Essays in Applied Econometrics and Behavioral Economics

Inaugural-Dissertation
zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von

Thomas Deckers

aus Bonn

Bonn 2014

Dekan:	Prof. Dr. Klaus Sandmann
Erstreferent:	Prof. Dr. Armin Falk
Zweitreferent:	Prof. Dr. Thomas Dohmen

Tag der mündlichen Prüfung: 18.08.2014

Acknowledgments

I would not have been able to complete this thesis without the help of numerous people to which I want to express my deepest gratitude.

First of all to my supervisor Armin Falk: You have been an enormous source of inspiration for my work. I could always count on your advice when I needed it, be it in scientific or "general" questions. It has always been very motivating talking to you or working together with you, but even more importantly, it has also been a lot of fun. I have learned heaps from you which I benefit from even outside the scientific world. Thank you very much for being my supervisor for the last four years!

To my second supervisor Thomas Dohmen: Your initials, your hometown and the city you have studied in made you an obvious choice for being my second supervisor. But definitely more importantly, I benefited a lot from having a supervisor who also has a passion for econometrics and with whom I could engage in exciting discussions about the econometric parts of this thesis. And, of course, the same as for Armin applies to you as well: Working with you was certainly always fun. Thank you very much!

To my co-authors Anke Becker, Fabian Kosse, Christoph Hanck and Hannah Schildberg-Hörisch: You are great people and I enjoyed working together with you a lot. Moreover, you are also friends which I reckon is the best you can say about your co-authors. Thank you so much!

To the BGSE, especially to Silke Kinzig, Pamela Mertens and Urs Schweizer for making the BGSE such a nice place to study and work at and for being supportive in any kinds of questions. Also to the whole team of the institute for applied micro which was a great team to work in. Especially to Birgit Jendrock and Stephanie Sauter who supported me in numerous ways, from getting signatures to cutting marble lanes. I am also grateful to Sebastian Kube, Matthias Wibral, Steffen Altmann, Hans-Martin von Gaudecker, Pia

Pinger and Katarina Kuss for helpful discussions.

To the whole 2008/2009 BGSE cohort with special thanks to the best office mate Harun: All of you have made the time at Bonn University to an unforgettable experience and of course we have been the best year!

Last but not least, to Meike, Mats and Luk: You have certainly made clear that there is more important stuff to do than writing this thesis. It was always good to have that perspective and good to have all of you. And finally to family for supporting me in whatever I do. Thanks to all of you!

Contents

List of Figures

List of Tables

Introduction

This thesis is essentially divided into two separate parts. The first part deals with a problem frequently encountered in applied econometrics, namely variable selection in long regressions, i.e., regressions with many candidate regressors. For this problem a new solution is proposed and carefully compared to existing solutions. The second part of this thesis is concerned with modelling heterogeneity between individuals in form of personality and preferences. In particular, it provides a thorough comparison of different concepts to model heterogeneity between individuals. Moreover, it is investigated how different environments influence individuals' life satisfaction and also their personalities.

PART I

In applied econometrics one often deals with the situation of performing large a number of hypothesis tests simultaneously. Usually, each of these hypotheses is rejected at some predefined significance level α, implying that for each test the probability of committing a type I error, i.e., to falsely reject the hypothesis, is equal to α. However, given that a large number of tests are performed simultaneously, the data are given many chances of falsely selecting a hypothesis. This problem is frequently referred to as 'multiple testing problem'. The approach to overcome this problem that is investigated in Chapter 1 of this dissertation is to limit the ratio of erroneously rejected hypotheses to the total number of rejections in expected value. The expected value of this ratio is called the False Discovery Rate (FDR). One needs to choose a limit for the FDR, comparable to the choice of an individual significance level α.

Chapter 1 investigates whether controlling the FDR can be used as a model selection procedure in long regressions.[1] In particular, the properties of selecting variables using

[1]This chapter is based on joint work with Christoph Hanck. Our paper is forthcoming in the *Oxford*

different FDR controlling procedures are compared between each other and to other well-known model selection procedures such as Bayesian Model Averaging (BMA), PcGets / Autometrics or the Least Absolute Shrinkage and Selection Operator (Lasso) using extensive Monte Carlo simulations. It is found that, using conventional tuning parameters of the different model selection procedures, only the FDR controlling procedures consistently limit the FDR at predefined levels. This comes at the price of a somewhat lower power at comparable sizes than for BMA and PcGets / Autometrics. Finally, all methods are applied to the example of cross-sectional growth regressions. When controlling the FDR at very small levels few growth determinants beyond initial GDP are found, providing some evidence for conditional convergence.

PART II

In Chapter 2 different concepts of modelling heterogeneity between persons are compared to each other.[2] In particular, the degree of association between the concepts of psychology – the Big Five and the locus of control – and economic preferences is investigated making use of laboratory as well as representative data. Moreover, the explanatory power of the paradigms from psychology and economics are assessed separately as well as jointly. Overall, it is found that the concepts from psychology and economics are rather complements than supplements. This finding is particularly useful for applied researchers who seek to model heterogeneity of actions and resulting outcomes between individuals.

Chapter 3 studies how regional price levels influence individual life satisfaction.[3] The chapter employs a novel data set which makes use of over 7 million data points to construct local price indices on district level for Germany. Once individual and district heterogeneity are controlled for using the rich data of the German socio-economic panel study (SOEP), it can be shown that higher prices significantly reduce life satisfaction. Moreover, people seem to slightly overvalue prices in comparison to their nominal income. The results have policy implications in that they provide arguments in favor of regional indexation of government transfer payments or public sector salaries.

In Chapter 4 it is investigated how the socio-economic status (SES) of a family influences the development of a child's personality.[4] The facets of personality that are investigated

Bulletin of Economics and Statistics, see Deckers and Hanck (forthcoming).

[2]This chapter is based on joint work with Anke Becker, Thomas Dohmen, Armin Falk and Fabian Kosse. Our paper is published in the *Annual Review of Economics*, see Becker et al. (2012)

[3]This chapter is based on joint work with Armin Falk and Hannah Schildberg-Hörisch.

[4]This chapter is based on joint work with Armin Falk, Fabian Kosse and Hannah Schildberg-Hörisch.

encompass time preferences, risk preferences, and altruism that are important noncognitive skills, as well as crystallized, fluid, and overall IQ that represent cognitive skills. The results show that children from families with higher SES are on average more patient, less likely to be risk seeking, and score higher on IQ tests. Further analyses show that 20% to 40% of this effect can be explained by dimensions of a child's environment that are shown to differ by parental SES. The dependence of personality on a family's SES might offer an explanation for social immobility and it highlights the need for controlling for SES when analyzing the influence of personality on (later) life outcomes.

Chapter 1

Variable Selection in Cross-Section Regressions: Comparisons and Extensions

1.1 Introduction

Model uncertainty is one of the most frequent problems in applied econometric work. It describes the common situation in which the investigator is faced with a large number of candidate explanatory variables for some dependent variable of interest. Given the typical size of economic data sets, the investigator then needs a procedure to select the relevant determinants from the pool of candidate variables. That is, he or she needs to perform 'model selection', or equivalently, 'variable selection.' Unfortunately, there is no generally accepted, let alone efficient or most powerful, way to do so. A leading example of this situation is that of selecting growth determinants in a cross-section growth regression (e.g., Levine and Renelt, 1992).

A seemingly simple (and often adopted) solution to model selection is to test each variable j individually at some level α using appropriate p-values, rejecting hypothesis H_j if $\widehat{p}_j \leq \alpha$. But, given the large set of regressors, the data are given many chances of falsely rejecting ('multiplicity'). Hence, one is bound to erroneously declare some irrelevant variables to be significant using this approach.

The statistical literature has a long history of dealing with this issue of multiplicity (e.g., Holm, 1979). In this paper, we set out some of these so-called multiple testing procedures (MTPs), which we believe deserve more attention from applied econometricians. We compare these to other well-established model selection procedures (for details see below). More specifically, we follow Romano, Shaikh, and Wolf (2008b) who suggest to

employ MTPs to perform model selection when there is a large set of candidate variables, as is the case in for instance growth econometrics. These methods are fast and easy to implement. Concretely, we employ the widely-used procedures of Benjamini and Hochberg (1995) and Romano, Shaikh, and Wolf (2008a), which are reviewed in Section 1.2. For instance, all it takes to implement the Benjamini and Hochberg (1995) method is to compare the p-value of each variable in a regression including all candidate regressors to a specific cutoff $\alpha_j \leqslant \alpha$ rather than to compare each of them to α.

Commensurate with the practical relevance of model selection, a large number of proposals have been made to offer empirical researchers less arbitrary and more rigorous ways of selecting an empirical model. Prominent examples, reviewed in Section 1.2, include the 'two million regressions approach' of Sala-i-Martin (1997), Bayesian Model Averaging (e.g., Fernandez, Ley, and Steel, 2001, BMA), General-to-Specific/Autometrics (e.g., Hoover and Perez, 1999; Krolzig and Hendry, 2001) and the Lasso (Tibshirani, 1996; Zou, 2006). In practice, all methods require the user to specify some criterion, controlling for instance nominal type I errors in the General-to-Specific/Autometrics search paths, the tolerated 'multiple' type I error γ (see below) for the MTPs or some threshold for variable importance in BMA (e.g., the popular choice of a posterior inclusion probability of more than 50%). Employing common choices for these specifications, this paper provides a thorough comparison of the above-mentioned MTPs with these widely used model selection methods. This is done by means of a Monte Carlo study as well as by empirically investigating the prominent example of variable selection in cross-section growth regressions.

The Monte Carlo study in Section 1.3 investigates data generating processes (DGPs) that intend to mimic data sets often found in growth empirics. Unsurprisingly, the results demonstrate that there exists no uniformly best model selection procedure. Model selection—in fact, any inferential procedure—always implies a tradeoff between size and power. Specifically, higher tolerated size generally leads to higher power, where we define power to be the number of correctly selected variables. (Our concrete measure of size in the present multiple testing situation is explained in the following paragraph.) Hence, the notion of 'the best procedure' strongly depends on the researcher's preferences regarding the 'size-power tradeoff'. Our approach to investigating the effectiveness of the procedures is to compare their power in situations in which they have the same, or very similar, size. That is, we work with a measure of 'size-adjusted' power. Similarly, we look for constellations in which two procedures have very similar power, but differ in the required size to achieve that

power. Using these benchmarks, one central message of our paper is that some methods are practically dominated by others: among the MTPs, the bootstrap method of Romano, Shaikh, and Wolf (2008a) identifies most relevant variables for a given size requirement. Moreover, we find that General-to-Specific/Autometrics and some variants of BMA seem to be a bit more powerful given a certain size than the MTPs. The approaches of Sala-i-Martin (1997) and the Lasso are dominated, for example, by the General-to-Specific/Autometrics approach which identifies roughly as many relevant variables while having a much smaller size.

In a multiple testing situation such as model selection, one needs to suitably generalize the notion of a type I error. The notion we focus on in this paper is the false discovery rate (FDR, Benjamini and Hochberg, 1995). The FDR is defined as the expected value of the number of falsely rejected hypotheses divided by the overall number of rejections. We argue that FDR-control is a useful notion in the present problem: researchers may be willing to tolerate a small expected fraction γ of erroneously selected variables among all selected variables. That is, they wish to avoid overly many 'false positives.' For the example of selecting growth determinants, this implies that researchers are willing to expect that a small number of all growth determinants found significant actually do not drive economic growth. Hence, the FDR can also be used as a possible measure for 'size' in these multiple testing situations: the observed FDR in simulations gives an indication of the fraction of all rejections one needs to expect to be false positives. As such, it is related to the definition of size in a single hypothesis test, where size is defined as the probability of obtaining a false positive.

Concerning control of the FDR, the Monte Carlo study reveals that the MTPs are the only model selection procedures to consistently control the FDR. To the extent that the FDR is agreed to be a useful multiple type I error rate, we consider this to be an important finding: the well-established standard approach for single hypothesis tests is to focus on tests controlling size at some prespecified level over a wide range of DGPs. One then proceeds to look for tests with high power within this class. It therefore does not seem implausible to adopt an analogous strategy in multiple testing situations.

In the empirical application of Section 1.4, the variables jointly selected by all model selection procedures mostly have a plausible economic or cultural and religious motivation. The MTPs find few growth determinants beyond initial GDP when controlling the FDR at very small levels, providing some evidence for conditional convergence. We further find

that the MTPs, PcGets/Autometrics and Lasso identify similar variables, which may differ substantially from those identified by BMA. Initial GDP is included by all procedures.

In the following, Section 1.2 illustrates the problem of multiplicity using the example of cross-section growth regressions to then sketch the different model selection procedures including the MTPs. Section 1.3 presents the setup and findings of the Monte Carlo study. Section 1.4 applies the model selection procedures to the empirical example of cross-section growth regression. Section 1.5 concludes.

1.2 Problem and methods

Cross-section growth regressions

To further motivate the testing problem and to prepare the ground for the empirical application of Section 1.4, let us discuss the leading example of selecting growth determinants in some more detail.

Given the uncertainty surrounding the true drivers of growth, cross-section growth regressions simultaneously test many variables for significance. It is standard practice to regress the logarithm of real per capita output (in PPP terms) on two sets of variables. The first set includes variables measuring the initial position of an economy. The second set uses variables accounting for the difference in steady states across economies. Such a specification is consistent with a variety of neoclassical growth models where log-linearization around the steady-state leads to the expression (Barro and Sala-i-Martin, 1995)

$$\log y_T - \log y_0 = -(1 - e^{-\lambda T}) \log y_0 + (1 - e^{-\lambda T}) \log y^*, \tag{1.1}$$

where $\log y_t$ is the logarithm of the per capita gross domestic product (GDP for short) at time t, $\log y^*$ is its steady-state value, and λ is the convergence rate. Since heterogeneous economies have different steady states ('conditional convergence'), empirical counterparts of (1.1) employ additional variables $\boldsymbol{x}_i$ to proxy the steady state of an economy:

$$\log(y_{iT}/y_{i0}) = \mu + \delta \log(y_{i0}) + \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n, \tag{1.2}$$

with u_i an error term and n the number of observations. Although specification (1.2) is widely used in the literature, there is little agreement on which variables to include in $\boldsymbol{x}_i$. Hence, many variables are considered. The final model contains too many variables if

all those for which a level-α test rejects are included, that is, if no multiplicity control is performed. However, it is important to avoid spurious findings about growth determinants, given the relevance to policy: growth regressions can be used to identify growth-enhancing policies. Now, if variables are only spuriously found to influence growth, ineffective public expenditures may arise.

Controlling for Multiplicity via the FDR

This subsection describes how FDR control can be achieved in order to resolve the problem of multiplicity. As mentioned before, multiplicity arises if one tests a large number of hypotheses at the same time. If each test is evaluated individually, the data is given many chances of producing false rejections. The multiplicity issue is thus related to, but different from, data mining. To quote Lovell (1983), 'a data miner uncovers t-statistics that appear significant [...] by running a large number of alternative regressions on the same [...] data.' Doing so, 'the probability of a type I error of rejecting [a true] null hypothesis is much greater than the claimed 5%.' The multiple testing issue differs from data mining in that it arises even if the researcher is not 'mining' the data by trying many specifications to obtain a significant result, but simply because a large number of hypotheses is being tested in a single long regression.

We use multiple testing procedures (MTPs) that ensure that the FDR, the expected ratio of falsely-selected to the total number of selected variables, is no more than some small user-chosen γ in order to select variables in a cross-section regression. So far, the econometrics literature has paid only limited attention to solving the multiplicity issue via MTPs; for recent exceptions, see Hanck (2009) and Moon and Perron (2012). Textbooks (e.g. Mittelhammer, Judge, and Miller, 2000), if anything, usually present only the classic Bonferroni procedure as a solution to multiplicity, which only rejects a hypothesis if $\widehat{p}_j \leqslant \alpha/k$. This leads to rather low power, as for example shown in Section 1.3. White (2000) proposes a bootstrap 'reality check' to test whether the *best* model (e.g. a trading strategy in finance) beats a benchmark (e.g. the efficient market hypothesis). His approach would not be useful here as we are interested in selecting possibly, and presumably, more than one relevant variable from a set of candidate regressors.

We now describe the FDR as well as the MTPs considered here in more detail. This subsection partly draws on Benjamini and Hochberg (1995), Romano, Shaikh, and Wolf (2008b) and Deckers and Hanck (2013).

Table 1.1: Number of decisions when testing k null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**F**	k_0
Non-true null hypotheses	**T**	**S**	$k - k_0$
	k-**R**	**R**	k

Benjamini and Hochberg (1995) introduce the FDR as a useful notion of type I errors in multiple testing situations. Adapting a notation similar to theirs and referring to Table 1.1, there are k simultaneously tested hypotheses, out of which k_0 are true. Here, k might equal the number of t-statistics on the k variables $(\log(y_{i0}), \boldsymbol{x}_i')'$ from (1.2). $\mathbf{R}$, the total number of rejections, is an observable random variable. $\mathbf{U}$, the number of correctly accepted hypotheses, $\mathbf{F}$, the number of falsely rejected hypotheses, $\mathbf{S}$, the number of correctly rejected hypotheses and $\mathbf{T}$, the number of falsely accepted hypotheses, are clearly unobservable: we test precisely because we do not know if a hypothesis is true or not—if we knew the truth we would not need statistical inference. In the familiar setting of a single hypothesis test, $\mathbf{F}$ is 0 or 1, where $P(\mathbf{F} = 1)$ is the size of the test. Similarly, $P(\mathbf{S} = 1)$ then is the power of the single test. The proportion of falsely rejected null hypotheses to all rejected hypotheses can be described by $\mathbf{Q} = \mathbf{F}/(\mathbf{F}+\mathbf{S})$. Naturally, if no hypothesis is rejected (i.e., $\mathbf{R} = \mathbf{F}+\mathbf{S} = 0$), we take $\mathbf{Q} = 0$. The FDR is then defined as $\mathrm{E}(\mathbf{Q}) = \mathrm{E}(\mathbf{F}/(\mathbf{F} + \mathbf{S})) = \mathrm{E}(\mathbf{F}/\mathbf{R})$. Other notions of multiple error rates such as the Familywise Error Rate (FWER) (Romano and Wolf, 2010), the probability of one or more false rejections, correspond to $P(\mathbf{F} \geq 1)$. The FWER is a stricter notion than the FDR: whenever one controls the FWER at some level γ one also controls the FDR at the same level, as $P(\mathbf{F} \geq 1) \geq \mathrm{E}(\mathbf{Q})$.

There are many MTPs to control the FDR (or also the FWER), i.e., to ensure that FDR $\leqslant \gamma$ where γ is a user-specified level. From a theoretical perspective, MTPs are attractive as they enjoy certain optimality properties in a class of model selection procedures (Abramovich et al., 2006). These methods can be categorized into single-step methods (e.g., the Bonferroni procedure) that apply a single critical value to all test statistics, and sequential methods. Sequential methods first sort the hypotheses from most to least significant and start the decision process at either the largest or smallest test statistic.

In the following, we assume without loss of generality that large test statistics provide evidence against the null hypothesis and hence correspond to small p-values. Then, a step-up procedure first evaluates the least significant hypothesis (the smallest test statistic). If the hypothesis is accepted by the procedure, it proceeds to evaluating the second-least

significant one (steps-up to the second smallest test statistic). It continues to proceed to further hypotheses until it rejects a hypothesis for the first time. This and all hypotheses associated with larger test statistics are then rejected. A step-down procedure works the other way round. It starts by (possibly) rejecting the most significant hypothesis (corresponding to the largest test statistic). If it does, it proceeds to the hypothesis associated with the second-largest test statistic and continues to proceed (step-down) to further hypotheses until it accepts a hypothesis for the first time. This and all hypotheses associated with smaller test statistics are then accepted.

