.6-C.9 reveals how controls including cognitive skills improve the predictive performance given the set of noncognitive skill factors. Additional cross-comparisons can show whether cognitive and noncognitive skill overlap in predictive power for individual labor market outcomes.

Assuming that the grouping of the survey items for the index construction is known, Figure 1 represents the steps for the model estimation for each index type. Equally weighted indices can be calculated directly. For the PCA indices, one has to run the PCA analysis and calculate the indices using the first principal component for the weights. For the group lasso indices, a logit model with all survey items as covariates is estimated by group lasso imposing the known group

[2]In our implementation, we regularized also the controls, i.e. speaking in terms of notation used in (3), the covariates Z were integrated into X_j's with a correspondingly augmented group structure.

structure. The target-oriented index weights can be calculated from the β's yielding context specific weights. The optimal λ is obtained from the 10-fold-cross-validation.

Figure 1: Model Estimation Step for the different Index Schemes



After the group lasso estimation, the researcher can decide whether to continue the analysis with the group lasso model using the survey items as covariates or use the indices in a so-called post-lasso estimation.[3] The next step in the post-lasso estimation is to calculate the indices and re-estimate (4) with the selected indices and other controls by a standard maximum likelihood to get $\hat{\gamma}$ and $\hat{\delta}$. The potential advantage in this step is to get conditionally debiased prediction coefficients in a sparse model.

3 Intervention Optimal Classification

In order to use the model estimates for classification, a threshold parameter τ has to be chosen, which determines which class is assigned to each observation. For our binary logit model the classification rule is given by:

$$\hat{Y}_i(\tau) = \begin{cases} 1 & \text{if } g_i(\hat{\theta}) > \tau \\ 0 & \text{if } g_i(\hat{\theta}) \leq \tau \end{cases}. \tag{5}$$

Table 1 captures outcomes of a predictive exercise for any threshold parameter: (1) correctly predicted cases (TP = 'true positives' and TN = 'true negatives') and (2) misclassified cases (FP = 'false positives' and FN = 'false negatives'). In our empirical application correctly classified individuals are those who are unemployed and for whom the early warning system has predicted unemployment ('true positives') as well as individuals, who are employed and were not predicted to become unemployed ('true negatives'). An individual, who is falsely predicted as unemployed

[3] It is possible to identify the index weights ω and index coefficients γ from β in the group lasso model. Regarding group lasso model predictions, it is not necessary to do so as using β with the survey items would yield the same predictions as calculating the index and using γ. Identifying ω is necessary for the post-lasso estimation.

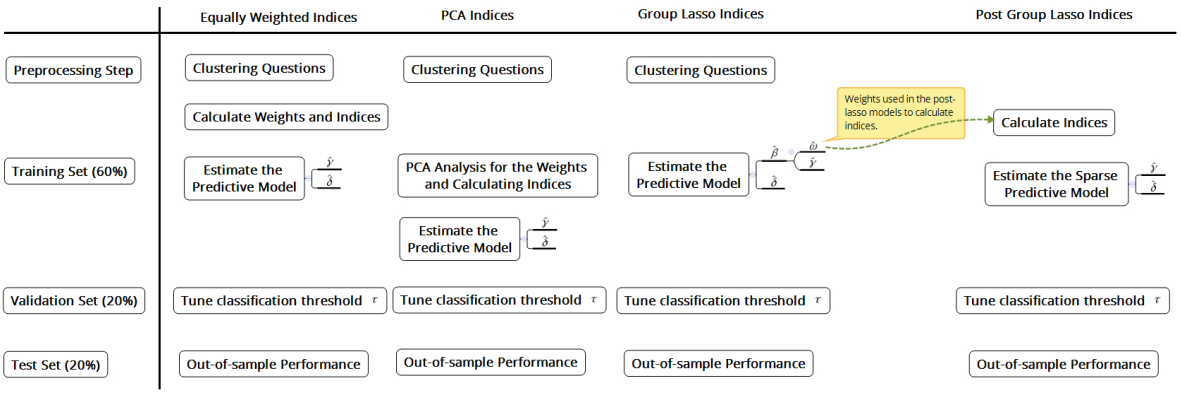
Table 1: Confusion Matrix

	true value	
prediction	$Y = 1$	$Y = 0$
$\hat{Y}(\tau) = 1$	True Positive (TP)	False Positive (FP)
$\hat{Y}(\tau) = 0$	False Negative (FN)	True Negative (TN)
total	All positives in the data (P)	All negatives in the data (N)

but will not experience unemployment belongs to the 'false positives'. Finally, the group of 'false negatives' consists of workers who are unemployed but were not detected to be at risk by the early warning system.

Following the goal to get a predictive model which performs well out-of-sample, the threshold parameter is tuned out-of-sample on a part of the data set which was not used to estimate the model coefficients θ. We refer to the data set used for estimation of θ as training set and the data set used for tuning τ as validation set. Having obtained the optimal τ, we then use it to perform the out-of-sample classifications for the test set data. The full procedure is captured in Figure 2.

Figure 2: Modeling Steps for the different Index Schemes



Statistically Optimal Intervention

From a statistical point of view, the optimal threshold parameter is chosen such that a certain statistical measure is optimized (e.g. accuracy or informedness). Conventional statistical measures to assess several threshold parameters are: (i) accuracy (share of correct predictions = (TP + TN)/(P+N)), (ii) sensitivity (true positive rate = $1 - \beta(\tau)$ = TP/P), (iii) specificity (true negative rate = $1 - \alpha(\tau)$ = TN/N) and (iv) positive predicted value (PPV, share of correctly

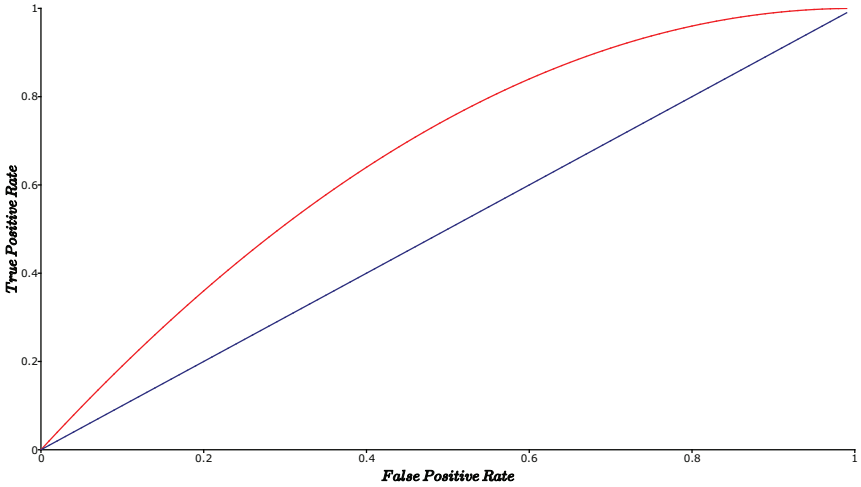
predicted positives among all predicted positives = TP/(TP+FP)). Such statistical criteria may fail to deal with several factors. One of them is an unbalanced sample. By focusing only on accuracy in a setting such as unemployment, an "accurate" predictive model might predict everyone as employed yielding 89-90% level of accuracy given that unemployment rates are between 10 and 11%. Such a scenario would lead to high accuracy, low sensitivity and high specificity. Therefore, it is important to focus also on sensitivity and specificity. The optimal scenario is to have a model with both, high specificity and sensitivity. High specificity implies that a positive prediction can be used to ruling in unemployment. Together with a high PPV, this would be an indication of a model which is reliable to predict unemployment. Low PPV indicates many false positives relative to true positives, i.e. predicting someone as unemployed might be a false alarm with a high probability. A standard approach for fine tuning of the threshold parameter in a statistical sense is to use the information in the Receiver Operating Characteristics (ROC) curve, which captures the sensitivity and specificity trade-off (see Fawcett (2006) for an introduction).

Figure 3 illustrates the out-of-sample ROC curve for a classifier. The ROC curve is used to assess statistical qualities of a classifier and to find the optimal threshold in a statistical sense. For our numerical example, the classification is effective, since the ROC curve lies above the 45-degree line representing the location of purely random sorting where the predicted outcomes are independent of the true outcomes. The area under the curve is $AUC = 0.66$ representing the probability that a randomly chosen unemployed person is classified as more likely to be unemployed than a randomly chosen employed person. The best threshold parameter τ in terms of Youden's index arises for a false positive rate, FPR $= \alpha(\tau_{opt}) = 0.5$. Geometrically, the best Youden's J-statistic maximizes the vertical distance of the ROC curve from the 45-degree line, i.e. the most informative model is chosen relatively to the random guess.

Economically Optimal Intervention

The statistical criteria do not capture the real price of misclassification. In case of predicting a low fraction of positives in the data, the typical trade-off is that by improving the number of true positives (high sensitivity), the number of false positives increases as well (low specificity). In an economic cost-benefit analysis this trade-off translates to the task of finding a point at

Figure 3: Receiver Operator Characteristics Curve



ROC curve (red) and 45-degree line (blue), AUC = 0.66, Youden's index = 0.25.

which the marginal cost equals marginal benefit to determine how many people should enter a program in order to achieve economic efficiency. In the unemployment context, we are thinking of any kind of early stage interventions, such as a manpower training program (MTP) for young adults, who are at risk of becoming unemployed in later years. The assignment into treatment and non-treatment is determined by choosing the threshold parameter of classification such that expected per-capita costs of the intervention are minimized in the same fashion as suggested by Metz (1978). The economically optimal threshold parameter could be very different from the statistically optimal one as it will be shown.

In our economic calculation, cases predicted as unemployed ('predicted positives') are assigned to an MTP and cases predicted as employed ('predicted negatives') are considered as in no need for an MTP. In this framework, the group of 'true negatives' does not generate any costs for the training program, while the 'true positives' receive an intervention with c as the treatment cost per capita. Moreover, assume that for the 'true positives' net unemployment costs per capita are φu, where u denotes unemployment benefits per capita and $0 < \varphi < 1$ is the percentage reduction of the unemployment benefit payments due to successful participation in the intervention. However, intervention cost also arises for an individual, who is falsely assigned to a training program but will not experience unemployment ('false positive'). Finally, for the group of 'false negatives', the unemployment benefits, u, have to be paid to full extent.

Table 2 reports the costs depending on the underlying (mis-)classification, where $Y = 1$

denotes a true unemployment status and $Y = 0$ employment. An individual is predicted to become unemployed $\hat{Y} = 1$, if the estimated predictor function $g_i(\hat{\theta})$ exceeds a given classification threshold τ.

Table 2: Confusion Matrix of Costs per Worker

	unemployment status	
predicted	$Y = 1$	$Y = 0$
$\hat{Y}(\tau) = 1$	$\varphi u + c$ (TP)	c (FP)
$\hat{Y}(\tau) = 0$	u (FN)	0 (TN)

Table 3 gives the classification probabilities in terms of true negative rate, $1 - \alpha(\tau) = \Pr\left[\hat{Y}(\tau) = 0 \,|Y = 0\right]$, i.e. specificity, and true positive rate $1 - \beta(\tau) = \Pr\left[\hat{Y}(\tau) = 1 \,|Y = 1\right]$, i.e. sensitivity, where $\pi = \Pr[Y = 1]$ is the probability of becoming unemployed (prevalence).

Table 3: Confusion Matrix of the Classification Probabilities

	unemployment status		
predicted	$Y = 1$	$Y = 0$	
$\hat{Y}(\tau) = 1$	$\pi(1 - \beta(\tau))$	$(1 - \pi)\alpha(\tau)$	$\pi(1 - \beta(\tau)) + (1 - \pi)\alpha(\tau)$
$\hat{Y}(\tau) = 0$	$\pi\beta(\tau)$	$(1 - \pi)(1 - \alpha(\tau))$	$\pi\beta(\tau) + (1 - \pi)(1 - \alpha(\tau))$
total	π	$1 - \pi$	1

For a given classification scheme with threshold value τ, expected costs of assignment into an early treatment are given by:

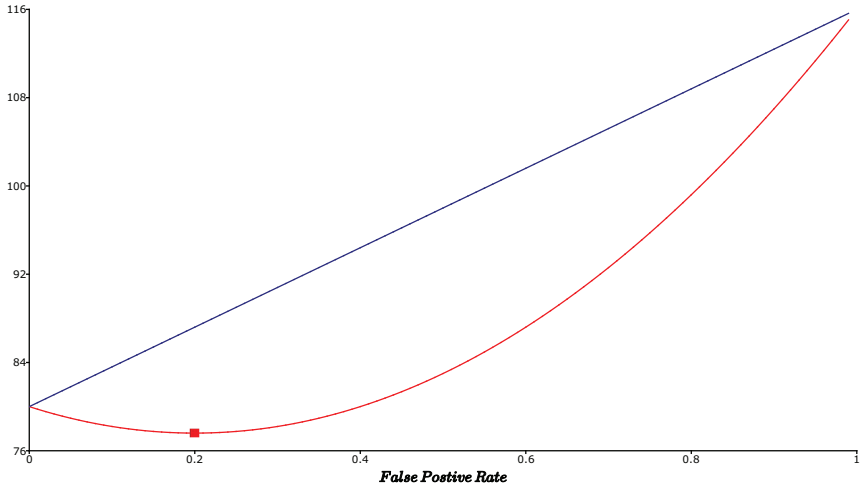
$$\mathrm{E}\left[C(\tau)\right] = \big(\pi(1 - \beta(\tau)) + (1 - \pi)\alpha(\tau)\big)c + \big(\pi(1 - \beta(\tau))\varphi + \pi\beta(\tau)\big)u. \tag{6}$$

Thus, optimal classification is defined by the threshold value that minimizes expected costs $\tau_{opt} = \arg\min_{\tau} \mathrm{E}\left[C(\tau)\right]$.

For our numerical example, expected assignment costs are depicted in Figure 4. The economically optimal threshold parameter is determined by the minimum of the expected assignment costs. In our example, the intervention optimal classification is obtained for a false positive rate of $\mathrm{FPR} = \alpha(\tau_{opt}) = 0.2$ and a true positive rate of $\mathrm{TPR} = 1 - \beta(\tau_{opt}) = 0.36$. In terms of these cost considerations, economically optimal assignment indicates that the requirements for assignment into treatment should be more restrictive than the statistically based optimal assignment, such that only 20 p.c. of the individuals not at risk of becoming unemployed will be assigned for the program (FPR = 0.2). Obviously the cost reduction results from not assigning

too many individuals to the program who do not need it. The gain from such a policy would be that around 36 p.c. of those who will face unemployment in the future can profit from the intervention. Compared to the statistically optimal classification, the cost minimizing assignment strategy is more restrictive as for the Youden's index maximizing threshold parameter there are considerably more false positives and more true positives assigned to the program.

Figure 4: Expected Assignment Costs



Expected cost of assignment as a function of the false positive rate (red line). The blue line indicates the expected cost under random assignment for $\pi = 0.04$, $u = 2000$, $c = 100$ and $\theta = 0.2$.

4 Data and Variable Construction

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 captured in 119 survey items. 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 longitudinal character of the data set enables us to analyze predictive power of early childhood environment and early cognitive and noncognitive skills measured at the age of 10 on adult labor market outcomes at the age of 34. Thus, our prediction horizon amounts to 24 years. After cleaning, imputing answers for the

missing answers of the noncognitive skills[4] and matching the data for males from 1980 to the adult wave of 2004 our sample consists of 2749 observations.

The BCS has been used in several empirical studies on the link between personality traits and individual 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.

Outcome Variable

The information on individual unemployment to be predicted with the variables measured at the age of 10 is taken from the 2004 wave of the BCS, i.e. when the study members were 34 years old. Similar information on unemployment is contained in the consecutive waves of 2008 and 2012. However, the attrition rate is around 18% relative to the 2004 wave. Due to this significant drop in the sample size, we refrained from predicting unemployment at even later stages of the life-cycle.

Moreover, we focus on male employees only, since for females at this particular stage of the life-cycle it is hard to distinguish between voluntary career breaks arising from family planning issues and involuntary ones. Predicting these frictional unemployment events due to voluntary career breaks is not in the focus of our study, in particular, if the goal is to optimize the predictive model with respect to the cost-minimal assignment into early intervention schemes. Employees who reported that they are: full-time or part-time employed and had a job in the last 4 years are coded as "Employed". Those who reported that they are currently unemployed and look for a job, receiving a job-seeker's allowance (JSA) or were unemployed at least once in the last 4 years are coded as "Unemployed". Thus, the outcome variable represents in this context a 4-year period prevalence of unemployment for male at 34 years of age. Study members who reported they are self-employed, 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. The unemployment rate of the sample is 9.79%.

[4]If more than 10 answers were missing, the observations were discarded from the data set, i.e. at most 10 missing values were estimated for each study member. The R package Amelia was used for the imputation.