Since sequential methods apply tailored critical values for each hypothesis, they are generally expected to be more powerful than single-step methods. It is however not generally clear whether step-up or step-down procedures are more powerful. We now present two sequential FDR controlling methods in more detail, viz. the step-up method from Benjamini and Hochberg (1995) (BH) and the bootstrap step-down method of Romano, Shaikh, and Wolf (2008a) (bootstrap method). We first present the algorithms to implement the two methods along with some intuition and give graphical illustration further below.

BH method

First, choose a level γ at which to control the FDR. The BH method works with p-values. Let $\widehat{p}_{(1)} \leq \ldots \leq \widehat{p}_{(k)}$ be the sorted p-values and $H_{(1)}, \ldots, H_{(k)}$ the corresponding null hypotheses, arranged from most to least significant. For $1 \leq j \leq k$, let $\gamma_j = j\gamma/k$. Then, the method rejects $H_{(1)}, \ldots, H_{(j^*)}$, where j^* is the largest j such that $\widehat{p}_{(j)} \leq \gamma_j$. If no such j exists, no hypothesis is rejected. Hence, BH is trivial to implement for the user: starting with the largest p-value, just compare each p-value to its cutoff γ_j. As soon as one p-value is smaller than its cutoff, reject the corresponding hypothesis and all hypotheses corresponding to lower p-values.

Benjamini and Yekutieli (2001) show control of the FDR under positive regression dependency, which under certain conditions includes coefficient test statistics in regressions. Hence, crucially, the procedure deals with the empirically relevant situation that the regressors, and hence p-values, are correlated. The Monte Carlo study in Section 1.3 shows that the FDR is also controlled under plausible assumptions about e.g. the DGP of a cross-section growth regression. Benjamini and Gavrilov (2009) also offer ample encouraging simulation evidence under general patterns of dependence.

Bootstrap method

As before, assume that a hypothesis $H_{(j)}$ is rejected for large values of its corresponding test

statistic $T_{(j)}$. Further, sort the statistics from smallest to largest, i.e., $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(k)}$. Let $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ denote the corresponding hypotheses. As explained earlier, a step-down procedure then compares the largest statistic $T_{(k)}$ with a suitable critical value c_k. (Clearly, the critical values depend on γ, but we typically suppress this dependence in what follows so as not to clutter the notation.) If $T_{(k)} < c_k$ the procedure rejects no hypothesis, as not even the largest test statistic exceeds its critical value. Otherwise it rejects $H_{(k)}$ and steps down to $T_{(k-1)}$. The procedure continues in this fashion until it either rejects $H_{(1)}$ or does not reject the current hypothesis. That is, a step-down procedure rejects hypotheses $H_{(k)}, H_{(k-1)}, \ldots, H_{(k-j^*)}$, where j^* is the largest integer j satisfying

$$T_{(k)} \geq c_k, T_{(k-1)} \geq c_{k-1}, \ldots, T_{(k-j)} \geq c_{k-j}$$

Some intuition for the bootstrap method is as follows. For any step-down procedure the FDR can be written as[1]

$$\begin{aligned} \text{FDR} &= E\left[\frac{\mathbf{F}}{\max\{\mathbf{R}, 1\}}\right] = \sum_{1 \leq r \leq k} \frac{1}{r} E[\mathbf{F}|\mathbf{R} = r] P\{\mathbf{R} = r\} \qquad (1.3) \\ &= \sum_{1 \leq r \leq k} \frac{1}{r} E[\mathbf{F}|\mathbf{R} = r] \cdot P\left\{T_{(k)} \geq c_k, \ldots, T_{(k-r+1)} \geq c_{k-r+1}, T_{(k-r)} < c_{k-r}\right\}, \end{aligned}$$

where the event $T_{(k-r)} < c_{k-r}$ is defined to be true when $r = k$. Of course, through $\mathbf{F}$, (1.3) depends on the number of true hypotheses k_0. Clearly, k_0 is unknown. Hence, in order for a procedure to control the FDR at level γ, (1.3) needs to be bounded above by γ for every possible k_0. That is precisely the condition used to recursively determine the critical values c_j that lead to an FDR controlling procedure. It is quite straightforward to show that (see Romano, Shaikh, and Wolf, 2008b) for e.g. $k_0 = 1$, (1.3) simplifies to $\text{FDR} = P\{T_{1:1} \geq c_1\}/k$, where $T_{r:k_0}$ is the rth smallest statistic of the k_0 true hypotheses. Hence, the first critical value is the smallest number such that (1.3) is bounded above by γ for $k_0 = 1$, i.e.,

$$c_1 = \inf\left\{x \in \mathbb{R} : P\{T_{1:1} \geq x\}/k \leq \gamma\right\}$$

[1]The first equality is simply a definition, where the max in the denominator serves to avoid division by zero if $r = 0$. The second equality follows from the law of iterated expectations, $E\left[\frac{\mathbf{F}}{\mathbf{R}}\right] = E\left[\mathrm{E}\{\frac{\mathbf{F}}{\mathbf{R}}|\mathbf{R}\}\right] = E\left[\frac{1}{\mathbf{R}}\mathrm{E}\{\mathbf{F}|\mathbf{R}\}\right]$. Since r is clearly discrete (in fact, integer), the expected value of the random variable $\frac{1}{\mathbf{R}}\mathrm{E}\{\mathbf{F}|\mathbf{R}\}$ is, as usual, given by the sum over all possible nonzero values of the random variable times its probability, $P\{\mathbf{R} = r\}$. Now, the third equality follows because $P\{\mathbf{R} = r\}$, the probability of obtaining r rejections is, for any step-down procedure, the same as the probability of the r largest statistics $T_{(k)}, \ldots, T_{(k-r+1)}$ exceeding their critical values, but that of the $(r-1)$th largest statistic $T_{(k-r)}$ not exceeding its critical value.

If $k\gamma > 1$, we look for the smallest x such that $P\{T_{1:1} \geq x\} \leq k\gamma > 1$. As any x trivially satisfies this condition one takes $c_1 = -\infty$. This would for instance obtain if $\gamma = 0.05$ and $k > 20$, which is plausible in the context of cross-section growth regressions. This means that if the step-down procedure has rejected the hypotheses $H_{(k)}, H_{(k-1)}, \ldots, H_{(2)}$ corresponding to the $k-1$ largest test statistics, then $H_{(1)}$ will be rejected, too. For $k_0 = 2$, (1.3) can again straightforwardly be shown to be equal to

$$\frac{1}{k-1}P\{T_{2:2} \geq c_2, T_{1:2} < c_1\} + \frac{2}{k}P\{T_{2:2} \geq c_2, T_{1:2} \geq c_1\} \tag{1.4}$$

Hence, having determined c_1, c_2 then simply is the smallest number for which (1.4) is bounded above by γ. The remaining critical values $c_3, \ldots, c_k$ can be found using further steps of the recursion.

In practice these c_i are unavailable, since the probability measure P is unknown—again, if we knew the truth we would not need statistical inference. We therefore approximate the joint null distribution of the T_j using the bootstrap. Appendix A1.1 gives details on an appropriate bootstrap procedure for the present problem. The key aim of the bootstrap is to properly preserve the dependence structure of the variables so as to provide valid inference in the relevant scenario of correlation.

Refer to Figure 1.1 for a graphical illustration of BH and the bootstrap method. It is based on the empirical results from Section 1.4 using the growth data set of Fernandez, Ley, and Steel (2001), which is based on that of Sala-i-Martin (1997). The bulleted line gives the $T_{(j)}$, that is, the absolute values of cross-section regression t-statistics sorted from large to small (equivalently, from more to less significant). We equivalently express the cutoffs for BH at the scale of test statistics rather than at that of p-values (i.e., $j\gamma/k$) as in the above description, allowing us to present the BH results in one figure with the bootstrap method results. This implies that rejections correspond to a test statistic *exceeding* a BH cutoff, as with the bootstrap methods, because a *small* p-value p_j corresponds to a *large* T_j.

Since BH is a step-up procedure, it starts at the right of the plot, asking if the smallest (i.e., least significant) test statistic exceeds the cutoff. Since this is not the case, it moves left until it first finds a test statistic that exceeds the corresponding cutoff, and declares this statistic as significant. The reasoning of a step-up procedure then is that if this statistic is significant, all statistics that are even larger are so, too. Note that this is the case although it is in principle possible—but not the case in this data—that the cutoff line again crosses

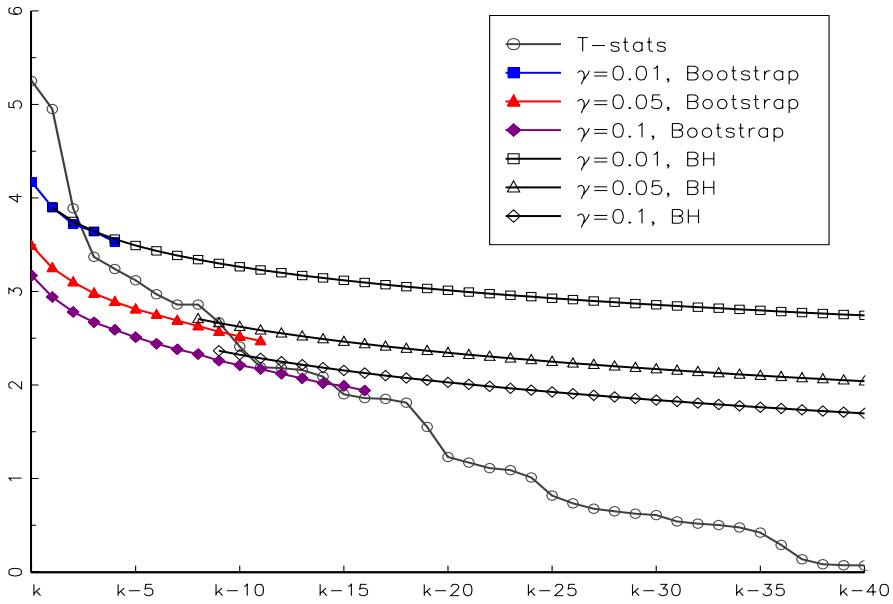


Figure 1.1: Test statistics and critical values for the FLS/Sala-i-Martin (1997) data

Note: The sorted test statistics $T_{(j)}$ and the relevant bootstrap critical values $\hat{c}_{\gamma,j}$ (i.e., those until and around $T_{(j)} < \hat{c}_{\gamma,j}$) and BH critical values for different γ plotted against the ranks of the T_j.

the sorted test statistics from below. Since cutoffs to the left of the first crossing from below thus do not matter anymore, we also do not plot all of them for better readability of the figure. If even the smallest test statistic lies above its cutoff, the procedure stops and rejects all hypotheses. Unsurprisingly, the cutoff curves lie higher the lower we choose γ: if we test at a more stringent level γ, larger test statistics T_j are necessary to declare a statistic to be significant. Had we chosen a γ so small that even $\widehat{p}_{(1)} > \gamma_1 = \gamma/k$, then the cutoff line (at the scale of test statistics) would lie entirely above the sorted test statistics, and no hypothesis would have been rejected.

The bootstrap method, in turn, is a step-down procedure that therefore starts at the left of the figure: it first asks if the largest statistic exceeds the cutoff. If not, then γ has been chosen to be so low (and/or the data lead to such low test statistics) that no statistic is declared as significant and the procedure stops. If the largest test statistic exceeds its cutoff, it is declared significant and the procedure continues rejecting for as long as the test statistics exceed the cutoff line. The first statistic that does not exceed its cutoff is declared non-significant. The reasoning of a step-down procedure then is that if this statistic is not significant, all statistics that are even smaller are not, either (even though it is possible—although not the case in this data—that the sorted test statistics again cross the cutoff line

from below further to the right).

Figure 1.1 also illustrates why the MTPs avoid the overly many rejections resulting from the 'classical approach' to hypothesis testing (i.e., rejecting H_j if, say, $\widehat{p}_j \leq 0.05$). Expressed at the scale of test statistics and using normal critical values, the cutoff line would be flat at 1.96, and (possibly spuriously) more test statistics exceed this cutoff line. MTPs, in turn, employ suitably larger cutoffs so as to avoid these spurious rejections.

Clearly, the very active multiple testing literature has proposed several other FDR- and FWER-controlling procedures. Classical and recent examples include Holm (1979), Storey (2002), Storey, Taylor, and Siegmund (2004), Benjamini, Krieger, and Yekutieli (2006), Sarkar (2006) and Romano and Wolf (2010).[2] We shall focus on the procedures described above as these arguably belong to the most popular ones, but also for brevity.

Other variable selection procedures

We compare variable selection via multiple testing with other popular selection procedures. The leading procedures considered here have all been applied to the problem of selecting variables in growth regressions in earlier studies (see also Section 1.4). Concretely, we investigate the General-to-Specific approach of Hoover and Perez (1999) and Krolzig and Hendry (2001), Bayesian Model Averaging as advocated in e.g. Fernandez, Ley, and Steel (2001) or Ley and Steel (2009), the two million regressions approach of Sala-i-Martin (1997) and the Lasso version employed in Schneider and Wagner (2012).[3] The following paragraphs briefly sketch these methods.

General to specific: Hoover and Perez (1999)

Hoover and Perez (1999) argue that if the general unrestricted model (GUM) provided by all available regressors nests a good approximation of the DGP, then 'General to simple' (Gets) selection would find the best model, a parsimonious model that conveys all of the information of the GUM in a more compact form. To go from the general model to the

[2]Indeed, we also obtained results for the procedures of Storey, Taylor, and Siegmund (2004) and Benjamini, Krieger, and Yekutieli (2006), which are very similar to the Bootstrap and BH results and hence we do not include them in the main text. These results may be found in the Appendix A1.2, see Tables 1.8-1.10.

[3]Due to space constraints, we again focus on what we believe are the most widely used techniques and omit several alternative suggestions. For instance, Acosta-González and Fernández-Rodríguez (2007) propose genetic algorithms based on the Bayesian Information Criterion to select regressors. In turn, building on general results in e.g. Goeman, van de Geer, and van Houwelingen (2006) and Huang, Horowitz, and Wei (2010), Jensen (2010) and Jensen and Würtz (2012) deal with the high-dimensional case in which the number of regressors $k > n$. Their approaches allow inference for either the maximal t-statistic or a specific variable of interest even if $k > n$. Our interest is however in all variables of a regression. As such, the procedures discussed here do not readily allow for the case $k > n$.

simplified model, a chain of simplifications such as eliminating insignificant variables is applied, while checking if the simplified model is a valid restriction of the general model at each step. For any given sample, it may be the case that a particular simplification results in an inappropriate model. Hoover and Perez (1999, HP) therefore suggest to conduct a multi-path search and to test the final models of the different search paths against each other. See Hoover and Perez (1999) and Hoover and Perez (2004) for a more detailed description. Hendry and Krolzig (2004) further note that if all regressors were orthogonal, then the ordered squared t-statistics from the GUM, $T^2_{(1)} \geq T^2_{(2)} \geq \ldots \geq T^2_{(k)}$, could be validly used for model selection. One would then select variables $1, \ldots, \tilde{k}$ with $T^2_{(\tilde{k})} \geq c_\alpha$ and discard the remaining ones, where $T^2_{(\tilde{k}+1)} \leq c_\alpha$. Since t-statistics in fact are not mutually orthogonal, multiple search paths are introduced. The HP approach is refined in Krolzig and Hendry (2001) (PcGets) through the addition of e.g. further misspecification tests; again refer to the original contribution for details. Our empirical application uses the Autometrics implementation of PcGets.

The general to specific approach is somewhat similar to the above MTPs, where the point of departure is also a long regression, decisions are also based on individual t-statistics and their dependence is taken into account using the techniques described earlier in this section. For example, it is easily seen that BH rejects $H_{(1)}, \ldots, H_{(\tilde{k})}$ if and only if

$$T^2_{(\tilde{k})} > \big(F^{-1}\big(0.5\tilde{k}\gamma/k\big)\big)^2 \qquad \text{but} \qquad T^2_{(j)} < \big(F^{-1}\big(0.5j\gamma/k\big)\big)^2, \quad j = \tilde{k}+1, \ldots, k,$$

with F^{-1} the quantile function of the null distribution of the t statistics (usually, the standard normal or t distribution). Hence, the $T_{(j)}$ are also compared to an 'adaptive' (i.e., depending on k) sequence of critical values, $F^{-1}(j\gamma/(2k))$. A potential drawback—not easily investigated analytically—resulting from the several regressions that need to be performed in the HP/PcGets/Autometrics multi-path searches is that such searches might suffer from post-model selection distortions in non-orthogonal data (Leeb and Pötscher, 2005).

Bayesian model averaging

Fernandez, Ley, and Steel (2001) (FLS) use BMA to account for model uncertainty in growth regressions. Given the k candidate regressors, there are 2^k possible correct models, assuming that these nest the true DGP. One assigns a prior distribution to each model as well as to the inclusion of a certain variable in each model. FLS then compute the marginal posterior probability of including a certain variable: '[It] is simply the sum of the

posterior probabilities of all models that contain this regressor.' The Bayesian framework of course has no critical value for a variable's marginal inclusion probability to declare it (non-)significant. Nevertheless, following e.g. Eicher, Papageorgiou, and Raftery (2011) (who draw on Jeffreys, 1961), a variable can be declared to be 'important' if its marginal inclusion probability exceeds 50%. We adopt this choice in what follows.[4] In further work, Ley and Steel (2009) demonstrate that different choices of prior densities of the parameters as well as choices on the expected model size highly influence the model selection results. In particular, Ley and Steel (2009) consider BMA on regression models M_j with $k_j \leq k$ regressors grouped in $\boldsymbol{X}_j$ leading to

$$\boldsymbol{y}|a, \boldsymbol{\beta}_j, \sigma \sim \mathrm{N}(a\boldsymbol{\iota} + \boldsymbol{X}_j\boldsymbol{\beta}_j, \sigma^2\boldsymbol{I}),$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{k_j}$ and $\sigma \in \mathbb{R}_+$ is a scale parameter. For the prior density of the parameters, Ley and Steel (2009) use a combination of a 'non-informative' improper prior on a and σ as well as a 'g-prior' on $\boldsymbol{\beta}$. Fernandez, Ley, and Steel (2001) employ $g = 1/\max\left\{n, k^2\right\}$. As shown by Feldkircher and Zeugner (2009), this choice of g leads BMA to act in a model selection fashion, that is, to place a high weight on the best-performing model. The prior model probabilities can be written as $P(M_j) = \theta^{k_j}(1-\theta)^{k-k_j}$, assuming that each regressor enters a model with equal and fixed probability and independently of the others. Then, the expected model size m equals $k\theta$. Ley and Steel (2009) also use a hierarchical prior with random θ's drawn from a beta distribution with parameters (a_1, b_1). Fixing $a_1 = 1$ and choosing a prior model size m, one can then calculate b_1 as $(k-m)/m$. Hence, for both fixed and random θ, one only needs to specify the expected model size m. In our simulation study, we investigate all eight combinations of $g = \{n^{-1}, k^{-2}\}$, fixed or random θ and expected model sizes equal to $\{7, k/2\}$, adopting the choices from Ley and Steel (2009).

Two million regressions: Sala-i-Martin (1997)

Sala-i-Martin (1997) flags a variable as significant if, controlling for three variables deemed to be important by default and a varying selection of three additional control variables, its coefficient estimate ± twice its standard error lies to one side of zero in more than 95% of the regressions. We follow Hoover and Perez (2004) and, unlike in Sala-i-Martin's empirical

[4] Clearly, other measures of evaluating the importance of a variable could be, and are, used. For example, a variable could be deemed to be important if its posterior inclusion probability exceeds its prior inclusion probability. Brock and Durlauf (2001, pp. 252-3) provide further discussion of the merits and drawbacks of the above common choice in the BMA literature.

application (cf. Section 1.4), do not include any variables by default in the Monte Carlo study of Section 1.3. We also follow Sala-i-Martin (1997) and Hoover and Perez (2004) and employ three control variables. The outcome of this approach depends on a large set of misspecified regressions, in particular via the inherent omitted variable bias when conducting many overly short regressions.