Survey Items

There are several questionnaires in the BCS measuring behavior and noncognitive skills of the study members. There are three sources of measurements: questionnaires filled out by (1) the study members, (2) their mothers and (3) their teachers. Answers to these survey items are traditionally collected into indices (scales) representing 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 in the Appendix A.1.1, Tables 12-16.

At the age of 10, 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. The Locus of Control Scale is supposed to capture how much one believes that he is in control of his life (Rotter, 1966). Higher scores represent an internalizer, a person who thinks that he has control over the outcomes in his life and that he can influence them by his own actions.

In the 1980 wave, mothers of the study members were asked to answer a set of survey items about the behavior of their 10 years old children. In total, there were 38 items which are listed in Table 14. Based on these items, two well known instruments - the Rutter Behavior Scale (Rutter et al., 1970) and the Conners Rating Scale (Conners, 1969; Goyette et al., 1978) - 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 apply cluster analysis in a similar spirit as Butler et al. (1982) and Prevoo and ter Weel (2015) and go beyond the two standard measures yielding a more detailed structure of noncognitive skills from the mother's questionnaire.

Teachers of the study members also answered several survey items about the behavior of the children. In total, the teachers answered 53 items listed in Tables 15 and 16. As in the case of the parents, there are teacher's versions of the Rutter Behavior Scale (Rutter, 1967) and the Conners Rating Scale (Conners, 1969; Goyette et al., 1978). The cluster analysis was applied to identify the scales based on the teacher's questionnaire.

As controls, we use variables listed in Table 11. Social class of the family captures home

environment and enters every model as 5 dummies (the first social class level is left out). Cognitive skills are captured in two "Ability" variables based on a Friendly Math Test and Edinburgh Reading Test.

Cluster Analysis

The results of the cluster analysis are captured in Figures 5 and 6 for the assessments by the mothers and the teachers, respectively. We use hierarchical clustering based on Ward's method, which starts with single item 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 5, the procedure can be illustrated as follows. The algorithm starts with 38 clusters, i.e. each item 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 items. The order of merging steps is represented in Figures 5 and 6 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 (Belsis et al., 2014). 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 "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 for the mother-rated items since going from 8 to 7 clusters is relatively costly, see Figure 8 in the Appendix A.2. For the teacher-rated items, a jump occurs at merging 12 to 11 clusters. The clusters are represented by red boxes in Figures 5 and 6 and got the following labels collected in Table 4.

In total, we have 22 indices for the analysis: (a) 8 clusters from the mother-rated items, (b) 12 clusters from the teacher-rated items, (c) self-esteem index and the locus of control index.

Figure 5: Clusters of Mother-rated Items

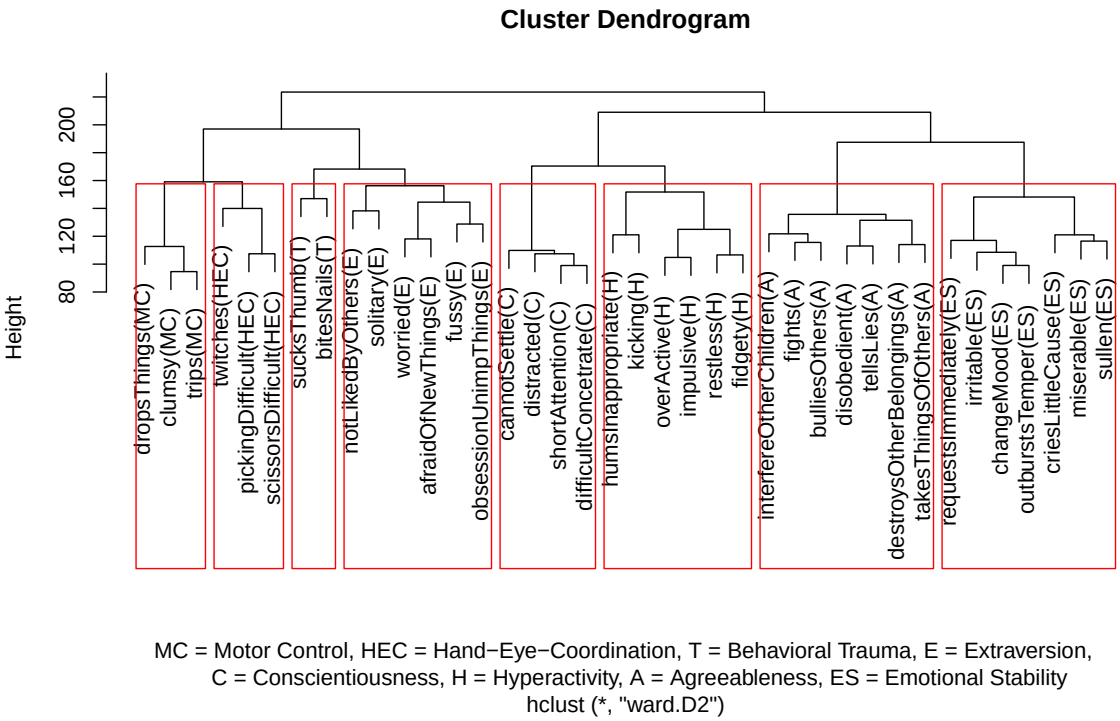


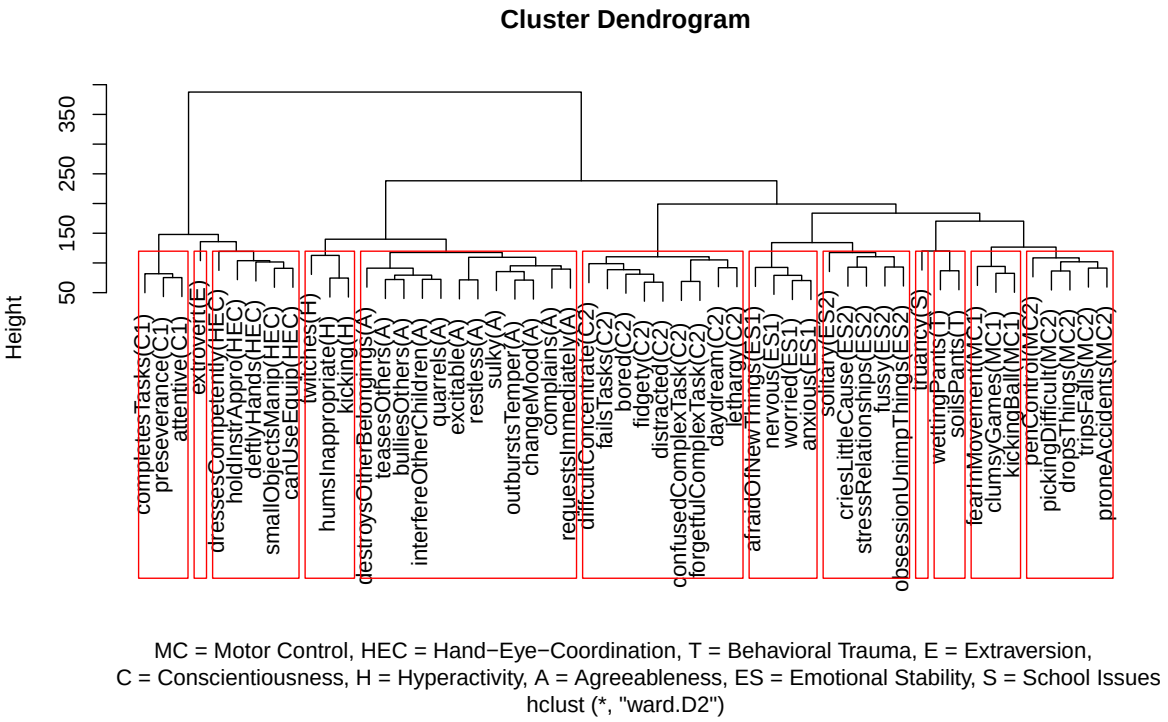
Table 4: Index Labels

Mother-rated items (8 clusters)	Teacher-rated items (12 clusters)
MC_M = Motor Control,	$MC1_T$ = Motor Control - fast movement,
HEC_M = Hand-Eye-Coordination,	$MC2_T$ = Motor Control - coordination,
T_M = Behavioral Trauma,	HEC_T = Hand-Eye-Coordination,
E_M = Extraversion,	T_T = Behavioral Trauma,
C_M = Conscientiousness,	E_T = Extraversion,
H_M = Hyperactivity,	$C1_T$ = Conscientiousness - (+) formulation,
A_M = Agreeableness,	$C2_T$ = Conscientiousness - (-) formulation,
ES_M = Emotional Stability.	H_T = Hyperactivity,
	A_T = Agreeableness,
	$ES1_T$ = Emotional Stability - inner,
	$ES2_T$ = Emotional Stability - outer,
	S_T = School Behavior.

Subscripts M an T indicate survey items from the mother's and teacher's questionnaire respectively.

The four personality traits coming from Big Five (E, C, A and ES) are very similar to the clusters obtained in Prevoo and ter Weel (2015) regarding the mother's survey items. Blanden et al. (2007) point out that the relevant variables in the BCS70 are rather close to the variables of Five Factor model. For a preliminary analysis, Tables 17 and 18 contain pairwise correlations

Figure 6: Clusters of Teacher-rated Items



between the 22 equally weighted indices separated for mothers and teachers respectively and two *Ability* indices. The results show that there is a slightly positive correlation between *Locus of Control* and both *Ability* indices indicating a potential that there might be a channel between noncognitive and cognitive skills.

By using information from both mother's and teacher's questionnaires, several noncognitive skills are measured from two different sources. This allows us to look into a question whether the source of information plays an important role for predictive purposes. In other words, who should be asked to evaluate the children in order to predict whether they would profit from an early MTP. Low levels of pairwise correlations between the personality traits measured in mother's and teacher's questionnaires in Table 19 indicate that including the views of the parents and teachers is not leading to redundant measurements but to additional information for the model.

5 Empirical Results

5.1 Unemployment Classifications

In the following, we will evaluate the different model specifications and index constructions solely in terms of their statistical and economic predictive performance. We use the multi-step procedure described in Figure 2. The overall data set was split randomly into training, validation and test data sets with the shares close to 60%, 20% and 20%, see Table 5. We apply a specific splitting procedure which divides the three subsets such that the unemployment rate in each subset is close to the overall unemployment rate, i.e. in the first step, the training set is drawn such that the unemployment rate is close to 9.79% in approximately 60% of the data. In the second step, the test set is drawn such that its unemployment rate is close to 9.79% in approximately 20% of the data. The validation set consists of the remaining observations not selected for the training and the test set. This yields an imprecise 50:50 split between validation and test set meanwhile assuring that the unemployment rate is as close to 9.79% as possible in each subset. Our splitting procedure guarantees that training set and validation set are representative in terms of the unemployment rate, so that choice of the tuning parameter τ selected on the observations of the validation set is based on a representative sample. After applying this splitting procedure, our test sample consists of 549 observations with an unemployment rate of 9.65%, for which we predict their employment status 24 years ahead.

Table 5: Training, Test and Validation Samples

	Training Set	Validation Set	Test Set	Total
Sample Size	1649	551	549	2749
Unemployment Rate (%)	9.76	9.98	9.65	9.79

Male workers at age 34, source: BCS wave 2004.

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 λ chosen by a 10-fold cross validation based on a margin-based loss function which approximates the upper bound of misclassification (Lin, 2004). Table 6 contains various classification measures for our nine different model variants where τ was tuned based on Youden's statistics. The results show that models with high overall accuracy

(over 80%) have a tendency to predict majority of cases as employed translating into low true positives (low sensitivity) and a high false negative rate. The results show the typical trade-off for binary outcome models, i.e. when the model detects many true positives, the number of false positives increases as well (making an unemployment program less efficient as many people take it unnecessarily) and when the model detects many true negatives, the number of false negatives increases (i.e. it misses many of truly unemployed people increasing the social cost of unemployment).

Table 6: Out-of-Sample Classification Measures: Statistical Tuning

	M_0	onlyEW	onlyPCA	onlyGL	post only GL	EW	PCA	GL	post GL
Accuracy	0.65	0.53	0.51	0.65	0.81	0.61	0.60	0.62	**0.82**
Sensitivity	0.42	0.47	**0.49**	0.43	0.15	0.42	0.47	**0.49**	0.17
Specificity	0.68	0.54	0.52	0.67	0.88	0.64	0.61	0.64	**0.89**
PPV	0.12	0.10	0.10	0.12	0.12	0.11	0.11	0.13	**0.14**
Costs	204.98	203.94	201.47	200.69	249.34	208.61	197.39	**190.74**	243.96

Classification with τ chosen based on Youden's statistic evaluated on the validation set. The first four rows contain the standard classification measures. The last row contains the expected economic costs for the classifiers for $\theta = 0.1$, $c = 100$ GBP, $u = 2800$ GBP (average annual JSA payments per JSA claimant) and $\pi = 0.0979$, the corresponding unemployment rate in the 2004 BCS sample. Bold numbers indicate the best values.

In terms of prediction quality, the models solely based on noncognitive skill factors are comparable with the richer models including in addition conventional regressors such as socio-economic controls (social class) and cognitive ability, i.e. even without conventional economic predictors the correct classification is rather high and improves only slightly for most of the classifiers. As a matter of fact, for the group lasso case including additional predictors turns out to be counterproductive in terms of accuracy. Comparing the predictive performance with respect to the type of index construction, the group lasso based index unambiguously dominates the two more standard procedures of index construction. The comparison of the models based on the area under the curve (AUC) described in Section 3 also underlines the statistical in- and out-of-sample dominance of the models based on the group lasso target-oriented indices as displayed in Table 7. The model based on conventional regressors (M_0) yields good results in terms of overall accuracy. However, in terms of sensitivity the majority of models including noncognitive skills yield a better result, i.e. they predict better the truly unemployed making noncognitive skills an important predictor for unemployment.

For illustration purposes the last row of Table 6 contains the economic costs of unemployment implied by sorting based on a purely statistically trained program. For both groups of models,

Table 7: AUC for all Models, Male, 34Y sample

	M_0	onlyEW	onlyPCA	onlyGL	post onlyGL	EW	PCA	GL	post GL
In-sample (Training Set)	0.574	0.658	0.655	0.739	0.756	0.667	0.664	0.736	**0.760**
Out-of-sample (Test Set)	0.533	0.528	0.533	0.567	0.563	0.535	0.538	**0.576**	0.568

AUC calculated for each model's ROC curve. The bold figures represent the best values.

the models containing additional predictors and the models including noncognitive skill factors only, the two group lasso approaches reveal the largest potential to improve on the economic costs without optimizing the tuning parameter τ in this respect.

Table 8 contains various classification measures for our nine different model variants where τ was tuned regarding the expected assignment costs. It is not too surpring that prediction accuracy drops compared to the classifiers which are tuned w.r.t. to Youden's statistic. However, comparing the costs between the models with economically tuned threshold parameter τ in Table 8 and statistically tuned threshold parameter τ in Table 6, economic tuning yields a considerable 33% reduction in economic costs on average. In the chosen illustrative setting, the cost of the program is relatively low ($c = 100$ GBP) in comparison to the JSA ($u = 2800$ GBP), therefore all the models predict many cases as becoming unemployed (high sensitivity and low specificity) as it is economically more efficient to send people to the training then to pay high unemployment benefits even though many participants actually did not need the training. As a consequence, the statistical measures will now crucially depend on the parameters of the expected costs function.

Table 8: Out-of-Sample Classification Measures: Economic Tuning

	M_0	onlyEW	onlyPCA	onlyGL	post only GL	EW	PCA	GL	post GL
Accuracy	0.12	**0.39**	0.35	0.13	0.12	0.24	0.26	0.15	0.11
Sensitivity	0.94	0.70	0.72	0.96	**0.98**	0.89	0.91	0.96	0.96
Specificity	0.03	**0.35**	0.31	0.04	0.02	0.17	0.19	0.06	0.02
PPV	0.09	0.10	0.10	0.10	0.10	0.10	**0.11**	0.10	0.10
Costs	138.26	167.06	166.59	132.52	**129.68**	138.75	132.46	131.06	134.15