Lasso

The Lasso (Least Absolute Shrinkage and Selection Operator) was introduced by Tibshirani (1996). The Lasso is the solution to the penalized linear regression

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}\left|\beta_j\right| \tag{1.5}$$

Due to the penalty function, the estimator sets some coefficients exactly equal to zero and thus performs model selection. The penalty function also induces a bias towards zero for large λ. Hence, one commonly performs model selection via the Lasso and reestimates the final model using e.g. OLS to obtain point estimates and standard errors. As one can see from (1.5), the selection of variables depends on the choice of λ. A common practice is to inspect plots of coefficients for different λ's and in that way choose important variables. Such a procedure is infeasible in a Monte Carlo simulation. Thus, in the simulation study of Section 1.3 λ is chosen by fivefold cross-validation, following Tibshirani (1996). We use code designed by Friedman, Hastie, and Tibshirani (2010), who use cyclical coordinate descent calculated along a regularization path. The algorithm is designed for the more general case of generalized linear models with elastic-net penalties of which the Lasso is a special case. The algorithm is especially suited for large data sets, since it performs estimation faster than its competitors (Friedman, Hastie, and Tibshirani, 2010).

In their application to growth regressions, Schneider and Wagner (2012) use the adaptive Lasso estimator (see Zou, 2006, for a detailed discussion) for variable selection. The adaptive Lasso is the following modification of the traditional Lasso estimator:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{k}\left|\frac{\beta_j}{\tilde{\beta}_j}\right|,$$

where $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator. One typically obtains $\tilde{\boldsymbol{\beta}}$ from estimating the full model with all k candidate variables by OLS. A potential drawback of the adaptive Lasso is that a tuning parameter needs to be chosen by the user (or some technical procedure),

possibly resulting in choices suboptimal for model selection. Schneider and Wagner (2012) choose λ via generalized cross validation, thus minimizing the squared error, as also done in fivefold cross-validation. Leng, Lin, and Wahba (2006) point out that choosing λ to maximize predictive performance may result in poor model selection performance for some data structures.

1.3 Monte Carlo study

This section provides a simulation study comparing the effectiveness of the model selection procedures presented above. We first describe the setup of the simulation study to then compare the 'size'—measured by the observed FDR—and 'power'—that is, the capability of identifying relevant variables—of the different procedures. Moreover, we present robustness checks for the results of the Monte Carlo simulation. Unsurprisingly, it will be seen that there is a certain size-power tradeoff involved in choosing a model selection procedure: a method that discovers many relevant variables typically does so at the cost of having higher size, that is, finding more false positives. For example, using an MTP with a higher γ or BMA with a cutoff of less than 50% for the posterior probability of course leads to both more true and false discoveries. This also implies that no method dominates all the others in terms of power. Nevertheless, to still be able to compare the methods in a meaningful way, we apply a measure of 'size-adjusted' power. That is, we compare their power in situations in which the methods have the same, or very similar, size. Alternatively, we look at cases in which procedures have similar power, but differ in the required size to achieve that power.

More specifically, our main findings are as follows. We show that 'classical testing' (i.e., rejecting H_j if $\widehat{p}_j \leq \alpha$) results in very large FDRs and confirm that the Bonferroni correction leads to very low power. All MTPs are shown to consistently control the FDR at the intended level. They are the only model selection procedures to do so. Given that size-control is a widely accepted property a testing procedure should have, we consider this to be an important result. Among the MTPs, the bootstrap is the most powerful procedure. We present the performance of BMA under different choices of prior. We corroborate the findings of Ley and Steel (2009), who show that the performance heavily depends on the choice of priors. In instances where the 'size' of the BMA variant is comparable to that of the MTPs, BMA appears to be slightly more powerful. The same also holds for

Table 1.2: Implied population R^2's of DGP (1.6)

Population R^2 as every ... β_ℓ equals 0.5				
		...tenth	*...fifth*	*...second*
	0	0.556	0.714	0.862
ρ	0.3	0.733	0.902	0.981
	0.5	0.789	0.932	0.988

the PcGets/Autometrics approach. In turn, PcGets/Autometrics and the BMA variants achieve similar power in these situations. The approach by Sala-i-Martin seems to be practically dominated by, for example, the PcGets/Autometrics approach, since the latter finds as many relevant variables while achieving a much smaller FDR. The Lasso, with λ being chosen via fivefold cross-validation, requires a much larger FDR to find the same number of correct rejections as, e.g., the PcGets/Autometrics approach.

Design

We employ a cross-section regression framework. The design extends the setup of Romano, Shaikh, and Wolf (2008a) who only consider relatively simple location testing problems and find large power gains of the bootstrap approach relative to other MTPs, including BH. We estimate the FDR through the average proportion of false rejections. To investigate the power of the procedures we consider the average number of correct rejections. Our DGP is as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \qquad \boldsymbol{u} = (u_1, \ldots, u_{n=100})' \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_{n=100}), \tag{1.6}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_k')'$ is a 100×50 regressor matrix. (Unreported experiments draw qualitatively similar pictures for different n and k.) Each row $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{ki})$ is multivariate normal with mean zero, variances one and common correlation ρ. We consider $\rho = \{0, 0.3, 0.5\}$, which induces correlation in $\boldsymbol{X}$ and hence in the test statistics, which the procedures must be able to handle. We investigate three scenarios for $\boldsymbol{\beta}$:

(i) Every tenth $\beta_\ell = 0.5$, and the remaining $\beta_\ell = 0$, such that there are 5 false hypotheses.

(ii) Every fifth $\beta_\ell = 0.5$, and the remaining $\beta_\ell = 0$, such that there are 10 false hypotheses.

(iii) Every second $\beta_\ell = 0.5$, and the remaining $\beta_\ell = 0$, such that there are 25 false hypotheses.

The value 0.5 allows discrimination between the power of the procedures. An extremely high β_ℓ for example, results in all false hypotheses being rejected by all procedures. A low

β_ℓ results in the opposite. This simulation design implies population R^2's that are realistic for data sets typically encountered in for instance growth econometrics. To see this, recall that the population R^2 is given by $\rho^2 = 1 - \mathrm{Var}(u_i)/\mathrm{Var}(y_i)$. Here,

$$\mathrm{Var}(y_i) = \mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta} + u_i) = \boldsymbol{\beta}' \mathrm{Var}(\boldsymbol{x}_i)\boldsymbol{\beta} + \mathrm{Var}(u_i) = \boldsymbol{\beta}' \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \boldsymbol{\beta} + 1$$

Table 1.2 shows that all implied R^2's are at least 0.5, often considered a lower bound in growth empirics. Several combinations of ρ and $\boldsymbol{\beta}$ investigate R^2's in the range 0.7-0.9, while the case of many relevant highly correlated regressors studies the case of almost perfect explanatory power.

We use 2,000 replications for all except the BMA procedures, which were run with 1,000 replications.[5]

Results

Tables 1.3-1.5 show the results of the Monte Carlo study for all procedures described in Section 1.2. Our interpretation of the results is as follows.

MTPs, Bonferroni and classical testing

The MTPs control the FDR for any ρ and any configuration of relevant and irrelevant variables.[6] Classical testing does not control the FDR; FDRs substantially higher than the nominal level of the individual tests result. Tables 1.3-1.5 suggest that the fraction of false to total rejections can be very high for classical testing: at the 10% level, up to one in two rejections can be expected to be false in the present DGP (Table 1.5). Further, FDR violation increases in ρ. One reason is that the larger ρ, the more correlated the regressors. The regressors then more readily substitute for each other, or exhibit *negative jointness* as defined by Doppelhofer and Weeks (2009). This roughly means that two (or more) variables capture similar underlying determinants of the dependent variable. The

[5]Due to the heavy computational cost the BMA experiments were performed on the Groningen Millipede cluster in R for Linux 2.13.1 using twelve 2.6 GHz AMD Opteron cores. Using 50,000 burn-ins and 100,000 MCMC iterations, the exercise requires around four days of computation time.

[6]Besides, the well-known conservativeness of the BH method is also visible in our results. Benjamini and Hochberg (1995) show that the FDR of the BH method is smaller than the nominal level γ by at least a factor k_0/k, that is, the ratio of true to all hypotheses. Formally, FDR $\leq k_0 \cdot \gamma/k$. Clearly, this expression equals γ only if all hypotheses are true, $k_0 = k$. For instance, there are $k - 10 = k_0 = 40$ true hypotheses for the case of 10 nonzero β_ℓ, and the FDR of BH never exceeds $40 \cdot \gamma/50 = 0.8 \cdot \gamma$ in this case.

Table 1.3: Monte Carlo results: 25 False hypotheses

	$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
	FDR	*CR*	*FDR*	*CR*	*FDR*	*CR*
Classical: $\alpha = 0.01$	.011	19.83	.013	15.39	.017	11.32
Classical: $\alpha = 0.05$	.048	23.11	.054	20.61	.063	17.27
Classical: $\alpha = 0.1$	.090	24.01	.098	22.38	.103	19.98
Bonferroni: $\alpha = 0.01$	.000	8.79	.004	4.55	.001	2.28
Bonferroni: $\alpha = 0.05$	.001	13.20	.002	8.27	.003	4.78
Bonferroni: $\alpha = 0.1$	.003	15.20	.004	10.20	.005	6.40
BH: $\gamma = 0.01$	.005	16.67	.004	10.18	.004	5.16
BH: $\gamma = 0.05$	.025	21.61	.026	17.34	.024	11.81
BH: $\gamma = 0.1$	.050	23.07	.050	20.10	.048	15.77
Bootstrap: $\gamma = 0.01$	.008	18.15	.008	11.32	.006	5.60
Bootstrap: $\gamma = 0.05$	.046	22.80	.042	19.31	.034	13.18
Bootstrap: $\gamma = 0.1$	.095	23.93	.087	21.82	.075	17.27
PcGets/Autometrics	.072	24.31	.087	22.97	.109	20.56
HP	.022	22.19	.037	20.01	.059	17.19
Bayesian Model Averaging						
$m = k/2$. $g = k^{-2}$, random θ	.005	4.17	.035	17.26	.044	13.97
$m = 7$. $g = k^{-2}$, random θ	.004	3.02	.035	16.52	.045	13.45
$m = k/2$. $g = k^{-2}$, fixed θ	.028	17.05	.036	19.04	.046	16.06
$m = 7$. $g = k^{-2}$, fixed θ	.008	2.86	.037	12.82	.047	11.45
$m = k/2$. $g = n^{-1}$, random θ	.107	24.13	.034	19.70	.034	14.26
$m = 7$. $g = n^{-1}$, random θ	.077	23.45	.031	18.75	.033	13.56
$m = k/2$. $g = n^{-1}$, fixed θ	.068	23.98	.036	20.40	.037	16.57
$m = 7$. $g = n^{-1}$, fixed θ	.020	12.15	.030	13.93	.035	10.83
Sala-i-Martin	.155	17.21	.494	24.96	.499	25.00
Lasso	.285	24.80	.255	24.93	.263	24.80

Notes: This table reports the results of a Monte Carlo simulation (2,000 replications) using a DGP as described in the beginning of Section 1.3, for 25 (out of 50) false hypotheses (relevant variables). Further specific parameter values are indicated in the table. The different procedures for model/variable selection are described in Section 1.2. FDR is estimated through the average proportion of false rejections. CR denotes the average number of correct rejections.

Table 1.4: Monte Carlo results: 10 False hypotheses

	$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
	FDR	*CR*	*FDR*	*CR*	*FDR*	*CR*
Classical: $\alpha = 0.01$	.039	7.94	.047	6.19	.062	4.57
Classical: $\alpha = 0.05$	.156	9.26	.176	8.27	.199	6.90
Classical: $\alpha = 0.1$	.268	9.61	.284	8.95	.299	8.00
Bonferroni: $\alpha = 0.01$	.003	3.48	.003	1.87	.003	0.94
Bonferroni: $\alpha = 0.05$	.005	5.28	.008	3.34	.010	1.96
Bonferroni: $\alpha = 0.1$	.010	6.06	.014	4.11	.019	2.51
BH: $\gamma = 0.01$	.008	5.39	.008	2.96	.006	1.41
BH: $\gamma = 0.05$	.040	7.68	.038	5.45	.036	3.27
BH: $\gamma = 0.1$	.079	8.50	.076	6.77	.072	4.63
Bootstrap: $\gamma = 0.01$	.011	5.50	.010	3.10	.008	1.52
Bootstrap: $\gamma = 0.05$	.050	7.89	.046	5.71	.044	3.42
Bootstrap: $\gamma = 0.1$	.101	8.67	.096	7.03	.091	4.74
PcGets/Autometrics	.202	9.88	.211	9.55	.234	8.83
HP	.050	9.66	.071	9.04	.106	7.90
Bayesian Model Averaging						
$m = k/2$. $g = k^{-2}$, random θ	.007	6.57	.022	7.90	.044	6.51
$m = 7$. $g = k^{-2}$, random θ	.006	6.23	.021	7.79	.044	6.41
$m = k/2$. $g = k^{-2}$, fixed θ	.037	9.50	.044	8.93	.064	7.64
$m = 7$. $g = k^{-2}$, fixed θ	.008	7.01	.021	7.72	.045	6.39
$m = k/2$. $g = n^{-1}$, random θ	.059	9.59	.047	8.92	.057	7.41
$m = 7$. $g = n^{-1}$, random θ	.050	9.48	.040	8.82	.053	7.28
$m = k/2$. $g = n^{-1}$, fixed θ	.164	9.88	.140	9.56	.142	8.62
$m = 7$. $g = n^{-1}$, fixed θ	.025	9.29	.029	8.55	.047	7.06
Sala-i-Martin	.328	9.44	.761	10.00	.780	10.00
Lasso	.336	9.85	.425	9.96	.451	9.89

Notes: This table reports the results of a Monte Carlo simulation (2,000 replications) using a DGP as described in the beginning of Section 1.3, for 10 (out of 50) false hypotheses (relevant variables). See also notes to Table 1.3.

Table 1.5: Monte Carlo results: 5 False hypotheses

	$\rho = 0$		$\rho = 0.3$		$\rho = 0.5$	
	FDR	*CR*	*FDR*	*CR*	*FDR*	*CR*
Classical: $\alpha = 0.01$	.083	3.96	.097	3.08	.125	2.25
Classical: $\alpha = 0.05$	.273	4.61	.302	4.12	.336	3.48
Classical: $\alpha = 0.1$	.426	4.81	.457	4.49	.473	3.98
Bonferroni: $\alpha = 0.01$	.003	1.76	.003	0.94	.006	0.46
Bonferroni: $\alpha = 0.05$	.010	2.57	.014	1.65	.020	0.96
Bonferroni: $\alpha = 0.1$	.021	3.03	.031	2.06	.035	1.27
BH: $\gamma = 0.01$	.010	2.27	.007	1.21	.009	0.55
BH: $\gamma = 0.05$	.038	3.35	.043	2.29	.042	1.32
BH: $\gamma = 0.1$	.086	3.81	.092	2.85	.080	1.80
Bootstrap: $\gamma = 0.01$	.011	2.36	.009	1.22	.010	0.59
Bootstrap: $\gamma = 0.05$	.052	3.45	.050	2.39	.048	1.35
Bootstrap: $\gamma = 0.1$	.100	3.94	.102	2.95	.092	1.88
PcGets/Autometrics	.334	4.95	.344	4.83	.363	4.51
HP	.087	4.87	.123	4.63	.165	4.12
Bayesian Model Averaging						
$m = k/2$. $g = k^{-2}$, random θ	.005	3.97	.020	3.97	.047	3.34
$m = 7$. $g = k^{-2}$, random θ	.004	3.91	.020	3.94	.047	3.31
$m = k/2$. $g = k^{-2}$, fixed θ	.056	4.85	.066	4.58	.093	4.01
$m = 7$. $g = k^{-2}$, fixed θ	.007	4.38	.021	4.12	.050	3.48
$m = k/2$. $g = n^{-1}$, random θ	.042	4.75	.049	4.46	.070	3.82
$m = 7$. $g = n^{-1}$, random θ	.036	4.72	.042	4.42	.066	3.77
$m = k/2$. $g = n^{-1}$, fixed θ	.272	4.97	.266	4.83	.279	4.49
$m = 7$. $g = n^{-1}$, fixed θ	.040	4.81	.047	4.50	.072	3.88
Sala-i-Martin	.488	4.98	.670	4.99	.689	4.97
Lasso	.246	4.85	.434	4.97	.500	4.90

Notes: This table reports the results of a Monte Carlo simulation (2,000 replications) using a DGP as described in the beginning of Section 1.3, for 5 (out of 50) false hypotheses (relevant variables). See also notes to Table 1.3.

significance of one of these may then depend on whether the other is also included in the model or not. Overall, this tends to increase model uncertainty, and therefore makes FDR control more challenging. (We are grateful to the editor of the Oxford Bulletin of Statistics, Jonathan Temple, for bringing this mechanism to our attention.)

The non-control of the FDR using classical testing becomes less severe as k_0 decreases when moving from Table 1.5 to 1.3 (i.e. as the number of nonzero β_ℓ increases). Recalling $\text{FDR} = \text{E}\left(\mathbf{F}/(\mathbf{F}+\mathbf{S})\right)$, this is not surprising: as the number of correct rejections $\mathbf{S}$ increases and the number of false rejections $\mathbf{F}$ is kept constant, the FDR is lower by construction.

The MTPs identify a reasonably large number of the relevant variables. The 'discovery rate' $\mathbf{S}/(k-k_0)$ appears to increase in the number of nonzero β_ℓ.[7] Among the MTPs, the bootstrap yields the higher number of correct rejections, at the expense of non-negligible additional computational cost. Higher power likely is due to the bootstrap method taking the dependence between the test statistics into account.

Of course, as in any hypothesis test, a smaller nominal type-I error translates not only into fewer false but also fewer correct rejections, that is, lower power. As such, the average number of correct rejections reported in Tables 1.3-1.5 decreases in γ.[8] The choice of a suitable γ, or in fact even of a suitable model selection approach (see below), may therefore ultimately depend on the loss function of the analyst, i.e. the relative costs of false rejections and false acceptances.

The Bonferroni correction, even being an FWER controlling procedure, unsurprisingly controls the FDR. However, it identifies substantially fewer false hypotheses, i.e., has low power. Thus, the Bonferroni correction does not seem to be a useful model selection device.

BMA

The eight different BMA variants considered in Tables 1.3-1.5 study the eight combinations of two expected model sizes and g-prior with fixed and random θ considered by Ley and Steel (2009). The choice $m = k/2$, $g = k^{-2}$, fixed θ corresponds to the settings of FLS. The so-called unit information prior (UIP) is obtained for $m = k/2$, $g = n^{-1}$ and fixed θ. The following conclusions can be drawn from the results:

1. The FDRs of the BMA variants vary quite widely from 0.4% (Table 1.3, $m = 7$,

[7] Focusing on $\gamma = 0.1$, $\mathbf{S}/(k-k_0)$ increases from a low of 36% (1.8/5) for BH with 5 false nulls (Table 1.5) and $\rho = 0.5$ to a maximum of over 95% (23.93/25) in the case of the bootstrap, 25 false nulls and $\rho = 0$ (Table 1.3).

[8] For instance, the highest power by any of the MTPs at $\gamma = 0.01$ is achieved by the bootstrap, with power of 73% (18.15/25) for 25 false nulls and $\rho = 0$ (Table 1.3)—a drop of roughly 22 percentage points relative to the corresponding power for $\gamma = 0.1$.

$g = k^{-2}$, random θ) to 27.9% (Table 1.5, $m = k/2$, $g = n^{-1}$, fixed θ). It does not seem to be the case that specific BMA versions have a consistently higher FDR than others. For example, $m = k/2$, $g = n^{-1}$, fixed θ (cited above with the extreme value of 27.9%) has a low FDR of 3.7% for $\rho = 0.5$ and 25 false hypotheses. This FDR is roughly the same for $m = 7$, $g = n^{-1}$, fixed θ. The latter's FDR however never exceeds 5%. The FDRs seem to increase in ρ, and show no perceptible pattern in the number of nonzero coefficients—some FDRs increase, others decrease.

2. Some BMA variants achieve a remarkably high number of correct rejections, with discovery rates of up to almost 99% (Table 1.4, 9.88/10) for the UIP ($m = k/2$, $g = n^{-1}$, fixed θ). There is no clear pattern for correct rejections in ρ.

3. The BMA variants also exhibit the expected size-power tradeoff between FDR and correct rejections (CR). Ley and Steel (2009) find the posterior distributions of the number of regressors chosen with $g = 1/n$ to lie to the right to those with $g = 1/k^2$ (cf. their Figure 5). In line with this result the former BMA variants identify more relevant variables here than the latter. This higher power for $g = 1/n$ comes at the expense of a higher FDR than under $g = 1/k^2$, which, in the case of $m = k/2$ and fixed θ, attains values of almost 30% (Table 1.5). This implies that we need to expect almost one false out of every three rejections. This high FDR results for the case of only 5 relevant variables, in which a prior with a mean prior model size of $m = 25$ yields many false positives. We further corroborate the finding of Ley and Steel (2009) that the choice of m is less influential if θ is random than if it is fixed, as the former leads to a less concentrated prior. For instance, in the above case of $m = k/2$ and 5 relevant variables, a random θ leads to a substantially smaller FDR of 4%-7% (Table 1.5). That said, the decision of whether to specify a fixed or random θ does not appear to have a perceptible effect on the number of correct rejections.

4. We next attempt to provide a comparison of the power of the BMA variants with that of the MTPs. Such a comparison is not without difficulties since, as documented above, the power of the MTPs is higher if we test at a higher 'level' γ. Similarly, all the BMA variants would of course select more variables if we lowered the cutoff threshold of 50% (see also footnote 4) or increased the expected model size (see previous paragraph). Hence, there cannot be such a statement as that an MTP would always be less powerful than BMA, as it would always be possible to find a

sufficiently liberal γ such that the statement is not true.

As stated above, our approach for comparing powers is to look at what is sometimes called 'size-adjusted' power in simulation studies. Using this metric, the power of the BMA variants is similar to each other and perhaps slightly higher than that of the MTPs. Consider e.g. the case of 25 nonzero β_ℓ and $\rho = 0.3$. The MTPs at $\gamma = 0.05$ have FDRs of 2.6% and 4.2% and achieve roughly 17 and 19 correct rejections (Table 1.3). In this particular case, all BMA variants have an FDR of around 3% and between 13 and 20 correct rejections.

Sala-i-Martin

We observe a very high number of correct rejections for the two million regressions approach of Sala-i-Martin (1997). It even identifies all relevant variables in all 2,000 replications in several cases, e.g. for 25 relevant variables and $\rho = 0.5$ (Table 1.3). However, this comes at the price of an extremely high FDR which ranges between 15% and 78%. The latter case (indeed, any value larger than 50%) implies that, on average, this technique needs to be expected to make more false than true discoveries. In fact, its expected fraction of false discoveries strongly exceeds that of classical testing even at the 10% level. These findings are in line with Hoover and Perez (2004) who find that a high proportion of included variables are irrelevant (i.e., high size) as well as a high 'power ratio'. The latter measures power of a search algorithm relative to the counterfactual situation in which the true regressors are known and only sampling uncertainty remains. If the decision maker is faced with a non-zero cost of false discoveries (e.g. through spurious policy recommendations) or is generally interested in a parsimonious model (as implied by e.g. model selection criteria like AIC or BIC) it seems difficult to advocate this technique for model selection.

General-to-Specific

The size-power tradeoff of a high FDR and many correct rejections appears substantially more favorable for the general-to-specific selection procedures of HP and PcGets/Autometrics. We use the rather common significance level of 0.05 for the latter version. Clearly, both the FDR and the number of correct rejections would be smaller (larger) had we chosen a smaller (larger) significance level for the Autometrics search algorithm.

They identify most of the relevant variables at rather low FDRs, again in line with Hoover and Perez (2004). Still, however, the FDR can attain values of more than 30% for PcGets/Autometrics (Table 1.5), implying that the General-to-Specific approaches do not offer uniform FDR control. HP is more conservative, but hence also less powerful than

PcGets/Autometrics. Trying to compare power, we observe that, for a given FDR equal to that of an MTP, the power of the general-to-specific approaches seems comparable or somewhat higher than that of the MTPs. To see this, consider for $\gamma = 0.1$ the case of 25 relevant variables and $\rho = 0.3$. Here, both PcGets/Autometrics and the bootstrap have an FDR of .087 and around 22 correct rejections. Likewise, in the case of 10 relevant variables and $\rho = 0.3$ HP and BH result in around 9 vs. 7 correct rejections and an FDR of roughly .075.

Lasso

Choosing λ via fivefold cross-validation leads to quite high FDRs (FDRs are even slightly higher than the ones observed for classical testing at the 10% level, Table 1.5), but to a very large number of correct rejections, which is higher than the one of classical testing at the 10% level. Hence, if FDR control is the goal, Lasso should not be applied, at least not with this choice of λ. It furthermore seems to be the case that e.g. the General-to-Specific approaches are capable of achieving a similar number of correct rejections with a substantially lower FDR.

Overall, we find a robust size-power tradeoff between the different model selection approaches. Some methods appear to be practically dominated by others. For instance, Sala-i-Martin's approach achieves similar power to that of e.g. PcGets/Autometrics or Lasso only with a substantially higher FDR. Only the MTPs consistently control the FDR, at the expense of lower power than e.g. PcGets/Autometrics or some BMA variants even for comparable FDRs.

Robustness checks

In order to investigate the robustness of these findings, we additionally draw the correlation coefficients $\rho_{\ell\jmath}$, $\ell \neq \jmath$ randomly as $\rho_{\ell\jmath} \sim N(0, 1/9)$ and truncate if $|\rho_{\ell\jmath}| > 0.95$. We only consider correlation matrices whose maximal element exceeds 0.8 but discard a draw if the resulting matrix is not positive definite (pd). This is to mimic the empirically relevant scenario where some regressors are strongly positively correlated, and others negatively correlated. In fact, we thus roughly match the regressor correlation matrix from our empirical application in Section 1.4. The MTPs again achieve FDR control and similar power in this practically relevant scenario (cf. Figure 1.3 in the Appendix). In particular, this implies that the techniques also work in the presence of some highly correlated regressors.[9]

[9]To make this exercise practical, we choose $k = 10$ as it is very unlikely to draw a pd correlation matrix for larger k. We also experiment with correlations such as shifted Beta$(1/2, 1/2)$ variates $B_{\ell\jmath}$,

We furthermore also tested selected methods on the DGPs of Eicher, Papageorgiou, and Raftery (2011) and Krolzig and Hendry (2001). Overall, the findings are in line with the results presented here. Under the DGP of Eicher, Papageorgiou, and Raftery (2011) all studied MTPs identify all the relevant variables, indicating that their DGP does not discriminate well between the powers of the different variable selection techniques. Under the DGP of Krolzig and Hendry (2001) we find, in line with the above results, PcGets/Autometrics to be somewhat more powerful than the MTPs in designs resulting in similar size, at the expense of somewhat higher size overall, i.e., when not only inspecting instances with equal size. Detailed results are available in the Appendix A1.2.

1.4 Empirical growth models revisited

This section applies the above-mentioned model selection procedures to the well-known data set of Fernandez, Ley, and Steel (2001) (FLS), extracted from Sala-i-Martin (1997). We first provide a brief literature review and describe the data. We then present the results from the multiple testing procedures (MTPs). For small tolerated FDRs γ, these select initial GDP, Fraction Confucian and Fraction Hindu, the latter two being variables related to cultural and religious factors. We also compare these results with those from classical testing and the remaining model selection techniques. We shall see that the MTPs, PcGets/Autometrics and Lasso suggest fairly similar models, that differ somewhat from those proposed by BMA using the settings of FLS and Sala-i-Martin's approach. We also follow Fernandez, Ley, and Steel (2001) in calculating the predictive performance as measured by the log predictive score (LPS) and find most BMA variants to have a better predictive performance than the BH procedure. In line with neoclassical growth theory, initial GDP is robustly included by all procedures. The section concludes with a robustness check that verifies that similar results are found when controlling the 'false discovery proportion'—the fraction of false to all rejections—rather than the FDR.

Previous literature

Building on earlier work of e.g. Kormendi and Meguire (1985), Grier and Tullock (1989) and Barro (1991), Mankiw, Romer, and Weil (1992) popularized so-called cross-section

$\rho_{\ell j} = -0.1 + B_{\ell j}$. Then, $\mathrm{E}(B_{\ell j}) = 1/2$ and $\mathrm{Var}(B_{\ell j}) = 1/20$ and thus, the $\rho_{\ell j} \in [-0.1, 0.9]$. Qualitatively similar results, which show that FDR control also obtains under random correlation, are available in Tables 1.11 and 1.12 in the Appendix.

growth regressions, in which average growth over a sample period for different countries is regressed on a list of candidate explanatory variables. Mankiw, Romer, and Weil (1992) find a negative relation of growth to initial GDP and population growth, and a positive one to investment and level of schooling. Unfortunately, other authors find quite diverse components to determine economic growth (Durlauf, Kourtellos, and Tan, 2008). A far from exhaustive list includes initial GDP, fertility rates, high school enrollment rates, political instability and quality of institutions. Econometrically, this inconclusiveness arises because there is no single correct procedure to identify which of the many possible explanatory variables really are drivers of growth. Subsequent work therefore paid more attention to systematically tackling this inherent model uncertainty.

As an early example, Sala-i-Martin (1997) uses an approach aiming to assign a level of certainty to each variable (results are discussed further below). Fernandez, Ley, and Steel (2001) (FLS), Sala-i-Martin, Doppelhofer, and Miller (2004) and Eicher, Papageorgiou, and Raftery (2011) use BMA to account for the fact that growth theories are not mutually exclusive and that even if one knew the true theories, it would be unclear which variable to include for each. Sala-i-Martin, Doppelhofer, and Miller (2004) find 11 variables to robustly explain growth, of which initial GDP again has the strongest impact.[10]

Results using multiple testing procedures

We now apply the MTPs from Section 1.2 to the widely employed data set of FLS.[11] The latter comprises $n = 140$ countries for which growth is computed from 1960 to 1992, and $k = 62$ regressors. FLS only use those countries for which observations on all 25 variables flagged important by Sala-i-Martin (1997) are available, yielding $n = 72$. They then add all variables which do not induce any missing observations in these countries. This yields $k = 41$. We use this data to compare model selection using MTPs to the findings of FLS, Sala-i-Martin (1997), Hendry and Krolzig (2004), Ley and Steel (2009) and Schneider and

[10]In early work, Levine and Renelt (1992) apply extreme bounds analysis to select variables. As this method is quite severe, they only find initial GDP, the investment rate, secondary school enrollment and population growth to have explanatory power. More recently, Magnus, Powell, and Prüfer (2010) employ a related, but computationally less expensive, weighted average least squares approach that splits regressors into 'focus' regressors which are to be in the model on theoretical grounds, and auxiliary regressors whose relevance is less certain. The results are similar to those of BMA (discussed further below).

[11]These data are available at http://econ.queensu.ca/jae/. We also performed an analogous exercise for the data set first used in Masanjala and Papageorgiou (2006), and report the qualitatively similar results in the Appendix A1.4. We also investigated the data set of Sala-i-Martin, Doppelhofer, and Miller (2004) (a very similar data set was used in Magnus, Powell, and Prüfer, 2010), but only found one variable ('Investment price') to be significant at the 5% level using classical testing. Controlling the FDR at any level, no variable was found to explain growth.

Table 1.6: Results for the FLS/Sala-i-Martin (1997) data set

	Regressor	$\hat{\beta}_\ell$	*p-value*	*Classical*	*BH*	*Boot*	*BMA*	*S-i-M*	*H&K*	*Lasso*	*FDP*
1	GDP level 1960	−0.017	0.00001	1%	1%	1%	1.000	1.00**	✓	✓	1%
2	Fraction Confucian	0.075	0.00003	1%	1%	1%	0.995	1.00*	✓	✓	1%
3	Life expectancy	0.001	0.003	1%	5%	5%	0.946	.999**	✓	✓	20%
4	Equipment investment	0.127	0.008	1%	5%	5%	0.942	1.00*	✓	✓	-
5	Sub-Saharan dummy	−0.02	0.006	1%	5%	5%	0.757	.997*	✓	✓	-
6	Fraction Muslim	0.011	0.227	-	-	-	0.656	1.00*	-	✓	-
7	Rule of law	0.012	0.068	10%	-	-	0.516	1.00*	-	✓	-
8	Number of years open economy	−0.003	0.620	-	-	-	0.502	1.00*	✓	-	-
9	Degree of capitalism	0.001	0.284	-	-	-	0.471	.987*	-	-	-
10	Fraction Protestant	−0.003	0.677	-	-	-	0.461	.966*	-	-	-
11	Fraction GDP in mining	0.04	0.008	1%	5%	5%	0.441	.994*	✓	✓	-
12	Non-Equipment investment	0.037	0.081	10%	-	-	0.431	.982*	-	✓	-
13	Latin American dummy	−0.013	0.039	5%	-	10%	0.190	.998*	✓	✓	-
14	Primary school enrollment, 1960	0.02	0.045	5%	-	10%	0.184	.992**	✓	✓	-
15	Fraction Buddhist	0.007	0.276	-	-	-	0.167	.964*	-	-	-
16	Black-market premium	−0.007	0.075	10%	-	-	0.157	.825	-	-	-
17	Fraction Catholic	0.003	0.593	-	-	-	0.110	.963*	-	-	-
18	Civil liberties	−0.002	0.321	-	-	-	0.100	.997*	-	-	-
19	Fraction Hindu	−0.097	0.001	1%	1%	1%	0.097	.654	✓	✓	5%
20	Political rights	0.0002	0.934	-	-	-	0.071	.998*	-	-	-
21	Primary exports, 1970	−0.006	0.421	-	-	-	0.069	.990*	-	-	-
22	Exchange rate distortions	−0.00002	0.538	-	-	-	0.060	.968*	-	-	-
23	Age	−0.00001	0.774	-	-	-	0.058	.903	-	-	-
24	War dummy	−0.001	0.548	-	-	-	0.052	.984*	-	-	-
25	Size labor force	3.e-07	0.004	1%	5%	5%	0.047	.835	✓	✓	-
26	Fraction speaking foreign language	−0.002	0.468	-	-	-	0.047	.831	-	-	-
27	Fraction of pop speaking English	−0.007	0.131	-	-	-	0.047	.910	-	-	-
28	Ethnologic fractionalization	0.014	0.012	5%	5%	5%	0.035	.643	✓	✓	-
29	Spanish colony dummy	0.013	0.022	5%	10%	10%	0.034	.938*	✓	-	-
30	SD of black-market premium	−0.000001	0.892	-	-	-	0.031	.993*	-	-	-
31	French colony dummy	0.009	0.038	5%	-	10%	0.031	.702	✓	-	-
32	Absolute latitude	−0.0001	0.521	-	-	-	0.024	.980*	-	-	-
33	Ratio of workers to population	−0.001	0.945	-	-	-	0.024	.766	-	-	-
34	Higher education enrollment	−0.129	0.002	1%	5%	5%	0.024	.579	✓	✓	20%
35	Population growth	−0.119	0.609	-	-	-	0.022	.807	-	-	-
36	British colony dummy	0.007	0.072	10%	-	-	0.022	.579	✓	-	-
37	Outward orientation	−0.005	0.036	5%	-	10%	0.021	.634	-	-	-
38	Fraction Jewish	−0.001	0.942	-	-	-	0.019	.747	-	-	-
39	Revolutions and coups	0.003	0.503	-	-	-	0.017	.995*	-	-	-
40	Public education share	0.137	0.249	-	-	-	0.016	.580	-	-	-
41	Area (scale effect)	3.e-07	0.637	-	-	-	0.016	.532	-	-	-
42	Intercept	0.0207	0.000								

Notes: Results are sorted on the BMA column. 'Classical' reports if a standard t-test rejects at level $\alpha = \{1\%, 5\%, 10\%\}$ for each variable. Columns 'BH' and 'Boot' show if the variables are found significant when controlling the FDR at the levels $\gamma = \{1\%, 5\%, 10\%\}$. The procedures are described in Section 1.2. 5,000 bootstrap iterations. 'BMA' denotes the marginal posterior probability of inclusion from Fernandez, Ley, and Steel (2001); 'S-i-M' shows the inclusion frequency from Sala-i-Martin (1997), where $**$ indicate variables always included and $*$ indicate variables found significantly related to growth; in the columns 'H&K' and 'Lasso' a ✓ denotes that a variable is included by the procedures of Hendry and Krolzig (2004) and the adaptive Lasso procedure from Schneider and Wagner (2012). FDP denotes values for τ at which a variable is found to be significant when controlling FDP in the sense that $P\{FDP > \nu\} \leq \tau$, where $\nu = 0.05$ and $\tau = \{1\%, 5\%, 10\%, 20\%\}$ is considered; cf. the robustness check at the end of Section 1.4.

Wagner (2012). To ensure comparability with these alternative model selection strategies, we estimate (1.2) by OLS. BH runs in fractions of a second. The bootstrap procedure is still quite fast and produces results within a few minutes.

Table 1.6 shows the results for the FLS/Sala-i-Martin (1997) data.[12] Classical testing declares 19 variables (excluding the intercept) significant at the 10% level, 15 at the 5% level and 9 at the 1% significance level. The MTPs draw a different picture. The more powerful procedure according to the Monte Carlo study, the bootstrap method, declares

[12]The results are based on usual OLS standard errors. We also calculate results using the heteroscedasticity robust standard errors HC_2 and HC_3 (MacKinnon and White, 1985). Results using HC_2 are similar to the OLS results; using HC_3 resulted in substantially fewer rejections. In the bootstrap approach we use the wild bootstrap as heteroscedasticity robust method. Out of the 15 variables which are significant at a 10% FDR rate when not accounting for possible heteroscedasticity we can confirm 11 using the wild bootstrap. Performing a White test for heteroscedasticity is pointless here, as the null would never be rejected because of the high number of variables in the model. The 5% critical value of the 800 degrees of freedom test is 866.9, whereas $n \cdot R^2$ of the test regression is bounded above by 72. Refer to Tables 1.20, 1.21 and 1.22 in the Appendix for details.

15 variables significant at $\gamma = 0.1$, and 10 at $\gamma = 0.05$. At $\gamma = 0.01$, only 3 variables are found to be significantly related to growth. We elaborate on these variables below. For illustration, see Figure 1.1, which plots the sorted statistics $T_{(j)}$, the $\hat{c}_{j,\gamma}$ as well as the implied cutoffs of BH (when expressed at the scale of test statistics rather than that of p-values). As expected from the identical results (cf. Table 1.6) the critical values are similar for $\gamma = 0.01, 0.05$. The power advantage of the bootstrap manifests itself for $\gamma = 0.1$, where the smaller $\hat{c}_{j,0.1}$ yield four additional variables to be included (viz. Latin American and French colony dummies, Primary school enrollment in 1960 and Outward orientation).[13]

This comparison reveals that classical testing may spuriously declare several variables to be significant because of the large number of tests performed. Additional simulations (cf. Table 1.9 in the Appendix) as in Section 1.3 with 10 (out of 50) false hypotheses find an FDR of classical testing of around 1/3. To the extent that the MC study of Section 1.3 is representative for the present application, we should expect roughly 1 false for every 3 rejections when testing at $\alpha = 0.1$. This would mean that we should expect around $13 (= 19 - 6)$ of the 19 rejections to be correct. This is in line with the bootstrap results, where around $(1 - 0.1) \times 15 \approx$ 13-14 of the rejections can be expected to be correct. Of course, such corrective reasoning for classical testing can only be applied if the Monte Carlo design accurately mirrors the data structure of the empirical application. A key advantage of multiple testing procedures is that FDR control is obtained under more general conditions, as evidenced by our and many other simulation studies and theoretical contributions.