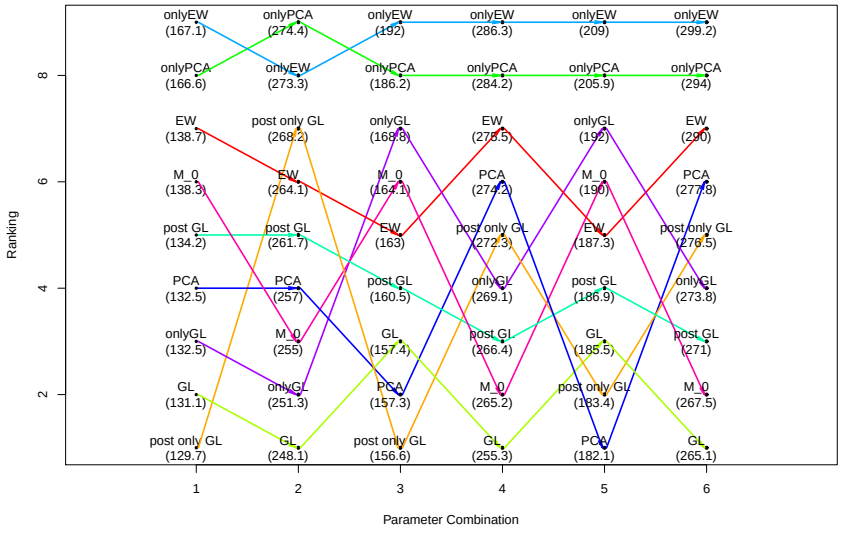
Classification based on cost-minimal τ evaluated on the validation set. The first four rows contain the standard classification measures. The last row contains the expected economic costs for the classifiers for $\theta = 0.1$, $c = 100$ GBP, $u = 2800$ GBP (average annual JSA payments per JSA claimant) and $\pi = 0.0979$ the corresponding unemployment rate in the 2004 BCS sample. Bold numbers indicate the best values.

The group lasso based models show clear cost reductions as Table 8 suggests. To get a more robust picture, several configurations of the following parameters were run to analyze the potential of each method in order to reduce economic costs of employment training:

- expected percentage reduction in unemployment benefit payments: $\varphi = \{0.1, 0.2, 0.3\}$,
- per capita cost of intervention program: $c = \{100, 250\}$ GBP,
- average annual JSA payment per JSA claimant: $u = 2800$ GBP,
- unemployment rate in the 2004 BCS sample (incidence): $\pi = 0.0979$.

The parameter φ is chosen based on the estimates of O'Connell (2002) who showed that the placement rate of low-skilled males in Ireland increased after a participation in an active labor market program by 9-30 percentage points. The values listed above yield 6 different combinations which were analyzed in terms of economic costs. Figure 7 captures the ranks over all parameter combinations for the 34 year old males. The group lasso model (GL) shows the lowest most stable expected costs. Table 9 confirms further the results of the graphical analysis. The median and mean ranks of the GL model are far ahead of the other methods underlying the potential of regularization in terms of yielding economically efficient predictive models over a range of parameter combinations.

Figure 7: Ranks - Economically Trained Models, Male 34Y Sample



Results for 6 parameter combinations. $\theta = 0.1$ in combinations 1 and 2 $\theta = 0.2$ in 3 and 4, $\theta = 0.3$ in 5 and 6. $c = 100$ in 1, 3 and 5 and $c = 250$ in 2, 4 and 6. $u = 2800$ and $\pi = 0.0979$.

5.2 Selected Skill Factors

In comparison to a pure machine learning approach or a kitchen-sink regression our group lasso approach yields interpretable prediction equations which may provide valuable information

Table 9: Ranks for the Economic Costs, Economically Trained, Male, 34Y sample

	M_0	onlyEW	onlyPCA	onlyGL	post onlyGL	EW	PCA	GL	post GL
Mean	4.17	8.83	8.17	4.50	3.50	6.17	3.67	1.83	4.00
Median	4.50	9.00	8.00	4.00	3.50	6.50	3.50	1.50	4.00

Rank counts based on the economic costs across all parameter combinations.

which skill factors are highly predictive and which may then be relevant for the design of appropriate intervention programs. Table 10 contains the estimates of M_0, *onlyPCA*, *onlyGL*, *PCA* and *GL* representing the main combinations of model covariate sets and types of weighting schemes. The models with equal indices were left out as their coefficient estimates are similar to PCA models. In the following, the main focus will be on the group lasso results as they perform economically well. The regularized models selected 16 indices out of 22 indicating that noncognitive skills play an important role for predictions and speaking against the models picking only one specific statistically significant noncognitive skill as in Prevoo and ter Weel (2015). Except of the extraversion (E_M) index, the signs of the covariates are the same in regularized and unregularized models. Interestingly, the regularized models do not include social background, *SocialClass*, in the model suggesting that this typically considered unemployment channel would not be helpful for prediction. In other words, social background has no predictive discriminatory effect for unemployment as well as the reading ability, *AbilityRead*. The typical predictor for the unemployment, mathematical ability (*AbilityMath*), was picked up by the group lasso with a negative effect on the predicted probability of unemployment similar to Feinstein (2000).

Looking into the Big Five indices, *Agreeableness* decreases the probability of the 4-year prevalence of being unemployed as it would be expected from people who get along well with others. The same logic holds for *Emotional Stability* and *Extraversion* which are in general highly appreciated noncognitive skills in the society and therefore expected to have a positive effect on employment. Going beyond Big Five, *Hand Eye Coordination (HEC)* measured at the age of 10 predicts lower unemployment probability. This could mean that a child with better developed fine physical coordination is at a higher cognitive level which is then indirectly measured through the *HEC* index. A survey devoted to studies which analyze eye and reaching movements as measures of real-time cognitive processing can be found in Spivey et al. (2008). Conversely, *Motor Control* increases the probability of unemployment. By a thorough look at the survey items, the motor control indices could be proxies of competitiveness and self-confidence

Table 10: Unemployment Equation Estimates, Sample 34Y

	M_0	onlyPCA	onlyGL	PCA	GL
Intercept	-1.9599^{***}	-2.3424^{***}	-2.2867	-2.0996^{***}	-2.2614
Noncognitive skills indices based on self-reported items					
SE		0.1098	0.0589	0.1083	0.0741
LC		-0.0738	-0.1274	-0.0309	-0.0774
Noncognitive skills indices based on mother's assessment					
MC_M		-0.0197	—	-0.0295	—
HEC_M		0.0398	—	0.0449	—
C_M		0.2772^{*}	0.1677	0.3100^{**}	0.1863
H_M		-0.0082	—	0.0018	—
A_M		-0.2390^{*}	-0.165	-0.2352^{*}	-0.1541
ES_M		-0.0750	-0.0765	-0.0757	-0.0762
TR_M		-0.0557	—	-0.0552	—
E_M		0.0218	-0.0022	0.0312	-0.0002
Noncognitive skills indices based on teacher's assessment					
$MC1_T$		0.2218	0.0399	0.2190	0.0291
$MC2_T$		0.1412	0.0436	0.1307	0.0423
HEC_T		-0.1405	-0.0447	-0.1203	-0.0358
$C1_T$		0.2708	—	0.2778	—
$C2_T$		-0.5155^{**}	-0.2458	-0.4618^{*}	-0.1742
H_T		0.0760	0.0494	0.0797	0.0503
A_T		-0.1677	-0.0783	-0.1794	-0.0824
$ES1_T$		0.0717	—	0.0888	—
$ES2_T$		-0.0063	-0.01	-0.0214	-0.0168
TR_T		0.0681	0.0246	0.0705	0.0237
E_T		-0.2642^{*}	-0.084	-0.2702^{*}	-0.0768
S_T		0.0421	0.0195	0.0413	0.0182
Control variables					
SocialClass10Y2	-0.3082			-0.3211	—
SocialClass10Y3	-0.2814			-0.2682	—
SocialClass10Y4	-0.2756			-0.1942	—
SocialClass10Y5	-0.1398			-0.0954	—
SocialClass10Y6	-0.1458			0.0734	—
AbilityMath10Y	-0.2640^{*}			-0.2586	-0.1362
AbilityRead10Y	-0.0112			0.0868	—

ML logit estimates of the unemployment equation. Dependent variable: employed = 0, unemployed = 1. Models M_0, *onlyPCA* and *PCA* are estimated by standard ML. The stars represent the following p-values: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Models *onlyGL* and *GL* are estimated by the regularized ML. The regularization parameter λ is chosen by the 10-fold cross-validation. The index coefficients are identified by the sum up constraint $\sum_j \hat{\omega}_j^I = 1$. See Appendix B for more details. Currently there is no standardized significance test for regularized coefficients allowing to assess the statistical significance. Subscripts of the NCS indices indicate whether it is built based on mother-rated questions (M) or teacher-rated questions (T). — means that lasso excluded the variable from the model.

as indirectly measured by focusing on the coordination during games and what impression a child leaves when performing a focused movement. Piek et al. (2006) study an impact of gross and fine motor skills in children and adolescents on their perceived self-worth concluding that poor motor control negatively impacts self-perception. In this light, both *Motor Control* indices

could be then predictors for a higher probability of the 4-year prevalence of being unemployed as more self-confident people are not afraid of short-term unemployment with an outlook of getting a position they want as it would be predicted by the Expectancy-Value Theory which is concisely outlined in Vansteenkiste et al. (2005). This would also justify the logic behind the positive prediction coefficient for *Self-Esteem*. Children with higher level of *Hyperactivity* and signs of *Trauma* have a higher predicted probability of being unemployed as they could have problems with concentration leading to difficulties on the labor market as descriptively analyzed in a phone survey by Biederman and Faraone (2006). Last predictor is related to *School Issues*, in particular to truancy. A truant child has a higher probability of being unemployed as it is probably a first sign of a potential dropout leading to a lower level of education with its consequences for the labor market.

6 Conclusion

This paper takes a closer look on the predictive power of non-cognitive skills for unemployment 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 and applying several novel steps in the modeling procedure.

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 target-oriented 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 be persistently outperformed by standard approaches using equally weighted indices or PCA based indices, especially in terms of economic costs. In comparison to other black box classification techniques, our predictive model is easily interpretable and can contribute to a discussion about important predictors of unemployment and help to design selection processes at an early stage of life cycle.

We also investigated two different perspectives for the optimal classification: statistical and economic one. Meanwhile statistical tuning of classification models is the dominant approach

in the machine learning literature, ignoring economic costs in applied economic research can lead to inefficient policy recommendations in the larger cost and benefit perspective. The idea of focusing on economic tuning of classification models instead of the statistical tuning can be easily extended to other classification techniques such as support vector machines or decision trees yielding potentially new interesting results.

Our empirical findings are based on data from the British Cohort Study (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 years later is very ambitious. Recognizing the fact that the survey items were not even designed for a purpose of individual unemployment prediction, the predictive quality of the factors is astonishing and leading to reduction of economic costs if the group lasso indices are used.

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 10-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 could 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.

Belsis, P., A. Koutoumanos, and C. Sgouropoulou (2014): "PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development," Requirements engineering, 19, 213–225.

Biederman, J. and S. V. Faraone (2006): "The effects of attention-deficit/hyperactivity disorder on employment and household income," Medscape General Medicine, 8, 12.

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.

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.

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," American 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.

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

Fawcett, T. (2006): "An introduction to ROC analysis," Pattern Recognition Letters, 27, 861 – 874, rOC Analysis in Pattern Recognition.

Feinstein, L. (2000): "The relative economic importance of academic, psychological and behavioural attributes developed on childhood," CEPDP (443), Centre for Economic Performance, London School of Economics and Political Science.

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

Goyette, C. H., C. K. Conners, and R. F. Ulrich (1978): "Normative data on Revised Conners Parent and Teacher Rating Scales," Journal of Abnormal Child Psychology, 6, 221–236.

Hastie, T., R. Tibshirani, and M. Wainright (2015): Statistical Learning with Sparisty The Lasso and Generalizations, Monographs on Statistics and Probability, CRC Press.

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.

Lin, Y. (2004): "A note on margin-based loss functions in classification," Statistics & Probability Letters, 68, 73 – 82.

Meier, L., S. van de Geer, and P. Bühlmann (2008): "The group Lasso for logistic regression," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70, 53 – 71.

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

Metz, C. E. (1978): "Basic principles of ROC analysis," Seminars in Nuclear Medicine, 8, 283 – 298.

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.

O'Connell, P. J. (2002): "Are They Working? Market Orientation and the Effectiveness of Active Labour-Market Programmes in Ireland," European Sociological Review, 18, 65–83.

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.

Piek, J. P., G. B. Baynam, and N. C. Barrett (2006): "The relationship between fine and gross motor ability, self-perceptions and self-worth in children and adolescents," Human Movement Science, 25, 65 – 75, approaches to Sensory-Motor Development in Infants and Children.

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.

Shmueli, G. (2010): "To Explain or to Predict?" Statististical Science, 25, 289–310.

Spivey, M., D. Richardson, and R. Dale (2008): Oxford handbook of human action, New York: Oxford University Press, chap. 12: The movement of eye and hand as a window into language and cognition, 225–249.

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.

Vansteenkiste, M., W. Lens, H. De Witte, and N. T. Feather (2005): "Understanding unemployed people's job search behaviour, unemployment experience and well-being: A comparison of expectancy-value theory and self-determination theory," British Journal of Social Psychology, 44, 269–287.

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 11: 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
AbilityMath10Y	standardized score of a "Friendly Math Test" taken at the age of 10
AbilityRead10Y	standardized score of an "Edinburgh Reading Test" taken at the age of 10

Table 12: 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 13: 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 14: Non-cognitive Skills Measures: Mother-rated Items (hierarchical 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 - recoded as (100 - answer) to capture hyperactivity	
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 - recoded as (100 - answer) to capture behavioral trauma	
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 15: Non-cognitive Skills Measures: Teacher-rated Items (hierarchical clustering)

Motor Control 1 (MC1) Items - recoded as (48 - answer) to capture motor control	
1	Child is noticeably clumsy in formal or informal games
2	Child finds it difficult to kick a ball forward
3	Child is fearful in movements, requires much encouragement to move faster
Motor Control 2 (MC2) Items - recoded as (48 - answer) to capture motor control	
4	Child trips or falls easily or bumps into objects or other children
5	Child shows difficulty when picking up small objects
6	Child drops things which are being carried
7	Child shows inadequate control when handlling a pencil or paint brush
8	Child experiences classroom or playground accidents
Hand Eye Coordination (HEC) Items	
9	Child works deftly with his or her hands
10	Child dresses and undresses competently (e.g. for P.E)
11	Child manipulates small objects easily with his/her hands
12	Child can use scissors and similar manipulative equipment competently
13	Child holds writing and drawing instruments appropriately
Conscientiousness 1 (C1) Items	
14	Child shows preservance, persists with difficult or routine work
15	Child pays attention to what is being explained in class
16	Child completes tasks which are started
Conscientiousness 2 (C2) Items - recoded as (48 - answer) to capture conscientiousness	
17	Child is given to daydreaming
18	Child cannot concentrate on any particular task, even though the child may return to it frequently
19	Child becomes bored during class
20	Child becomes confused or hesitant when given a complex task
21	Child is squirmy or fidgety
22	Child is easily distracted
23	Child is forgetful when given a complex task
24	Child shows lethargic and listless behaviour
25	Child fails to finish things he/she starts
Hyperactivity (H) Items	
26	Child hums or makes other odd noises at inappropriate times
27	Child is given to rhythmic tapping or kicking during class
28	Child has twitches mannerisms or tics of the face and body
Agreeableness (A) Items - recoded as (48 - answer) to capture agreeableness	
29	Child complains about things
30	Child displays outbursts of temper, explosive or unpredictable behaviour
31	Child teases other children to excess
32	Child interferes with the activity of other children
33	Child changes mood quickly and drastically
34	Child is excitable, impulsive
35	Child shows restless or over-active behaviour
36	Child quarrels with other children
37	Child destroys own or other children's belongings
38	Child bullies other children
39	Child is sullen or sulky
40	Child's requests must be met immediately, easily frustrated

Answers coded on a scale from 1 to 47 where 1 = "not at all" and 47 = "a great deal".

Table 16: Non-cognitive Skills Measures: Teacher-rated Items (hierarchical clustering) - cont.

Emotional Stability 1 (ES1) Items - recoded as (48 - answer) to capture emotional stability	
41	Child is fearful or afraid of new things or situations
42	Child behaves nervously
43	Child is worried and anxious about many things
44	Child is anxious (opposite: unworried)
Emotional Stability 2 (ES2) Items - recoded as (48 - answer) to capture emotional stability	
45	Child cries for little cause
46	Child is fussy or over particular
47	In relations with others child appears to be miserable, unhappy, tearful or distressed
48	Child becomes obsessional about unimportant things
49	Child tends to do things on his own
Behavioral Trauma (TR) Items	
50	Child has problems with wetting pants during class
51	Child has problems of soiling pants during class
Extraversion (E) Items	
52	Child is an extrovert (opposite: introvert)
School Issues (S) Items	
53	Child truants from school

Answers coded on a scale from 1 to 47 where 1 = "not at all" and 47 = "a great deal".