The three variables most robustly found to influence growth, namely, the ones significant at $\gamma = 0.01$, are GDP level 1960, Fraction Confucian and Fraction Hindu (Table 1.6). The rationale for the impact of initial GDP on growth is well established since the work of Mankiw, Romer, and Weil (1992). The two other variables capture heterogeneity related to culture and religion.

Comparison to previous studies using FLS data

We now compare the MTP results with those of the other popular model selection strategies reviewed in Section 1.2.

PcGets/Autometrics

The similarity between multiple testing and general-to-specific also becomes visible when

[13]Results using the procedures of Storey, Taylor, and Siegmund (2004) and Benjamini, Krieger, and Yekutieli (2006) can be found in Table 1.19 in the Appendix.

comparing the variables found significant (Table 1.6).[14] Hendry and Krolzig (2004) find 16 significant variables, compared to 15 for the bootstrap approach at $\gamma = 0.1$. Of these, 14 coincide. Only Outward orientation is included by the bootstrap but not by PcGets/Autometrics, whereas the latter includes the Number of years as an open economy, unlike the bootstrap method.[15] It may seem wise not to overstate even these small differences between the two models. Outward orientation and Number of years as an open economy are variables that rather readily substitute for each other, that is, exhibit *negative jointness* in the terminology of Doppelhofer and Weeks (2009). Hence, the final specifications are likely to be even more similar to each other than the overlap of 14 significant variables suggests.

To the extent that our Monte Carlo study is informative about cross-section growth data sets, this result is plausible: consider the column $\rho = 0.5$, which arguably comes closest to real-world growth data sets. In e.g. Table 1.3, PcGets/Autometrics has a higher power (at the expense of a higher FDR) than the MTPs and should hence be expected to produce slightly more rejections.

Bayesian model averaging

Table 1.6 shows that BMA using the settings of FLS yields eight variables with a posterior inclusion probability above the threshold of 50% employed here. This roughly compares with the 10 significant variables using the MTPs at $\gamma = 0.05$. Again, these results are plausible given the findings of our Monte Carlo simulations: consider once more column $\rho = 0.5$ and Table 1.3. BMA using the FLS settings ($m = k/2$, $g = k^{-2}$, fixed θ), BH and the bootstrap method find around 10 relevant variables, with the MTPs being somewhat more powerful in this particular scenario. The five variables with the highest marginal posterior probability in FLS are also significant according to the bootstrap method at $\gamma = 0.05$.

Again, some of the apparent differences in results between the methods are likely driven by jointness effects. For instance, specific religion variables are often assigned some importance by one method but not another. E.g., Fraction Hindu is significant at $\gamma = 0.01$ while its marginal posterior probability is only 0.097. Conversely, the fraction of Muslims has a rather high posterior inclusion probability, but is not included by any of the MTPs nor PcGets/Autometrics. This suggests that the relevance of a variable should not only be

[14] Hoover and Perez (2004) also apply their variant of Gets to growth data. The dataset they work however differs somewhat from the one used in FLS, and hence we prefer to compare the MTP procedures to the results of Hendry and Krolzig (2004) who use the same data as FLS.

[15] A J-test (Davidson and MacKinnon, 1981) rejects both models, although Hendry and Krolzig's model is rejected only at larger significance levels.

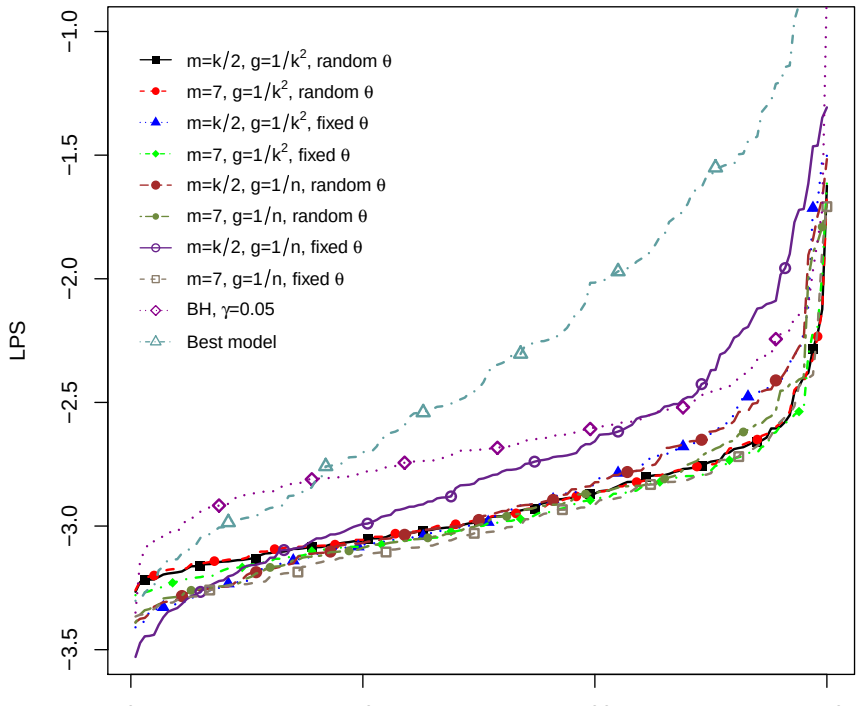


Figure 1.2: Log predictive scores

Note: Sorted log predictive scores for the eight priors from Ley and Steel (2009), BH and the Best model using the FLS prior settings. 150 subsamples.

analyzed marginally, but also jointly with that of other, related variables that may either complement or substitute for that variable. Consequently, the approach of Doppelhofer and Weeks (2009) may enrich the lessons that can be drawn from the marginal view inherent in both the BMA variant analyzed here and the MTPs.

Regarding the robustness of BMA results, Ley and Steel (2009) show that for the eight prior choices considered in their paper (cf. Section 1.2), the posterior mean model size ranges from 6.03 with $m = k/2 = 20.5$, random θ and $g = 1/k^2$ to 19.84 for $m = k/2 = 20.5$, fixed θ and $g = 1/n$. Their Table II shows the ranking of the marginal posterior probability of including a certain variable to also be highly sensitive to the prior settings. Comparing the prior choice $g = 1/n$, fixed θ and $m = 20.5$ with that of $g = 1/n$, fixed θ and $m = 7$ they note: 'Fraction Hindu, the Labor force size, and Higher education enrollment go from virtually always included with $m = 20.5$ to virtually never included with $m = 7$.'

Similarly, recent work by Eicher, Papageorgiou, and Raftery (2011) shows that some alternative 'default' priors can lead to rather different growth models using the FLS data, with 'as few as three and as many as 22 regressors' being found to influence growth. They recommend a unit information prior which is very closely related to the UIP discussed in Section 1.3, cf. their Table I. It will be interesting to see whether the BMA literature will

Table 1.7: Log predictive scores

	mean(LPS)	*max(LPS)*	*min(LPS)*
BH	−2.66	−0.81	−3.35
Bayesian Model Averaging			
$m = k/2$, $g = 1/k^2$, random θ	−2.90	−2.16	−3.21
$m = 7$, $g = 1/k^2$, random θ	−2.90	−2.12	−3.21
$m = k/2$, $g = 1/k^2$, fixed θ	−2.92	−0.99	−3.48
$m = 7$, $g = 1/k^2$, fixed θ	−2.94	−2.09	−3.25
$m = k/2$, $g = 1/n$, random θ	−2.94	−1.40	−3.44
$m = 7$, $g = 1/n$, random θ	−2.96	−1.39	−3.46
$m = k/2$, $g = 1/n$, fixed θ	−2.80	−1.43	−3.50
$m = 7$, $g = 1/n$, fixed θ	−2.98	−1.56	−3.41
Best Model	−2.31	0.11	−3.33

Notes: LPS scores are calculated using the FLS data, using 75% of the data (i.e. 54 observations) as a training sample, and the remainder of $n = 72$ as a holdout sample. 150 subsamples. The BMA variants are those considered by Ley and Steel (2009), the best model uses the settings of FLS.

henceforth adopt this choice, or whether different models continue to be put forward using different variants of BMA.[16]

These findings might help explain the differences between the MTPs and the marginal posterior inclusion probabilities of FLS, as well as the differences between the latter and the results of the other model selection procedures (see below). The findings of Ley and Steel (2009) and Eicher, Papageorgiou, and Raftery (2011) imply that the robustness of BMA must be interpreted with care.

We also follow FLS in calculating the predictive performance of the BH procedure as well as of all BMA procedures considered by Ley and Steel (2009). The procedures' predictive performance is measured by their log predictive scores, a statistic that increases in both lack of predictive fit and sampling uncertainty. We employ the R (R Core Team, 2012) package BMS of Feldkircher and Zeugner (2009) and follow the design of FLS. That is, we randomly split the $n = 72$ observations into a training (or 'inference') subsample of size $0.75 \cdot 72 = 54$ and a holdout (or 'prediction') subsample of size 18. Figure 1.2 reports results for 150 subsamples. Table 1.7 gives the corresponding minimum, mean and maximum LPS.

For BH at $\gamma = 0.05$, we find a minimum, mean and maximum LPS of -3.35, -2.66 and -0.81 over 150 subsamples. These values are higher, hence worse, than for instance those for the BMA prior settings of FLS ($m = k/2$, $g = 1/k^2$, fixed θ), i.e. -3.48, -2.92 and -0.99. This reflects, as is also known from the forecast combination literature, that

[16] As with the MTPs, our discussion of alternative variants of BMA is constrained by space considerations and focuses on those that we believe are most prominent in the literature. Other recent proposals include Liang et al. (2008), Feldkircher and Zeugner (2009) or Crespo Cuaresma (2011).

using evidence from multiple models tends to improve out of sample performance. That said, the differences seem to be modest.[17] In general, the different BMA settings have quite similar LPS in the center of the distributions, and hence also similar mean LPS. Thus, all of these have better mean LPS than BH. The different BMA settings however lead to rather different best and worst LPS. The cases $m = k/2$, $g = 1/k^2$, random θ as well as $m = 7$, $g = 1/k^2$, random θ for instance have worse best-case LPS than BH. On the other hand, all BMA settings lead to better worst-case LPS than BH. BH seems to be more competitive with the BMA procedures when bad LPS are considered than when looking at favorable ones.

Overall, the distance between BH and the BMA variants appears to be modest when compared to the variance of the LPS of the different procedures over the different subsamples: the predictive performance of an average BH model is much better than that of a poor BMA model. Hence, on average we expect BMA to predict more accurately, but there is no guarantee that this also holds true for any given sample one uses for prediction in practice. Finally, the BH procedure (as do the BMA variants) outperforms the best model, i.e., the one with the highest posterior probability, of the BMA exercise using the FLS settings.

Sala-i-Martin

In addition to the varying selection of three control variables, Sala-i-Martin (1997) imposes inclusion of three more variables deemed to be important by default—GDP level in 1960, Life expectancy and Primary school enrollment—in his empirical application. He finds 22 significant variables, but *assumes* relevance of the three default variables. There are five variables found significant by Sala-i-Martin (1997) (including defaults)—GDP level in 1960, Fraction Confucian, Life expectancy, Equipment investment, Sub-Saharan Dummy—that can be confirmed by the bootstrap approach ($\gamma = 0.05$), FLS and PcGets/Autometrics. Beyond that, there is little agreement with the MTPs. Using a tolerated FDR up to 10%, we can only confirm 9 of his 25 significant variables. In light of our Monte Carlo results, it is not implausible to interpret this high number of rejections as resulting from a high FDR of Sala-i-Martin's approach.

Lasso

Schneider and Wagner's (2012) model includes 15 variables, of which 12 coincide with those identified by the bootstrap method at $\gamma = 0.1$. These 12 variables are also among the 16 selected in Hendry and Krolzig (2004). The three additional variables that are selected by

[17]The values for the FLS settings are a bit larger than those reported by FLS. This suggests that the 20 subsamples drawn by FLS may have happened to be rather favorable to prediction.

Lasso are Fraction Muslim, Rule of law and Non-Equipment investment. The Lasso does not include the number of years as an open economy and the Spanish, French and British colony dummies. Of these, only the Spanish dummy is also included by both BH and the bootstrap. Overall, this indicates some robustness concerning the significance of the 12 variables that are selected by the bootstrap, PcGets/Autometrics and Lasso. Interestingly, some of these 12 variables have very low marginal posterior probabilities when BMA is used. The Lasso however agrees with BMA in including the fraction of Muslims.

Overall, there are five variables jointly significant in FLS, Sala-i-Martin (1997), Hendry and Krolzig (2004), with the Lasso and for the MTPs at $\gamma \geq 0.05$: GDP level 1960, Fraction Confucian, Life expectancy, Equipment investment and the Sub-Saharan dummy. Hence, the relevance of these variables appears quite robust. These variables mostly have a plausible economic or cultural and religious motivation.

Robustness check: Controlling the false discovery proportion

The FDR is defined as the expected value of the false discovery proportion (FDP), i.e., the expected value of the number of false over total rejections. As pointed out in Romano, Shaikh, and Wolf (2008b), even when the expected value of the FDP is controlled at level γ, the realized value of the FDP can lie above γ with possibly high probability. Therefore, as a robustness check for our results, we also employ techniques controlling the FDP. In particular, we aim to ensure that $P\{FDP > \nu\} \leq \tau$. That is, the probability of the fraction of false to all rejections exceeding ν is to be no more than τ. We use the procedure to control the FDP introduced by Romano and Shaikh (2006), which is a step-down procedure based on individual p-values of each test statistic.[18] The p-values are ordered from smallest to largest and are compared against step-down constants o_j. Starting with $\widehat{p}_{(1)}$, the procedure rejects and steps down to the next p-value as long as $\widehat{p}_{(j)} \leq o_j$. It stops rejecting hypotheses and accepts the remaining ones as soon as $\widehat{p}_{(j)} > o_j$. Lehmann and Romano (2005) propose to use $o_j = \frac{(\lfloor \nu \cdot j \rfloor + 1)\tau}{k + \lfloor \nu \cdot j \rfloor + 1 - j}$. For this choice of o_j to be applicable some assumptions about the joint dependence of the p-values have to be made (Lehmann and Romano, 2005). Hence, Romano and Shaikh (2006) introduce a constant D with which one divides the o_j. In our case of $k = 41$ tests and $\nu = 0.05$, $D \approx 1.48$. This then guarantees control of the FDP

[18] We could have alternatively used the bootstrap based FDP controlling procedure introduced in Romano, Shaikh, and Wolf (2008b). In view of the moderate differences between p-value- and bootstrap-based FDR controlling procedures found above, we use the computationally less expensive p-value approach in this robustness check.

under arbitrary dependence of the p-values. The procedure is thus suitable for our empirical application.

In general, control of the FDR and the FDP is linked and Romano and Shaikh (2006) derive connections between procedures controlling either of the two. A procedure that controls the FDR at level γ also controls the FDP in the sense that $P\{FDP > \nu\} \leq \gamma/\nu$. Moreover, if the FDP is controlled in the sense that $P\{FDP > \nu\} \leq \tau$, the FDR is controlled at level $\tau(1-\nu)+\nu$. The first approximation provides valuable insights for our empirical application. Consider controlling the FDR at the 1% level. At this level, the probability of obtaining a FDP of above 10% is lower than 10%. In the empirical application we select three variables when controlling the FDR at the 1% level. If one of the three hypotheses were falsely rejected, this would result in an FDP of 1/3. Given that the probability of obtaining an FDP of above 10% is no more than 10%, we view the selection of the three variables as extremely robust.

To gain insights into how control of the FDP is linked to the FDR we study FDP control in our main Monte Carlo setup as described in Section 1.3. We take $\nu = 0.05$ and $\tau \in \{0.01, 0.05, 0.1\}$. Detailed results are available in Table 1.14 in the Appendix. We find that the realized $P\{FDP > \nu\}$ is always smaller than the required value τ and thus the Romano and Shaikh (2006) procedure controls the FDP in the desired sense. FDP control is achieved for all degrees of dependence among the regressors. Unsurprisingly, the number of correct rejections increases with τ, as the control becomes less strict the higher τ. Correct rejections decrease with higher correlations among the regressors. The FDR is controlled at quite low levels, i.e., at levels substantially lower than the upper bound $\tau(1-\nu)+\nu$ would allow for (here the upper bounds are 5.95%, 9.75% and 14.5% for τ equal to 1%, 5% and 10%, respectively). The number of correct rejections roughly compares to the number of correct rejections when controlling the FDR at 1% for $\tau = 0.05$ and to controlling the FDR at 5% for $\tau = 0.1$.

We next return to the above empirical application. We again aim to control the FDP in the sense that $P\{FDP > \nu\} \leq \tau$. We set $\nu = 0.05$ and take 1%, 5%, 10% and 20% as possible values for τ. As can be seen in the last column of Table 1.6, for $\tau = 0.2$, 5 variables are found to explain economic growth; 3 variables are found for τ equal to 10% or 5% and only two variables are found when $\tau = 0.01$. Controlling $P\{FDP > 0.05\}$ at the 5% level, we can corroborate the significance of the three variables which were already found to be significant by the MTPs at the 1% level. This result is not entirely surprising given the

connection of FDR and FDP control described above. Next to these three variables, Life expectancy and Higher education enrollment are additionally found to be significant when employing $\tau = 0.2$. With respect to the set of five robust variables found above, FDP control does not lead to inclusion of Equipment investment and the Sub-Saharan dummy. Overall, we conclude that the main conclusions from our empirical application are robust to using the FDP rather than the FDR as a measure of multiple type I error rate to control.

1.5 Conclusion

In this paper, we have shown how techniques from the statistical multiple testing literature can usefully be applied to economic data. Concretely, controlling the false discovery rate, we employ multiple testing procedures (MTPs) to select explanatory variables in for instance cross-section growth regressions in a simple and fast way, requiring only the p-values of for instance the test statistics of a long regression. Alternatively, the exercise can be performed using a bootstrap approach which takes the dependence structure of the test statistics into account.

We compare the MTPs to other popular variable selection approaches, viz. PcGets/ Autometrics, Bayesian Model Averaging (BMA), the 'two million regressions approach' of Sala-i-Martin (1997) and the Lasso. This is done using simulations and via empirical cross-section growth regressions. We compare the power of the approaches in situations in which methods have the same size or look at constellations in which two procedures have very similar power, but differ in the required size to achieve that power. Using these benchmarks, some approaches are practically dominated by others. The bootstrap method is most powerful among the MTPs given the same size requirement. The approach of Sala-i-Martin (1997) is dominated, for example, by the General-to-Specific/Autometrics approach. The General-to-Specific/Autometrics and some variants of BMA seem to be a bit more powerful than the MTPs. Another important finding is that the MTPs are the only model selection procedure to consistently control the FDR. To the extent that the FDR is agreed to be a useful multiple type I error rate, MTPs would appear to be the only variable selection procedures which yield consistent size control. For single hypothesis tests, this is widely considered to be a desirable property.

The variables jointly selected by all model selection procedures in the empirical application mostly have a plausible economic or cultural and religious motivation. One of the

variables selected by the MTPs at low significance levels in our application to growth data—initial GDP—supports neoclassical growth theory and its implied convergence of countries to a steady-state output level.

Controlling the false discovery rate is important in more literatures in applied econometrics. Closely related to the present application is the question of whether output time series of different countries converge over time, i.e., whether their output gap contains neither unit root nor deterministic trend. When studying n countries, there are $n(n-1)/2$ pairs to consider, such that a non-negligible number of pairs will be found to be convergent with individual tests even if they are not (Deckers and Hanck, 2013). Furthermore, modeling returns to schooling or some forecasting exercises involve a large number of candidate explanatory variables. As such, the techniques studied here may prove fruitful there, too.

A1 Appendix to Chapter 1

A1.1 Bootstrap procedure

We use the following semi-parametric bootstrap:

1. Estimate $\boldsymbol{\beta}$ using $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and calculate the residuals $\hat{\boldsymbol{u}}$ using $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. (In regressions without constant, get demeaned residuals $\tilde{\boldsymbol{u}}$. In regressions with constant, assume w.l.o.g. that the constant is in column $k+1$ of the $n \times (k+1)$ matrix $\boldsymbol{X}$.)
2. For each element of $\boldsymbol{\beta}$ corresponding to a non-constant element of $\boldsymbol{X}$, calculate the t-statistic $T_j = \hat{\beta}_j / (\widehat{\mathrm{Var}}(\hat{\beta}_j))^{1/2}$ for $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$, $j = 1, \ldots, k$. Here $\widehat{\mathrm{Var}}(\hat{\beta}_j) = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}$ and $s^2 = \sum_i \hat{u}_i^2/(n-k-1)$.
3. Resample non-parametrically with replacement from $\hat{\boldsymbol{u}}$ (or $\tilde{\boldsymbol{u}}$ if necessary) to obtain the bootstrap errors u_i^* and build the bootstrap sample $y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + u_i^*$. Under heteroscedasticity it is recommended to use the wild bootstrap which was introduced by Wu (1986).[19] The bootstrap sample is built as $y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + f(\hat{u}_i)v_i^*$, where $f(\hat{u}_i) = \hat{u}_i/(1-h_i)^{1/2}$, where h_i is the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, and v_i^* is chosen as follows:

$$v_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability} \quad (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability} \quad (\sqrt{5}-1)/(2\sqrt{5}) \end{cases}$$

4. Calculate $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^*$ and $\boldsymbol{u}^* = \boldsymbol{y}^* - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*$.
5. For each element of $\hat{\boldsymbol{\beta}}^*$ corresponding to a non-constant element of $\boldsymbol{X}$, construct the bootstrapped version of each individual t-statistic using $T_j^* = (\hat{\beta}_j^* - \hat{\beta}_j)/(\widehat{\mathrm{Var}}(\hat{\beta}_j^*))^{1/2}$, where $\widehat{\mathrm{Var}}(\hat{\beta}_j^*) = s^{2*}(\boldsymbol{X}'\boldsymbol{X})^{-1}_{jj}$ and $s^{2*} = \sum_i u_i^{*2}/(n-k-1)$. Repeat steps 3 – 5 B times.
6. Given $\hat{P}$, the critical values are defined recursively as follows: Having determined $\hat{c}_1, \ldots, \hat{c}_{j-1}$, the jth critical value is determined using the minimization rule (Romano, Shaikh, and Wolf, 2008a):

[19] An alternative would be the pairs bootstrap. We find, however, that the pairs bootstrap does not perform well in the present setting of many regressors in terms of power and size using HC2 or HC3 standard errors, respectively. Refer to Table 1.13 in the next section for details.

$$\hat{c}_j = \inf \Bigg\{ c \in \mathbb{R} : \sum_{k-j+1 \leq r \leq k} \frac{r-k+j}{r} \times \hat{P} \left\{ T^*_{j:j} \geq c, \ldots, T^*_{k-r+1:j} \geq \hat{c}_{k-r+1}, T^*_{k-r:j} < \hat{c}_{k-r} \right\} \leq \gamma \Bigg\} \tag{1.7}$$

(Note the meaning of meaning of $T^*_{r:t}$. The index t stems from the ordering of the original statistics, whereas r corresponds to the bootstrapped statistics. So $T^*_{r:t}$ has the following meaning: Out of the t smallest original statistics pick the rth smallest of the corresponding bootstrap statistics.)

7. Use the $\hat{c}_j$ from 6 and compare them step-down to the t-statistics from step 2.

We thus bootstrap by estimating the original model under the alternative and calculating the bootstrap t-statistic accordingly. This leads to a large gain in computation time, since one only needs to rebuild the DGP once and not k times. It further preserves the dependence of the statistics in each iteration. This is important, since when applying (1.7), one makes statements about the joint distribution of the statistics, and these are (possibly strongly) correlated through the correlation of the regressors. It is therefore crucial for the success of the bootstrap that the resampling procedure preserves the dependence in the statistics.

A1.2 Additional simulation results

The present additional tables also contain complementary results for the multiple testing procedures of Storey, Taylor, and Siegmund (2004) and Benjamini, Krieger, and Yekutieli (2006), which are briefly reviewed here. The results are qualitatively very similar to the Bootstrap and BH results whence we do not include them in the main text.

Storey method

The BH method is conservative: it can be shown that FDR $\leq k_0\gamma/k$. So, unless $k_0 = k$, the power of BH can be improved by redefining $\gamma_j = j\gamma/k_0$. Storey, Taylor, and Siegmund (2004) estimate k_0 by $\widehat{k}_0 = (\#[\widehat{p}_j > \lambda] + 1)/(1 - \lambda)$, where $\lambda \in (0, 1)$ is user-specified. The idea behind this estimator is that p-values of true null hypotheses are approximately uniform$[0, 1]$ distributed. Therefore $k_0(1 - \lambda)$ of these should lie in the interval $(\lambda, 1]$. Storey, Taylor, and Siegmund (2004) argue that this procedure typically controls the FDR whenever BH does.[20] It can however be quite liberal for positively dependent p-values.

BKY algorithm

Benjamini, Krieger, and Yekutieli (2006) propose a two step improvement of BH. First, apply the BH procedure at $\gamma^* = \gamma/(1 + \gamma)$. If no, or all, hypotheses are rejected the procedure stops. If $0 < r < k$ hypotheses are rejected, continue using BH replacing γ_j by $j\gamma^*/(k - r)$. Benjamini, Krieger, and Yekutieli (2006) prove FDR control for independent statistics and also provide encouraging simulations under dependence.

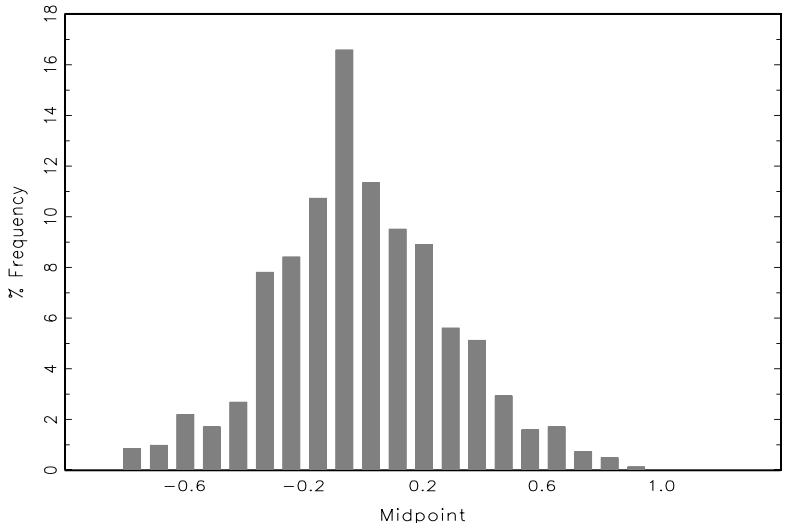


Figure 1.3: Histogram of the Regressor Correlation Matrix of Section 1.4

[20] Storey, Taylor, and Siegmund (2004) also suggest to estimate the FDR after a sequence of tests at a fixed significance level. This has the Bayesian interpretation of the probability of a given rejection coming from a true null hypothesis.

Table 1.8: Linear regression model with 5 false hypotheses

	# False hypotheses: 5		*Sample size: 100*		*# Regressors: 50*	
	$\rho = 0$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.083	3.96	0.273	4.61	0.426	4.81
Bonferroni	0.003	1.76	0.010	2.57	0.021	3.03
BH	0.010	2.27	0.038	3.35	0.086	3.81
Storey	0.013	2.32	0.049	3.40	0.102	3.84
BKY	0.010	2.29	0.040	3.36	0.086	3.79
Bootstrap	0.011	2.36	0.052	3.45	0.100	3.94
	$\rho = 0.3$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.097	3.08	0.302	4.12	0.457	4.49
Bonferroni	0.003	0.94	0.014	1.65	0.031	2.06
BH	0.007	1.21	0.043	2.29	0.092	2.85
Storey	0.008	1.26	0.051	2.35	0.113	2.91
BKY	0.007	1.22	0.043	2.28	0.090	2.80
Bootstrap	0.009	1.22	0.050	2.39	0.102	2.95
	$\rho = 0.5$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.125	2.250	0.336	3.48	0.473	3.98
Bonferroni	0.006	0.457	0.02	0.96	0.035	1.27
BH	0.009	0.547	0.042	1.32	0.080	1.80
Storey ($\lambda = 0.5$).	0.011	0.594	0.050	1.35	0.096	1.87
BKY	0.009	0.549	0.042	1.30	0.078	1.75
Bootstrap	0.010	0.592	0.048	1.35	0.092	1.88

Notes: This table shows the results of a Monte Carlo simulation as specified in Section 1.3 with 2,000 replications and the indicated parameter settings when controlling the FDR at $\gamma = \{1\%, 5\%, 10\%\}$. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

Table 1.9: Linear regression model with 10 false hypotheses

	# False hypotheses: 10		*Sample size: 100*		*# Regressors: 50*	
	$\rho = 0$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.039	7.94	0.156	9.26	0.268	9.61
Bonferroni	0.003	3.48	0.005	5.28	0.010	6.06
BH	0.008	5.39	0.040	7.68	0.079	8.50
Storey	0.011	5.65	0.053	7.88	0.101	8.66
BKY	0.009	5.52	0.045	7.78	0.087	8.57
Bootstrap	0.011	5.50	0.050	7.89	0.101	8.67
	$\rho = 0.3$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.047	6.19	0.176	8.27	0.284	8.95
Bonferroni	0.003	1.87	0.008	3.34	0.014	4.11
BH	0.008	2.96	0.038	5.45	0.076	6.77
Storey	0.011	3.22	0.053	5.76	0.104	7.05
BKY	0.009	3.03	0.042	5.54	0.083	6.83
Bootstrap	0.010	3.10	0.046	5.71	0.096	7.03
	$\rho = 0.5$					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.062	4.57	0.199	6.90	0.299	8.00
Bonferroni	0.003	0.94	0.010	1.96	0.019	2.51
BH	0.006	1.41	0.036	3.27	0.072	4.63
Storey ($\lambda = 0.5$)	0.009	1.61	0.050	3.57	0.097	4.97
BKY	0.006	1.42	0.037	3.30	0.073	4.62
Bootstrap	0.008	1.52	0.044	3.42	0.091	4.74

Notes: This table shows the results of a Monte Carlo simulation as specified in Section 1.3 with 2,000 replications and the indicated parameter settings. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

Table 1.10: Linear regression model with 25 false hypotheses

	# False hypotheses: 25		*Sample size: 100*		*# Regressors: 50*	
			$\rho = 0$			
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.011	19.83	0.048	23.11	0.090	24.01
Bonferroni	0.0003	8.79	0.001	13.2	0.003	15.2
BH	0.005	16.67	0.025	21.61	0.050	23.07
Storey	0.001	18.69	0.051	22.86	0.100	23.93
BKY	0.007	17.88	0.043	22.63	0.087	23.81
Bootstrap	0.008	18.15	0.046	22.80	0.095	23.93
			$\rho = 0.3$			
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.013	15.39	0.054	20.61	0.098	22.38
Bonferroni	0.004	4.55	0.002	8.27	0.004	10.2
BH	0.004	10.18	0.026	17.34	0.050	20.10
Storey	0.010	12.78	0.051	19.68	0.102	22.01
BKY	0.006	11.03	0.039	18.74	0.080	21.37
Bootstrap	0.008	11.32	0.042	19.31	0.087	21.82
			$\rho = 0.5$			
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.017	11.32	0.063	17.27	0.103	19.98
Bonferroni	0.001	2.28	0.003	4.78	0.005	6.4
BH	0.004	5.16	0.024	11.81	0.048	15.77
Storey ($\lambda = 0.5$)	0.008	7.21	0.047	14.67	0.093	18.65
BKY	0.004	5.49	0.032	12.78	0.065	17.02
Bootstrap	0.006	5.60	0.034	13.18	0.075	17.27

Notes: This table shows the results of a Monte Carlo simulation as specified in Section 1.3 with 2,000 replications and the indicated parameter settings. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

Table 1.11: Linear regression model with random correlation matrix, truncated normal r.v.

	Sample size: 100				*# Regressors: 10*	
	# False hypotheses: 5					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.012	1.92	0.055	2.78	0.103	3.17
BH	0.003	1.31	0.023	2.09	0.047	2.47
Storey	0.010	1.44	0.050	2.29	0.087	2.71
BKY	0.005	1.35	0.031	2.17	0.062	2.56
Bootstrap	0.007	1.45	0.038	2.35	0.078	2.81
	# False hypotheses: 2					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.036	0.760	0.133	1.11	0.225	1.28
BH	0.007	0.454	0.031	0.705	0.073	0.851
Storey	0.018	0.469	0.058	0.740	0.118	0.895
BKY	0.008	0.455	0.033	0.703	0.076	0.845
Bootstrap	0.011	0.503	0.048	0.792	0.101	0.954
	# False hypotheses: 1					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.041	0.388	0.182	0.558	0.316	0.637
BH	0.004	0.219	0.034	0.337	0.068	0.392
Storey ($\lambda = 0.5$)	0.014	0.230	0.066	0.352	0.119	0.416
BKY	0.004	0.219	0.034	0.337	0.067	0.386
Bootstrap	0.008	0.250	0.044	0.380	0.095	0.452

Notes: This table shows the results of a Monte Carlo simulation as specified in Section 1.3 with 2,000 replications and the indicated parameter settings. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

Table 1.12: Linear regression model with random correlation matrix, beta r.v.

	Sample size: 100				*# Regressors: 10*	
	# False hypotheses: 5					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.016	1.56	0.063	2.42	0.099	2.97
BH	0.006	0.92	0.021	1.59	0.045	2.09
Storey	0.012	1.07	0.043	1.83	0.082	2.37
BKY	0.007	0.95	0.028	1.65	0.054	2.17
Bootstrap	0.008	1.05	0.036	1.85	0.075	2.44
	# False hypotheses: 2					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.033	0.62	0.130	1.00	0.228	1.17
BH	0.006	0.30	0.030	0.54	0.067	0.69
Storey	0.013	0.33	0.060	0.60	0.113	0.74
BKY	0.006	0.31	0.032	0.54	0.068	0.68
Bootstrap	0.009	0.34	0.044	0.62	0.088	0.81
	# False hypotheses: 1					
	1%		5%		10%	
	FDR	CR	FDR	CR	FDR	CR
Classical	0.041	0.32	0.179	0.49	0.322	0.57
BH	0.006	0.14	0.040	0.24	0.077	0.33
Storey ($\lambda = 0.5$)	0.014	0.15	0.080	0.26	0.123	0.34
BKY	0.006	0.14	0.039	0.24	0.070	0.32
Bootstrap	0.008	0.17	0.054	0.28	0.103	0.37

Notes: This table shows the results of a Monte Carlo simulation as specified in Section 1.3 with 2,000 replications and the indicated parameter settings. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

Table 1.13: Size and power properties of the bootstrap under heteroscedasticity

			$N = 100$, $k = 50$, 10 false hypotheses		
	Pairwise using HC2	Classical using HC2	Pairwise using HC3	Classical using HC3	Wild bootstrap
Size	0.025	0.064	0.120	0.006	0.077
Power	0.602	0.732	0.766	0.470	0.738

			$N = 500$, $k = 10$, 2 false hypotheses		
	Pairwise using HC2	Classical using HC2	Pairwise using HC3	Classical using HC3	Wild bootstrap
Size	0.046	0.053	0.046	0.048	0.070
Power	0.997	1.000	0.997	0.999	1.000

Notes: This table shows the results of a size and power study for single hypothesis testing. The DGP is designed as described in Section 1.3, so there is a constant correlation ($\rho = 0.3$) among the regressors. Size is evaluated using the results from a single hypothesis test of one of the by design insignificant variables; power is evaluated for a corresponding test of a significant variable ($\beta_i = 0.5$ for significant variables). Heteroscedasticity is induced by multiplying the error term with $X_i X_j$, where X_i and X_j are the regressors corresponding to the significant and insignificant variable under study, respectively. The bootstrap methods that are compared are pairwise bootstrap using HC2 and HC3 standard errors and the wild bootstrap (for details see Appendix A1.1). For comparison, also the results of standard tests using HC2 and HC3 standard errors are provided.

Table 1.14: Linear regression model with 5 false hypotheses: FDP control

	# False hypotheses: 5			*Sample size: 100*			*# Regressors: 50*		
	$\tau = 1\%$			$\tau = 5\%$			$\tau = 10\%$		
	$P\{FDP > \nu\}$	FDR	CR	$P\{FDP > \nu\}$	FDR	CR	$P\{FDP > \nu\}$	FDR	CR
$\rho = 0$	0.005	0.001	1.584	0.026	0.007	2.429	0.050	0.013	2.835
$\rho = 0.3$	0.009	0.004	0.776	0.028	0.011	1.470	0.058	0.021	1.878
$\rho = 0.5$	0.005	0.003	0.374	0.021	0.010	0.819	0.048	0.021	1.107

Notes: This table shows the results of a Monte Carlo simulation as specified in section 3 with 2,000 replications and the indicated parameter settings. The FDP is controlled using the step-down procedure of Romano and Shaikh (2006). ν is always set to 5%.

DGP of Eicher, Papageorgiou, and Raftery (2011)

We replicate the DGP ('Model 2') of Eicher, Papageorgiou, and Raftery (2011):

$$y = i_n + \sum_{\ell=1}^{k/2} z_{(k/2+\ell)} + 2\epsilon, \qquad n = 100, \quad \epsilon \sim N(\mathbf{0}, \boldsymbol{I}_n) \tag{1.8}$$

Table 1.15: Eicher, Papageorgiou, and Raftery (2011) DGP

	$\gamma = 0.01$		$\gamma = 0.05$		$\gamma = 0.1$	
	FDR	*CR*	*FDR*	*CR*	*FDR*	*CR*
BH	0.001	19.65	0.009	19.86	0.020	19.93
Storey ($\lambda = 0.5$)	0.002	19.72	0.019	19.91	0.046	19.96
BKY	0.002	19.74	0.021	19.92	0.046	19.97
Bootstrap	0.003	19.74	0.023	19.93	0.050	19.97
Eicher *et al.* (2011)	# selected variables using $> 50\%$ posterior inclusion probability as cutoff:					
'best' prior	19		'worst' prior	3		

Notes: Monte Carlo results using the DGP from (1.8). 2,000 replications. 20 out of 40 hypotheses are false. The sample size is 100. The MTPs are described in Section 1.2. 'CR' stands for 'Correct Rejections'.

The model comprises $k = 40$ mean zero normal regressors with standard deviation two (of which one half thus matters) and equicorrelation of 0.5.[21] Eicher, Papageorgiou, and Raftery (2011) find that, using a posterior inclusion probability of more than 50% as a threshold, three to 19 of the 20 relevant variables are selected, depending on the choice of the prior. Given that the choice of a suitable prior is fraught with uncertainty (Ley and Steel, 2009, see also Sec. 1.4), the average expected performance of BMA probably lies somewhere between these extremes. Table 1.15 reports results for the MTPs. Even when controlling the FDR at the 1% level, thus virtually making no wrong rejections, we find around 19.7 correct rejections using any of our MTPs. This is slightly more than the result for the best of Eicher, Papageorgiou, and Raftery's priors.

[21] We do not use this DGP in our simulations above as the large implied theoretical t-ratios (Krolzig and Hendry, 2001) of $T_\ell^* = 1/(n^{-1/2}[\mathrm{Cov}(\boldsymbol{x})^{-1/2}]_{\ell\ell}\sigma_\epsilon) \approx 10/2 = 5$ make it somewhat too easy to identify true variables in our view. (To see this, write $\mathrm{Cov}(\boldsymbol{x}) = \mathrm{diag}(\mathrm{Var}(\boldsymbol{x}))^{1/2}\,\mathrm{Corr}(\boldsymbol{x})\,\mathrm{diag}(\mathrm{Var}(\boldsymbol{x}))^{1/2}$. From Graybill (1983, Thm. 8.3.4) we have $\mathrm{Corr}(\boldsymbol{x})^{-1} = 1/(1-\rho)\cdot[\boldsymbol{I}_k - \rho/(1+(k-1)\rho)\boldsymbol{J}_k]$ with $\boldsymbol{J}_k$ a $k \times k$ matrix of ones. Hence $[\mathrm{Cov}(\boldsymbol{x})^{-1/2}]_{\ell\ell} = \sqrt{[1-\rho/(1+(k-1)\rho)]/[(1-\rho)\,\mathrm{Var}(\boldsymbol{x})]} = 0.9877$.)