Table 17: Correlation Matrix for Equally Weighted Indexes (Mother's Questionnaire) based on the Hierarchical Clustering for Male - 34Y

	MC_M	HEC_M	C_M	H_M	A_M	ES_M	TR_M	E_M	SE	LC	A-M
HEC_M	0.46										
C_M	0.35	0.28									
H_M	-0.36	-0.26	-0.56								
A_M	0.34	0.35	0.47	-0.48							
ES_M	0.32	0.29	0.42	-0.52	0.59						
TR_M	-0.17	-0.16	-0.14	0.15	-0.20	-0.16					
E_M	0.29	0.30	0.26	-0.33	0.30	0.51	-0.15				
SE	0.09	0.03	0.21	-0.13	0.14	0.14	-0.06	0.08			
LC	0.09	0.08	0.25	-0.13	0.19	0.14	-0.07	0.09	0.41		
A-M	0.09	0.10	0.34	-0.16	0.21	0.16	-0.06	0.08	0.26	0.46	
A-R	0.09	0.09	0.32	-0.14	0.20	0.15	-0.05	0.06	0.25	0.48	0.75

A-M = Ability Math, A-R = Ability Read, subscript M indicates questions in the Mother's Questionnaire.

Table 18: Correlation Matrix for Equally Weighted Indeces (Teacher's Questionnaire) based on Hierarchical Clustering for Male - 34Y

	$C1_T$	$C2_T$	$ES1_T$	$ES2_T$	TR_T	A_T	$MC1_T$	$MC2_T$	HEC_T	S_T	E_T	H_T	SE	LC	A-M
$C2_T$	0.83														
$ES1_T$	0.32	0.47													
$ES2_T$	0.28	0.43	0.62												
TR_T	-0.06	-0.09	-0.08	-0.17											
A_T	0.48	0.57	0.26	0.47	-0.11										
$MC1_T$	0.29	0.41	0.42	0.53	-0.15	0.25									
$MC2_T$	0.42	0.54	0.37	0.49	-0.22	0.49	0.60								
HEC_T	0.46	0.44	0.30	0.30	-0.17	0.27	0.45	0.59							
S_T	-0.16	-0.20	-0.12	-0.20	0.36	-0.25	-0.16	-0.25	-0.18						
E_T	0.07	0.11	0.46	0.41	-0.05	-0.19	0.35	0.11	0.18	-0.03					
H_T	-0.41	-0.50	-0.27	-0.39	0.16	-0.63	-0.30	-0.55	-0.32	0.27	0.07				
SE	0.24	0.26	0.15	0.20	-0.03	0.21	0.14	0.17	0.16	-0.07	0.08	-0.16			
LC	0.35	0.37	0.19	0.17	-0.01	0.19	0.13	0.20	0.20	-0.09	0.09	-0.15	0.41		
A-M	0.47	0.50	0.25	0.14	-0.04	0.22	0.16	0.25	0.28	-0.11	0.08	-0.19	0.26	0.46	
A-R	0.44	0.49	0.23	0.11	-0.03	0.22	0.13	0.26	0.28	-0.10	0.07	-0.18	0.25	0.48	0.75

A-M = Ability Math, A-R = Ability Read, subscript T indicates questions in the Teacher's Questionnaire.

Table 19: Comparing Mother's and Teacher's Equally Weighted Indexes based on Hierarchical Clustering, Male 34Y

	MC_M	$MC1_T$	$MC2_T$
MC_M	1.00	0.16	0.19
$MC1_T$	0.16	1.00	0.60
$MC2_T$	0.19	0.60	1.00

	HEC_M	HEC_T
HEC_M	1.00	0.08
HEC_T	0.08	1.00

	C_M	$C1_T$	$C2_T$
C_M	1.00	0.38	0.38
$C1_T$	0.38	1.00	0.83
$C2_T$	0.38	0.83	1.00

	H_M	H_T
H_M	1.00	0.16
H_T	0.16	1.00

	A_M	A_T
A_M	1.00	0.24
A_T	0.24	1.00

	ES_M	$ES1_T$	$ES2_T$
ES_M	1.00	0.10	0.11
$ES1_T$	0.10	1.00	0.62
$ES2_T$	0.11	0.62	1.00

	TR_M	TR_T
TR_M	1.00	0.03
TR_T	0.03	1.00

	E_M	E_T
E_M	1.00	0.16
E_T	0.16	1.00

Table 20: Means of Model Covariates

	All		Train		Validation		Test	
	Empl	Unempl	Empl	Unempl	Empl	Unempl	Empl	Unempl
SocialClass10Y1	0.09	0.13	0.08	0.12	0.09	0.18	0.11	0.11
SocialClass10Y2	0.28	0.27	0.27	0.26	0.30	0.29	0.28	0.26
SocialClass10Y3	0.30	0.31	0.31	0.30	0.30	0.27	0.28	0.38
SocialClass10Y4	0.17	0.14	0.18	0.16	0.16	0.09	0.16	0.15
SocialClass10Y5	0.14	0.13	0.14	0.14	0.13	0.15	0.15	0.09
SocialClass10Y6	0.02	0.02	0.03	0.02	0.02	0.02	0.02*	0.00*
AbilityMath10Y	0.26*	−0.02*	0.24*	−0.04*	0.31*	−0.07*	0.28	0.06
AbilityRead10Y	0.13*	−0.06*	0.13*	−0.08*	0.15*	−0.18*	0.13	0.11
eqMC_M	0.01	−0.03	0.00	−0.02	0.02	0.12	0.03	−0.23
eqHEC_M	−0.00	−0.00	−0.00	0.03	−0.03	0.07	0.03	−0.19
eqC_M	0.01	−0.06	−0.00	0.04	−0.00	−0.27	0.04	−0.12
eqH_M	−0.01	0.07	−0.01	0.05	0.03	0.18	−0.05	0.01
eqA_M	0.02*	−0.23*	0.03*	−0.23*	−0.00	−0.33	0.02	−0.14
eqES_M	0.00	−0.11	0.02	−0.14	−0.03	−0.06	0.00	−0.05
eqTR_M	0.01	−0.01	0.00	−0.03	0.04	−0.04	−0.00	0.07
eqE_M	−0.01	−0.04	0.00	−0.04	−0.04	−0.03	−0.03	−0.03
eqSE	0.04	−0.04	0.00	−0.02	0.07	−0.22	0.12	0.06
eqLC	0.02*	−0.21*	0.02*	−0.17*	0.01*	−0.44*	0.05	−0.11
eqC1_T	0.03*	−0.19*	0.02*	−0.19*	0.03*	−0.34*	0.08	−0.04
eqC2_T	0.03*	−0.28*	0.03*	−0.28*	0.02*	−0.53*	0.06	−0.02
eqES1_T	0.02*	−0.18*	0.01	−0.14	0.03*	−0.40*	0.04	−0.07
eqES2_T	0.04*	−0.23*	0.02*	−0.19*	0.03*	−0.37*	0.12	−0.22
eqTR_T	−0.02	0.10	−0.01	0.11	0.03	0.22	−0.08	−0.04
eqA_T	0.02*	−0.21*	0.03*	−0.24*	−0.04	−0.34	0.06	0.01
eqMC1_T	0.03	−0.10	0.00	−0.02	0.10*	−0.23*	0.05	−0.21
eqMC2_T	0.02*	−0.19*	0.01	−0.13	0.02*	−0.42*	0.04	−0.10
eqHEC_T	0.02*	−0.23*	0.01	−0.14	0.01*	−0.43*	0.05*	−0.28*
eqS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
eqE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
eqH_T	−0.03*	0.21*	−0.02*	0.20*	−0.05*	0.46*	−0.05	−0.01
pcaMC_M	0.01	−0.03	0.00	−0.02	0.02	0.13	0.04	−0.23
pcaHEC_M	−0.01	−0.02	−0.00	0.01	−0.03	0.05	0.01	−0.17
pcaC_M	0.00	−0.06	−0.00	0.04	−0.00	−0.27	0.04	−0.12
pcaH_M	−0.01	0.07	−0.01	0.06	0.04	0.19	−0.05	0.01
pcaA_M	0.02*	−0.23*	0.03*	−0.23*	−0.00	−0.32	0.02	−0.15
pcaES_M	0.01	−0.11	0.02	−0.14	−0.03	−0.08	0.01	−0.06
pcaTR_M	0.01	−0.01	0.00	−0.02	0.04	−0.04	−0.00	0.06
pcaE_M	−0.02	−0.02	0.00	−0.02	−0.06	−0.04	−0.03	−0.01
pcaSE	0.04	−0.03	0.00	−0.02	0.07	−0.22	0.12	0.11
pcaLC	0.03*	−0.21*	0.02	−0.15	0.01*	−0.50*	0.08	−0.11
pcaC1_T	0.03*	−0.19*	0.02*	−0.18*	0.03*	−0.34*	0.08	−0.04
pcaC2_T	0.03*	−0.28*	0.03*	−0.28*	0.02*	−0.52*	0.06	−0.02
pcaES1_T	0.02*	−0.18*	0.01	−0.14	0.03*	−0.40*	0.04	−0.06
pcaES2_T	0.04*	−0.22*	0.02*	−0.17*	0.03*	−0.37*	0.13	−0.20
pcaTR_T	−0.02	0.10	−0.01	0.11	0.03	0.22	−0.08	−0.04
pcaA_T	0.02*	−0.22*	0.03*	−0.24*	−0.04	−0.34	0.06	−0.00
pcaMC1_T	0.03	−0.10	0.00	−0.02	0.10*	−0.23*	0.05	−0.21
pcaMC2_T	0.02*	−0.18*	0.01	−0.11	0.02*	−0.44*	0.04	−0.10
pcaHEC_T	0.02*	−0.23*	0.02	−0.14	0.01*	−0.44*	0.05*	−0.30*
pcaS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
pcaE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
pcaH_T	−0.03*	0.21*	−0.02*	0.20*	−0.05*	0.46*	−0.05	−0.02