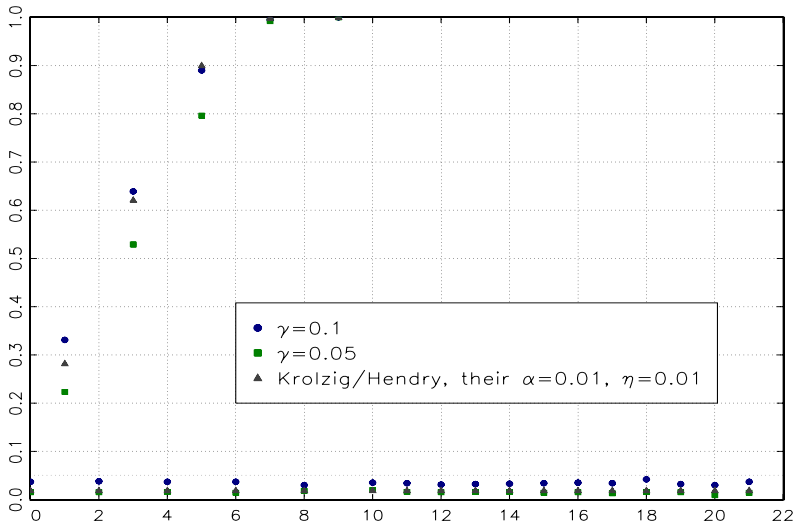


Figure 1.4: Rejection Rates Bootstrap Method—Krolzig and Hendry (2001) DGP

Note: The numbers on the horizontal axis indicate the variable from the GUM (1.10). Variables 1, 3, 5 and 7 correspond to the variables included in the DGP (1.9). The vertical axis plots the inclusion probability of the variables from (1.10) when controlling the FDR at $\gamma = 0.1$ and $\gamma = 0.05$ using the bootstrap. The dotted horizontal line is at 5%.

DGP of Krolzig and Hendry (2001)

We also compare our results to Krolzig and Hendry's (2001) PcGets approach, using their DGP:

$$y_t = \sum_{k=1}^{5} \beta_k x_{k,t} + \epsilon_t, \qquad \epsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1) \quad \text{for } t = 1, \ldots, T \tag{1.9}$$

$$\boldsymbol{x}_t \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_{10}); \quad \beta_1 = 2/\sqrt{T},\ \beta_2 = 3/\sqrt{T},\ \beta_3 = 4/\sqrt{T},\ \beta_4 = 6/\sqrt{T} \text{ and } \beta_5 = 8/\sqrt{T}$$

These β_k yield theoretical t-ratios of 2,3,4,6 and 8. The GUM also includes y_{t-1}, the irrelevant variables $x_{6,t}, \ldots, x_{10,t}$ and lags of all $x_{k,t}$:

$$y_t = \pi_{0,1} y_{t-1} + \sum_{k=1}^{10} \sum_{p=0}^{1} \pi_{k,p} x_{k,t-p} + \pi_{0,0} + u_t \tag{1.10}$$

Results are summarized in Figure 1.4. It is analogous to their Figure 4 and reproduces their results for convenience.[22] We focus on the bootstrap method, but similar results for

[22]We take the 'average' size of 0.0189 reported in their Table 4 for all irrelevant regressors as Krolzig and Hendry (2001) only graphically report these for each irrelevant regressor, but not the exact figure. This leads to no meaningful inaccuracy as their Figure 4 suggests that the size of all irrelevant regressors is very close.

the other MTPs are also.[23] As expected, higher theoretical t-ratios translate into higher selection probabilities. We again control the FDR at γ. The irrelevant variables are selected with probabilities of around 3% for $\gamma = 0.1$. This is slightly higher than the selection probability that is obtained if both t-tests and diagnostic tests are conducted at levels $\alpha = \eta = 0.01$ in the simulations of Krolzig and Hendry (2001), see the triangles in Figure 1.4. (The combined effect of variable selection and diagnostic testing in PcGets inevitably leads to selection probabilities for irrelevant variables of more than α.) The bootstrap method then is somewhat more powerful than PcGets employing this DGP. It would be roughly equally powerful if the sizes of both techniques were identical.

Table 1.16: Krolzig and Hendry (2001) DGP for $\alpha = \gamma = 0.01$

	Probability of including each variable				
Number of variable	*Classical*	*BH*	*Storey*	*BKY*	*Bootstrap*
0	0.011	0.003	0.006	0.004	0.003
1	0.199	0.086	0.097	0.091	0.097
2	0.008	0.001	0.003	0.002	0.002
3	0.523	0.308	0.324	0.321	0.324
4	0.009	0.002	0.003	0.002	0.002
5	0.813	0.603	0.623	0.616	0.635
6	0.009	0.001	0.002	0.001	0.002
7	0.991	0.957	0.959	0.958	0.960
8	0.011	0.003	0.005	0.004	0.003
9	1.000	0.999	0.999	0.999	0.999
10	0.013	0.003	0.005	0.004	0.004
11	0.008	0.002	0.004	0.002	0.003
12	0.009	0.003	0.003	0.003	0.004
13	0.011	0.001	0.002	0.001	0.002
14	0.008	0.002	0.002	0.002	0.001
15	0.012	0.003	0.003	0.003	0.003
16	0.014	0.005	0.006	0.005	0.006
17	0.010	0.002	0.002	0.002	0.002
18	0.010	0.004	0.005	0.005	0.005
19	0.009	0.002	0.003	0.002	0.003
20	0.014	0.004	0.006	0.004	0.004
21	0.010	0.003	0.004	0.003	0.004

Notes: This table shows the results of a Monte Carlo simulation with settings as in Krolzig and Hendry (2001) with 2,000 replications and $n = 100$. The MTPs are described in Section 1.2. For Storey, $\lambda = 0.5$.

[23] Please refer to Tables 1.16 to 1.18.

Table 1.17: Krolzig and Hendry (2001) DGP for $\alpha = \gamma = 0.05$

	Probability of including each variable				
Number of variable	*Classical*	*BH*	*Storey*	*BKY*	*Bootstrap*
0	0.046	0.014	0.018	0.014	0.015
1	0.414	0.201	0.221	0.211	0.223
2	0.050	0.011	0.019	0.013	0.015
3	0.739	0.499	0.524	0.515	0.529
4	0.053	0.013	0.023	0.017	0.016
5	0.923	0.777	0.798	0.786	0.796
6	0.046	0.011	0.015	0.013	0.014
7	0.998	0.989	0.990	0.989	0.992
8	0.055	0.013	0.022	0.018	0.018
9	1.000	1.000	1.000	1.000	1.000
10	0.052	0.015	0.023	0.017	0.020
11	0.048	0.011	0.018	0.016	0.016
12	0.048	0.011	0.016	0.012	0.015
13	0.057	0.010	0.019	0.014	0.016
14	0.046	0.014	0.017	0.015	0.016
15	0.048	0.009	0.014	0.011	0.014
16	0.055	0.011	0.019	0.014	0.015
17	0.049	0.010	0.016	0.013	0.013
18	0.048	0.012	0.016	0.014	0.016
19	0.055	0.012	0.018	0.013	0.015
20	0.039	0.004	0.009	0.007	0.009
21	0.047	0.013	0.019	0.013	0.014

Notes: This table shows the results of a Monte Carlo simulation with settings as in Krolzig and Hendry (2001) with 2,000 replications and $n = 100$. The MTPs are described in Section 1.2. For Storey, $\lambda = 0.5$.

Table 1.18: Krolzig and Hendry (2001) DGP for $\alpha = \gamma = 0.1$

	Probability of including each variable				
Number of variable	*Classical*	*BH*	*Storey*	*BKY*	*Bootstrap*
0	0.107	0.027	0.042	0.034	0.037
1	0.546	0.297	0.324	0.311	0.331
2	0.097	0.027	0.045	0.033	0.038
3	0.834	0.606	0.636	0.615	0.639
4	0.102	0.029	0.044	0.033	0.037
5	0.972	0.863	0.873	0.868	0.880
6	0.106	0.024	0.041	0.030	0.037
7	0.999	0.995	0.995	0.995	0.995
8	0.087	0.023	0.035	0.026	0.030
9	1.000	1.000	1.000	1.000	1.000
10	0.100	0.029	0.044	0.031	0.035
11	0.097	0.026	0.039	0.030	0.034
12	0.096	0.024	0.036	0.028	0.031
13	0.103	0.022	0.037	0.028	0.032
14	0.099	0.021	0.032	0.028	0.033
15	0.097	0.024	0.036	0.029	0.034
16	0.101	0.025	0.041	0.028	0.035
17	0.092	0.022	0.036	0.028	0.034
18	0.103	0.030	0.043	0.034	0.042
19	0.098	0.025	0.034	0.029	0.032
20	0.093	0.022	0.036	0.028	0.030
21	0.107	0.027	0.038	0.033	0.037

Notes: This table shows the results of a Monte Carlo simulation with settings as in Krolzig and Hendry (2001) with 2,000 replications and $n = 100$. The MTPs are described in Section 1.2. For Storey, $\lambda = 0.5$.

A1.3 Additional empirical results

Table 1.19: Results for the FLS/Sala-i-Martin (1997) data set

	Regressor	$\hat{\beta}_\ell$	*p-value*	*Classical*	*BH*	*Storey*	*BKY*	*Boot*
1	GDP level 1960	−0.017	0.00001	1%	1%	1%	1%	1%
2	Fraction Confucian	0.075	0.00003	1%	1%	1%	1%	1%
3	Life expectancy	0.001	0.003	1%	5%	5%	5%	5%
4	Equipment investment	0.127	0.008	1%	5%	5%	5%	5%
5	Sub-Saharan dummy	−0.020	0.006	1%	5%	5%	5%	5%
6	Fraction Muslim	0.011	0.224	-	-	-	-	-
7	Rule of law	0.012	0.070	10%	-	-	-	-
8	Number of years open economy	−0.003	0.634	-	-	-	-	-
9	Degree of capitalism	0.001	0.297	-	-	-	-	-
10	Fraction Protestant	−0.003	0.686	-	-	-	-	-
11	Fraction GDP in mining	0.040	0.007	1%	5%	5%	5%	5%
12	Non-Equipment investment	0.037	0.083	10%	-	-	-	-
13	Latin American dummy	−0.013	0.044	5%	-	10%	10%	10%
14	Primary school enrollment, 1960	0.020	0.045	5%	-	10%	10%	10%
15	Fraction Buddhist	0.007	0.273	-	-	-	-	-
16	Black-market premium	−0.007	0.076	10%	-	-	-	-
17	Fraction Catholic	0.003	0.589	-	-	-	-	-
18	Civil liberties	−0.002	0.312	-	-	-	-	-
19	Fraction Hindu	−0.097	0.001	1%	1%	1%	1%	1%
20	Political rights	0.0002	0.913	-	-	-	-	-
21	Primary exports, 1970	−0.006	0.422	-	-	-	-	-
22	Exchange rate distortions	−0.00002	0.530	-	-	-	-	-
23	Age	−0.00001	0.781	-	-	-	-	-
24	War dummy	−0.001	0.534	-	-	-	-	-
25	Size labor force	0.0000003	0.004	1%	5%	5%	5%	5%
26	Fraction speaking foreign language	−0.002	0.470	-	-	-	-	-
27	Fraction of pop speaking English	−0.007	0.130	-	-	-	-	-
28	Ethnologic fractionalization	0.014	0.012	5%	5%	5%	5%	5%
29	Spanish colony dummy	0.013	0.022	5%	10%	10%	10%	10%
30	SD of black-market premium	−0.000001	0.882	-	-	-	-	-
31	French colony dummy	0.009	0.037	5%	-	10%	10%	10%
32	Absolute latitude	−0.0001	0.523	-	-	-	-	-
33	Ratio of workers to population	−0.001	0.946	-	-	-	-	-
34	Higher education enrollment	−0.129	0.002	1%	5%	5%	5%	5%
35	Population growth	−0.119	0.595	-	-	-	-	-
36	British colony dummy	0.007	0.070	10%	-	-	-	-
37	Outward orientation	−0.005	0.037	5%	-	10%	10%	10%
38	Fraction Jewish	−0.001	0.960	-	-	-	-	-
39	Revolutions and coups	0.003	0.505	-	-	-	-	-
40	Public education share	0.137	0.254	-	-	-	-	-
41	Area (scale effect)	0.0000003	0.639	-	-	-	-	-
42	Intercept	0.0207	0.000					

Notes: For every regressor the table shows if the variable is found significant when controlling the FDR at the indicated level $\gamma = \{1\%, 5\%, 10\%\}$. The procedures are described in Section 1.2. 5,000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table 1.20: Results for the FLS data set using the wild bootstrap

	Regressor	$\hat{\beta}_i$	*Wild bootstrap*
1	GDP level 1960	−0.017	1%
2	Fraction Confucian	0.075	1%
3	Life expectancy	0.001	-
4	Equipment investment	0.127	5%
5	Sub-Saharan dummy	−0.020	5%
6	Fraction Muslim	0.011	-
7	Rule of law	0.012	-
8	Number of years open economy	−0.003	-
9	Degree of capitalism	0.001	-
10	Fraction Protestant	−0.003	-
11	Fraction GDP in mining	0.040	5%
12	Non-Equipment investment	0.037	-
13	Latin American dummy	−0.013	-
14	Primary school enrollment, 1960	0.020	-
15	Fraction Buddhist	0.007	-
16	Black-market premium	−0.007	-
17	Fraction Catholic	0.003	-
18	Civil liberties	−0.002	-
19	Fraction Hindu	−0.097	1%
20	Political rights	0.0002	-
21	Primary exports, 1970	−0.006	-
22	Exchange rate distortions	−0.00002	-
23	Age	−0.00001	-
24	War dummy	−0.001	-
25	Size labor force	0.0000003	5%
26	Fraction speaking foreign language	−0.002	-
27	Fraction of pop speaking English	−0.007	-
28	Ethnologic fractionalization	0.014	10%
29	Spanish colony dummy	0.013	10%
30	SD of black-market premium	−0.000001	-
31	French colony dummy	0.009	5%
32	Absolute latitude	−0.0001	-
33	Ratio of workers to population	−0.001	-
34	Higher education enrollment	−0.129	5%
35	Population growth	−0.119	-
36	British colony dummy	0.007	-
37	Outward orientation	−0.005	-
38	Fraction Jewish	−0.001	-
39	Revolutions and coups	0.003	-
40	Public education share	0.137	-
41	Area (scale effect)	0.0000003	-

Notes: For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in Section 1.2. We work with 5,000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table 1.21: Results for the FLS data set using HC_2 standard errors

	Regressor	$\hat{\beta}_i$	*p-value*	*BH*	*Storey*	*BKY*	*Pairs bootstrap*
1	GDP level 1960	−0.017	0.018	10%	10%	10%	-
2	Fraction Confucian	0.075	0.024	10%	10%	10%	-
3	Life expectancy	0.001	0.0001	1%	1%	1%	-
4	Equipment investment	0.127	0.095	-	-	-	-
5	Sub-Saharan dummy	−0.020	0.004	5%	5%	5%	-
6	Fraction Muslim	0.011	0.463	-	-	-	-
7	Rule of law	0.012	0.036	-	10%	-	-
8	Number of years open economy	−0.003	0.144	-	-	-	-
9	Degree of capitalism	0.001	0.654	-	-	-	-
10	Fraction Protestant	−0.003	0.032	-	10%	-	-
11	Fraction GDP in mining	0.040	0.325	-	-	-	-
12	Non-Equipment investment	0.037	0.906	-	-	-	-
13	Latin American dummy	−0.013	0.004	5%	5%	5%	-
14	Primary school enrollment, 1960	0.020	0.006	5%	5%	5%	-
15	Fraction Buddhist	0.007	0.555	-	-	-	-
16	Black-market premium	−0.007	0.620	-	-	-	-
17	Fraction Catholic	0.003	0.0002	1%	1%	1%	-
18	Civil liberties	−0.002	0.543	-	-	-	-
19	Fraction Hindu	−0.097	0.928	-	-	-	-
20	Political rights	0.0002	0.267	-	-	-	-
21	Primary exports, 1970	−0.006	0.682	-	-	-	-
22	Exchange rate distortions	−0.00002	0.788	-	-	-	-
23	Age	−0.00001	0.010	5%	5%	5%	-
24	War dummy	−0.001	0.927	-	-	-	-
25	Size labor force	0.0000003	0.054	-	10%	-	-
26	Fraction speaking foreign language	−0.002	0.463	-	-	-	-
27	Fraction of pop speaking English	−0.007	0.399	-	-	-	-
28	Ethnologic fractionalization	0.014	0.052	-	10%	-	-
29	Spanish colony dummy	0.013	0.563	-	-	-	-
30	SD of black-market premium	−0.000001	0.054	-	10%	-	-
31	French colony dummy	0.0009	0.0001	1%	1%	1%	-
32	Absolute latitude	−0.0001	0.001	5%	1%	5%	-
33	Ratio of workers to population	−0.001	0.001	5%	1%	5%	-
34	Higher education enrollment	−0.129	0.121	-	-	-	-
35	Population growth	−0.114	0.951	-	-	-	-
36	British colony dummy	0.007	0.417	-	-	-	-
37	Outward orientation	−0.005	0.043	-	10%	-	-
38	Fraction Jewish	−0.001	0.303	-	-	-	-
39	Revolutions and coups	0.003	0.514	-	-	-	-
40	Public education share	0.138	0.538	-	-	-	-
41	Area (scale effect)	0.0000003	0.054	-	10%	-	-

Notes: For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in Section 1.2. We work with 5,000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table 1.22: Results for the FLS data set using HC_3 standard errors

	Regressor	$\hat{\beta}_i$	*p-value*	*BH*	*Storey*	*BKY*	*Pairs bootstrap*
1	GDP level 1960	−0.017	0.253	-	-	-	5%
2	Fraction Confucian	0.075	0.136	-	-	-	5%
3	Life expectancy	0.001	0.007	-	-	-	5%
4	Equipment investment	0.127	0.307	-	-	-	5%
5	Sub-Saharan dummy	−0.020	0.121	-	-	-	5%
6	Fraction Muslim	0.011	0.662	-	-	-	-
7	Rule of law	0.012	0.218	-	-	-	5%
8	Number of years open economy	−0.003	0.395	-	-	-	5%
9	Degree of capitalism	0.001	0.808	-	-	-	-
10	Fraction Protestant	−0.003	0.215	-	-	-	5%
11	Fraction GDP in mining	0.040	0.553	-	-	-	10%
12	Non-Equipment investment	0.037	0.946	-	-	-	-
13	Latin American dummy	−0.013	0.087	-	-	-	5%
14	Primary school enrollment, 1960	0.020	0.077	-	-	-	5%
15	Fraction Buddhist	0.007	0.723	-	-	-	-
16	Black-market premium	−0.007	0.764	-	-	-	-
17	Fraction catholic	0.003	0.071	-	-	-	5%
18	Civil liberties	−0.002	0.717	-	-	-	-
19	Fraction Hindu	−0.097	0.995	-	-	-	-
20	Political rights	0.0002	0.503	-	-	-	10%
21	Primary exports, 1970	−0.006	0.806	-	-	-	-
22	Exchange rate distortions	−0.00002	0.864	-	-	-	-
23	Age	−0.00001	0.128	-	-	-	5%
24	War dummy	−0.001	0.958	-	-	-	-
25	Size labor force	0.0000003	0.239	-	-	-	5%
26	Fraction speaking foreign language	−0.002	0.688	-	-	-	-
27	Fraction of pop speaking English	−0.007	0.620	-	-	-	-
28	Ethnologic fractionalization	0.014	0.254	-	-	-	5%
29	Spanish colony dummy	0.013	0.745	-	-	-	-
30	SD of black-market premium	−0.000001	0.239	-	-	-	5%
31	French colony dummy	0.0009	0.022	-	-	-	5%
32	Absolute latitude	−0.0001	0.068	-	-	-	5%
33	Ratio of workers to population	−0.001	0.068	-	-	-	5%
34	Higher education enrollment	−0.129	0.397	-	-	-	5%
35	Population growth	−0.114	0.971	-	-	-	-
36	British colony dummy	0.007	0.665	-	-	-	-
37	Outward orientation	−0.005	0.307	-	-	-	5%
38	Fraction Jewish	−0.001	0.532	-	-	-	10%
39	Revolutions and coups	0.003	0.696	-	-	-	-
40	Public education share	0.138	0.713	-	-	-	-
41	Area (scale effect)	0.0000003	0.239	-	-	-	5%

Notes: For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in Section 1.2. We work with 5,000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

A1.4 MP data set

The data set covers 93 countries, for which average GDP growth was calculated from 1960 to 1992. It was developed and first used in Masanjala and Papageorgiou (2006). The data set consists of 32 basic variables out of which 22 are also combined with an interaction dummy for African countries. Thus, $k = 54$ here. To tackle the endogeneity issues Masanjala and Papageorgiou (2006) devote careful attention to only including predetermined variables. Furthermore, the interaction dummies address possible parameter heterogeneity.