Stars indicate mean differences which are significant at 5% significance level.

Table 21: Means of Model Covariates - cont.

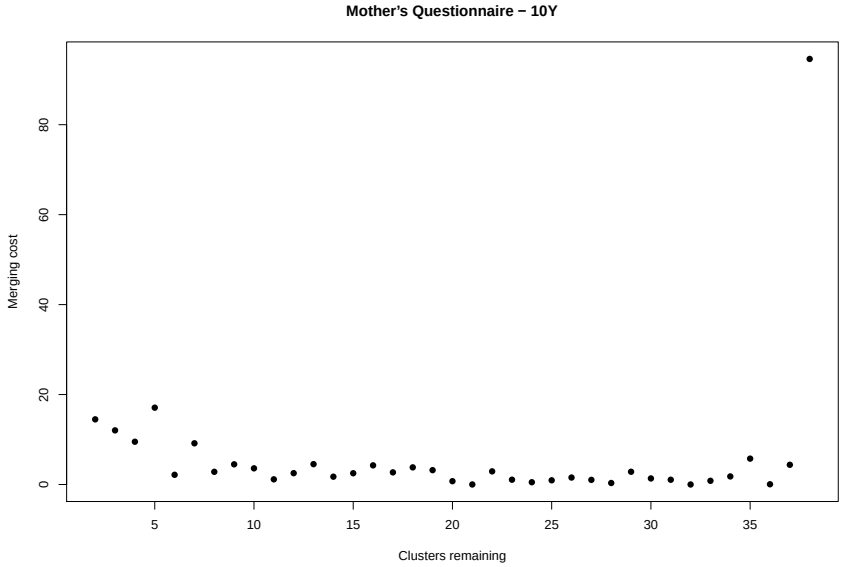
	All		Train		Validation		Test	
	Empl	Unempl	Empl	Unempl	Empl	Unempl	Empl	Unempl
XMC_M								
$XHEC_M$								
XC_M	−0.00	−0.01	−0.01	0.07	0.01	−0.21	0.01	−0.07
XH_M								
XA_M	0.02*	−0.16*	0.02*	−0.19*	0.01	−0.18	0.01	−0.06
XES_M	0.00*	−0.19*	0.03*	−0.25*	−0.06	−0.13	−0.01	−0.06
XTR_M								
XE_M	15.26*	−57.51*	7.86*	−72.64*	32.59	−44.34	20.11	−25.19
XSE	0.01*	0.77*	−0.10*	0.94*	0.23	0.57	0.11	0.45
XLC	0.03*	−0.14*	0.03*	−0.25*	−0.00	−0.01	0.07	0.06
$XC1_T$								
$XC2_T$	0.03*	−0.23*	0.03*	−0.23*	0.02*	−0.42*	0.04	−0.03
$XES1_T$								
$XES2_T$	0.03*	−0.24*	0.03*	−0.23*	0.03	−0.20	0.02	−0.29
XTR_T	−0.02	0.10	−0.01	0.12	0.02	0.17	−0.08	−0.04
XA_T	0.02*	−0.19*	0.02*	−0.22*	−0.01	−0.27	0.05	−0.02
$XMC1_T$	0.03	−0.07	0.00	−0.01	0.09*	−0.17*	0.05	−0.15
$XMC2_T$	−0.01	0.05	−0.03*	0.28*	0.02	−0.57	0.04	−0.00
$XHEC_T$	0.08*	−0.43*	0.08*	−0.71*	0.02	0.31	0.15	−0.34
XS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
XE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
XH_T	−0.03*	0.20*	−0.04*	0.33*	−0.05	0.30	0.00	−0.29

Group lasso indices using weights based on the C.8 specification. Stars indicate mean differences which are significant at 5% significance level.

A.2 Figures

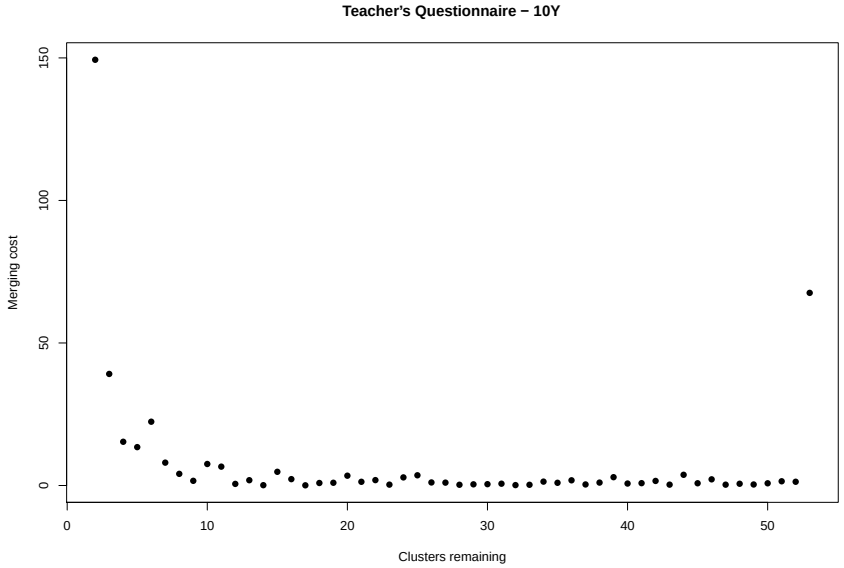
A.2.1 Cluster Analysis

Figure 8: Merging Costs of K clusters - Mother's Questionnaire



Each point represents the merging costs of merging the 'Clusters remaining + 1' to 'Clusters remaining'. The only exception is the point at the very right which represents the within-cluster variance when no merging occured.

Figure 9: Merging Costs of K clusters - Teacher's Questionnaire



Each point represents the merging costs of merging the 'Clusters remaining + 1' to 'Clusters remaining'. The only exception is the point at the very right which represents the within-cluster variance when no merging occured.

B Additional Material

B.1 Identification of Group Lasso Weights

By estimating the following logit model under a group lasso penalty

$$\begin{aligned} Y_i &= \Lambda(\alpha + \beta_1 C_{i1} + \beta_2 C_{i2} + \cdots + \beta_K C_{iK_C} + OtherControls_i'\delta) + \varepsilon_i, \\ &= \Lambda(\alpha + IC_i^C \gamma_C + OtherControls_i'\delta) + \varepsilon_i, \\ \text{where } IC_i^C &= \sum_{l=1}^{K_C} \omega_l^C C_{il}, \end{aligned}$$

we can derive situation specific weights (here for the Conscientiousness index) under the summing up to one constraint $\sum_l \hat{\omega}_l = 1$ as follows:

$$\begin{aligned} \hat{\beta}_l &= \widehat{\gamma_C \omega_l}, \\ \sum_{l=1}^{K_C} \hat{\beta}_l &= \sum_{l=1}^{K_C} \widehat{\gamma_C \omega_l}, \\ \sum_{l=1}^{K_C} \hat{\beta}_l &= \hat{\gamma}_C, \qquad \text{summming up to one constraint} \\ \Rightarrow \hat{\omega}_l &= \frac{\hat{\beta}_l}{\sum_l \hat{\beta}_l}. \end{aligned}$$