Table 1.23 shows the results when accounting for multiple testing by controlling the FDR in the MP data set.[24] Only three variables are found to be significantly related to GDP growth, namely ln GDP per capita, 1960 (at $\gamma = 0.05$), Life expectancy(at $\gamma = 0.01$) and 'A∗Mining'(at $\gamma = 0.01$) using the bootstrap method.[25] Again the difference to the number of significant variables using classical testing is substantial. Classical testing finds 13 significant variables at $\alpha = 0.1$. Compared to the FLS data, a possible reason for the larger difference between classical testing and MTP results is the larger number of explanatory variables. This results in stricter MTPs. (Of course, it is also simply other data.)

Comparing our findings to those of the BMA approach of Masanjala and Papageorgiou (2006), we can confirm two out of their three variables with a marginal posterior probability of inclusion of 100%. On the other hand, 'A*Mining' only has marginal posterior probability of 76.1, yet it is found significant by the MTPs. Hence, again the BMA results differ somewhat from the multiple testing results.[26] Ley and Steel (2009) also use this data set to investigate the influence of the priors on the outcome of BMA. Again, depending on the different choices, the average posterior model size ranges from 5.42 ($m = 7$, random θ, $g = 1/k^2$) to 17.90 ($m = 27$, fixed θ, $g = 1/n$).

[24] We also calculate robust standard errors. The results are qualitatively the same as for the FLS data set.

[25] For the MP data set, BH and BKY yield the same results; Storey finds significance of 'A*Mining' only at a FDR of 5%.

[26] For a comparison of the marginal posterior probability to our results refer to Table 1.24.

Table 1.23: MP data set

	Regressor	$\hat{\beta}_i$	*p-value*	*Classical*	*BH*	*Storey*	*BKY*	*Boot*
1	ln GDP per capita, 1960	−1.80	0.000960	1%	5%	5%	5%	5%
2	Life expectancy, 1960	0.196	0.000101	1%	1%	1%	1%	1%
3	A∗Mining	10.7	0.000504	1%	1%	5%	1%	1%
4	South-East Asia	2.15	0.0240	5%	-	-	-	-
5	Fraction muslim	2.62	0.0174	5%	-	-	-	-
6	A∗Land area	−0.00113	0.0744	10%	-	-	-	-
7	OECD	1.48	0.0337	5%	-	-	-	-
8	A∗Primary export, 1970	−4.73	0.0141	5%	-	-	-	-
9	Primary export, 1970	0.0783	0.942	-	-	-	-	-
10	A∗European language	3.98	0.278	-	-	-	-	-
11	European language	0.955	0.0575	10%	-	-	-	-
12	Fraction Confucian	3.74	0.0835	10%	-	-	-	-
13	A∗Colony	1.98	0.604	-	-	-	-	-
14	A∗Tropical fraction	−2.90	0.570	-	-	-	-	-
15	A∗Landlocked	−0.199	0.849	-	-	-	-	-
16	Latin America	−0.227	0.778	-	-	-	-	-
17	A∗Labor force, 1960	0.000	0.755	-	-	-	-	-
18	Malaria prevalence, 1960	0.335	0.723	-	-	-	-	-
19	Landlocked	−0.436	0.427	-	-	-	-	-
20	Fraction Catholic	0.835	0.383	-	-	-	-	-
21	A∗Malaria prevalence, 1960	−1.77	0.664	-	-	-	-	-
22	A∗Primary school, 1960	4.39	0.0574	10%	-	-	-	-
23	A∗Fraction Muslim	0.905	0.577	-	-	-	-	-
24	Fraction Protestant	0.534	0.619	-	-	-	-	-
25	A∗ln GDP per capita, 1960	0.300	0.725	-	-	-	-	-
26	Tropical fraction	0.559	0.519	-	-	-	-	-
27	Fraction Buddhist	0.707	0.539	-	-	-	-	-
28	Primary school, 1960	−1.42	0.400	-	-	-	-	-
29	A∗British colony	−0.483	0.887	-	-	-	-	-
30	Sub Saharan Africa	7.68	0.235	-	-	-	-	-
31	A∗Fraction urban pop, 1960	−6.49	0.105	-	-	-	-	-
32	A∗Life expectancy, 1960	−0.0356	0.666	-	-	-	-	-
33	A∗Secondary school, 1960	1.13	0.961	-	-	-	-	-
34	Mining	−0.319	0.892	-	-	-	-	-
35	Fraction Hindu	2.82	0.0957	10%	-	-	-	-
36	A∗Fraction Catholic	−2.43	0.753	-	-	-	-	-
37	Distance from equator	−0.00667	0.800	-	-	-	-	-
38	Land area	0.000161	0.0583	10%	-	-	-	-
39	Colony	−1.21	0.258	-	-	-	-	-
40	A∗Ethnoliguistic fractionalization	−0.559	0.645	-	-	-	-	-
41	Spanish colony	0.895	0.324	-	-	-	-	-
42	Secondary school, 1960	0.725	0.622	-	-	-	-	-
43	A∗French colony	−1.03	0.751	-	-	-	-	-
44	A∗Distance from equator	−0.103	0.142	-	-	-	-	-
45	Tertiary education, 1960	−4.50	0.401	-	-	-	-	-
46	British colony	0.554	0.582	-	-	-	-	-
47	A∗Fraction Protestant	−2.73	0.639	-	-	-	-	-
48	A∗Fraction English speaking	−3.71	0.847	-	-	-	-	-
49	French colony	0.861	0.369	-	-	-	-	-
50	Ethnoliguistic fractionalization	−0.107	0.876	-	-	-	-	-
51	Fraction English speaking	0.281	0.693	-	-	-	-	-
52	Labor force, 1960	−0.000002	0.810	-	-	-	-	-
53	Fraction Jewish	2.00	0.166	-	-	-	-	-
54	Fraction urban pop, 1960	0.0889	0.924	-	-	-	-	-
	Intercept	1.73	0.000	1%	1%	1%	1%	1%

Notes: For every regressor in the FLS data set this table shows whether the variable is found to be significant when controlling the FDR at the indicated level. The procedures are described in Section 1.2. We work with 5,000 bootstrap iterations. In the Storey approach $\lambda = 0.5$.

Table 1.24: Comparison MP data set

	Regressor	*Bootstrap approach*	*BMA post. prob.*
1	ln GDP per capita, 1960	5%	100.0
2	Life expectancy, 1960	1%	100.0
3	A∗Mining	1%	76.1
4	South-East Asia	-	93.9
5	Fraction muslim	-	83.5
6	A∗ Land area	-	62.8
7	OECD	-	45.3
8	A∗Primary export, 1970	-	43.1
9	Primary export, 1970	-	49.9
10	A∗European language	-	51.4
11	European language	-	24.4
12	Fraction Confucian	-	21.7
13	A∗Colony	-	54.6
14	A∗Tropical fraction	-	100.0
15	A∗Landlocked	-	72.2
16	Latin America	-	7.5
17	A∗Labor force, 1960	-	10.6
18	Malaria prevalence, 1960	-	12.8
19	Landlocked	-	7.1
20	Fraction Catholic	-	7.2
21	A∗Malaria prevalence, 1960	-	78.9
22	A∗Primary school, 1960	-	3.2
23	A∗Fraction Muslim	-	1.4
24	Fraction Protestant	-	3.8
25	A∗ln GDP per capita, 1960	-	2.8
26	Tropical fraction	-	3.5
27	Fraction Buddhist	-	2.0
28	Primary school, 1960	-	4.4
29	A∗British colony	-	1.1
30	Sub Saharan Africa	-	2.7
31	A∗Fraction urban pop, 1960	-	1.1
32	A∗Life expectancy, 1960	-	1.5
33	A∗Secondary school, 1960	-	0.9
34	Mining	-	4.9
35	Fraction Hindu	-	1.0
36	A∗Fraction Catholic	-	0.5
37	Distance from equator	-	0.5
38	Land area	-	0.6
39	Colony	-	0.9
40	A∗Ethnoliguistic fractionalization	-	1.0
41	Spanish colony	-	0.7
42	Secondary school, 1960	-	1.1
43	A∗French colony	-	5.1
44	A∗Distance from equator	-	0.2
45	Tertiary education, 1960	-	0.2
46	British colony	-	0.3
47	A∗Fraction Protestant	-	0.2
48	A∗Fraction English speaking	-	47.1
49	French colony	-	0.3
50	Ethnoliguistic fractionalization	-	0.2
51	Fraction English speaking	-	0.1
52	Labor force, 1960	-	0.1
53	Fraction Jewish	-	0.1
54	Fraction urban pop, 1960	-	0.3

Notes: 'Bootstrap approach' denotes significance when controlling the FDR at the indicated level; 'BMA post. prob.' denotes the marginal posterior probability of inclusion found in Masanjala and Papageorgiou (2006).

Chapter 2

The Relationship Between Economic Preferences and Psychological Personality Measures

2.1 Introduction

Both economists and personality psychologists seek to identify determinants of heterogeneity in behavior. Economists typically depict decision problems in a framework of utility maximization. An individual's utility is shaped by preferences such as risk, time, and social preferences.[1] These preferences, in combination with expectations of future events, perceptions, beliefs, strategic consideration, prices and constraints shape behavior. Personality psychology, the branch of psychology studying personality and individual differences, offers several frameworks describing universal traits and individual differences. Personality traits – defined by Roberts (2009, p. 140) as "the relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances" – are important determinants of personality (Roberts, 2006) and affect outcomes. There has been a long tradition in personality psychology to measure personality

[1] In the standard expected utility framework, risk preference is captured by the curvature of the utility function, whereas the degree of risk aversion is represented in the concavity of the utility function (e.g. Gollier, 2001). Time preference describes how an individual trades off utility at different points in time (Samuelson, 1937; Frederick, Loewenstein, and O'Donoghue, 2002). Social preferences capture the idea that an individual's utility does not depend only on his own material payoff, but that it is also shaped by others' behavior and material payoff. Social preferences include altruism (e.g. Eckel and Grossmann, 1996) and negative and positive reciprocity (e.g. Falk and Fischbacher, 2006). Finally, trust describes an individual's belief about others' trustworthiness combined with a preference to take social risks (e.g. Fehr, 2009). Another important economic preference is the preference for work versus leisure. This preference is difficult to measure in experiments and is therefore not part of our analysis.

traits. The Big Five or five-factor model is the most widely used taxonomy of personality traits. It originates from the lexical hypothesis of Allport and Odbert (1936), which postulates that individual differences are encoded in language (see Borghans et al. 2008). After years of research in this tradition, psychologists have arrived at a hierarchical organization of personality traits with five traits at the highest level. These Big Five traits, which are commonly labeled as openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, capture personality traits at the broadest level of abstraction. Each Big Five trait condenses several distinct and more narrowly defined traits. It has been argued that the bulk of items that personality psychologists have used to measure personality can be mapped into the Big Five taxonomy (see, e.g., Costa and McCrae, 1992)).[2] Another important concept in psychology focusing on individual beliefs and perceptions is the locus of control framework by Rotter (1966). It represents the framework of the social learning theory of personality and refers to the extent people believe they have control over events.

An integration of the different measures and concepts used by economists and personality psychologists promises much potential for amalgamating evidence about the drivers of human behavior which accumulated disjointedly in the fields of economics and psychology (Borghans et al., 2008). Recently, scholars have begun to integrate personality into economic decision making (e.g., Borghans et al., 2008). Almlund et al. (2011a) enrich theory by incorporating personality traits in a standard economic framework of production, choice, and information. Their model interprets measured personality as a "construct derived from an economic model of preferences, constraints, and information" (Almlund et al., 2011, p. 3). However, empirical knowledge is too limited to judge how personality traits relate to the concepts and parameters economists typically model to predict behavior.

To shed more light on the relationship between economic preferences and psychological measures of personality we therefore study how key economic preferences, such as risk, time and social preferences, are linked to conventional measures of personality, such as the Big Five and locus of control. We analyze this relationship in a coherent framework using two main approaches. The first approach focuses on assessing the magnitude of the correlations between psychological and economic measurement systems in three unique data sets. The second approach departs from the fact that both preference measures and measures of personality traits predict a wide range of important life outcomes. If these two measurement systems are closely linked, they are expected to be substitutes in explaining heterogene-

[2]For a more detailed description of the research on the development of the Big Five, criticism of the approach and alternative measurement systems see Borghans et al. (2008).

ity in behavior. If, however, preferences and personality traits capture different aspects of behavior, the two measurement systems may have complementary predictive power for important life outcomes. We therefore evaluate the individual as well as the joint explanatory power of economic preferences and psychological measures of personality in explaining health, educational and labor market outcomes.

We use three complementary datasets. First, we look at data from laboratory experiments. Using a student subject pool we conducted choice experiments on key economic preferences, namely risk taking, time discounting, altruism, trust, and positive and negative reciprocity. We incentivized decision-making and obtained multiple behavioral measures for each preference. We assessed the Big Five domains using the 60-item NEO-FFI (NEO Five Factor Inventory) (Costa and McCrae, 1989) and a 15-item subset, the so-called BFI-S (Gerlitz and Schupp, 2005). We also measured the locus of control using 10 items adapted from Rotter (1966). Our second data set comprises very similar incentivized experimental measures with respect to risk taking and time discounting using a representative sample of almost 1000 participants from the German population. We are therefore able to obtain incentivized preference measures for a representative population. Personality was assessed using the BFI-S. The third data set stems from the German Socio-Economic Panel Study (SOEP), comprising preference and personality measures for a representative sample of more than 14.000 individuals. Preference measures were obtained using subjective self-assessment survey items rather than incentivized experiments, and personality was measured by using the BFI-S and the locus of control questionnaire. Using this data set we analyze associations between important life outcomes, such as labor market success, subjective health status or life satisfaction, and individuals' preferences and personalities.

These three data sets allow for a comprehensive analysis. The first data set contains very detailed personality measures in combination with multiple experimental indicators for preferences. This student sample therefore provides a particularly accurate assessment of potential relations between economic preferences and personality. The second data set uses experimental measures for a limited set of preferences and a shorter version of the Big Five but a representative sample. A comparison of results of the two data sets therefore informs us about the generalizability of our findings from the student sample. The third data set additionally allows us to study an even larger sample and to explore the explanatory power of personality and preferences for important life outcomes.

We start by analyzing data on 489 university students. We relate all five factors that

capture personality according to the Big Five taxonomy and the measure of Locus of Control to our experimental preference measures. We generally find only small correlations between personality traits and preferences. In particular, only 11 of the 36 correlations in our student sample exceed 0.1 in absolute value and only one correlation exceeds 0.2 in absolute value. These eleven correlation coefficients are all significant at conventional levels, and eight of them involve correlations between social preferences and personality traits.

Next, we gauge whether the correlation patterns generalize to representative samples. We first turn to the data set that contains very similar experimental measures of risk and time preferences and survey measures of the Big Five approximately 1000 individuals, who were sampled to be representative of the adult population living in Germany (see Dohmen et al., 2010). The correlation structure between personality traits and risk and time preferences turns out to be similar to the one we find for students, with few exceptions.

Finally, we assess whether the empirical associations between preference parameters and personality traits are sensitive to the way in which preferences are measured. We compare correlations between personality traits and measures of preferences derived from the incentivized choice experiments in the student and the representative sample to correlations that are constructed based on the non-incentivized subjective self-assessments in a representative sample of 14.000 individuals from the SOEP. Our result on the pattern of correlations between preference measures and personality measures is again largely confirmed.

We then turn to a different type of analysis in which we assess the power of preferences and personality in explaining life outcomes, including health, life satisfaction, earnings, unemployment and education. Our analysis reveals that both measurement systems have similar explanatory power when used separately as explanatory variables. The explained fraction of variance increases by approximately 60% when life outcomes are regressed on both measurement systems. We therefore conclude that each measurement system captures distinct sources of the heterogeneity in life outcomes. A coherent picture emerges from our analysis. Both approaches strongly suggest that standard measures of preferences and personality are complementary constructs.

So far no clear picture concerning the relations between measures of personality and economic preferences has emerged in the literature (see Almlund et al., 2011a). For example, the study by Daly, Delaney, and Harmon (2009) suggests a negative relationship between conscientiousness and the discount rate, but such a negative correlation is not corroborated by Dohmen et al. (2010), who relate experimental measures of willingness to take risk and

impatience to survey measures of the Big Five in a representative sample of adults living in Germany, nor by Anderson et al. (2011), who relate a measure of delay acceptance to four of the Big Five domains in a sample of 1065 US trainee truckers.[3] In fact, Dohmen et al. (2010) find no significant relationship between personality traits and preference measures in a regression framework that includes controls for IQ, gender, age, height, education, and household income. Raw correlations between preference and personality measures, which are also reported in Almlund et al. (2011a), are weak; time preference is significantly correlated only to agreeableness (at the 10 percent level).[4] This finding is confirmed by the significant correlation between delay acceptance and agreeableness in the truck-driver sample of Anderson et al. (2011).

Evidence on the link between risk preferences and the Big Five domains is equally mixed. Raw correlations between a lottery-choice measure of risk preference and personality traits in the data from Dohmen et al. (2010) indicate significant relationships between risk preferences and openness to experience (at the 1 percent level) and agreeableness (at the 5 percent level). Anderson et al. (2011) do not measure openness to experience. They do not find a significant correlation for risk preference and agreeableness, but report a weak correlation between risk preference and neuroticism (0.05 in absolute value), which is significant at the 10 percent level. This finding is in line with the significant positive association between risk aversion and neuroticism reported by Borghans et al. (2009). Other researchers (e.g. Zuckerman, 1994) have related risk preferences to sensation seeking, a facet of extraversion in the Big Five taxonomy, and found mixed evidence. Whereas Bibby and Ferguson (2010) report a significant correlation between a measure of loss aversion and sensation seeking ($r = 0.27$), Eckel and Grossmann (2002) find no evidence of an association between risk preferences and sensation seeking.

Evidence on the link between social preferences and personality is somewhat stronger. Dohmen et al. (2008) relate survey measures of social preferences to measures of the Big Five using data from the SOEP and find significant associations between trust, as well as positive and negative reciprocity and personality traits. Trust is related positively to agreeableness and openness to experience, and negatively to conscientiousness and neuroticism; while positive reciprocity is positively associated with all five personality factors, negative reciprocity is related negatively to conscientiousness and extraversion, and positively to

[3]The effect sizes of the correlations between preference and personality measures are all smaller than 0.1 in absolute value.

[4]We report this data in Table 2.3.

neuroticism. A link between extraversion and behavior in the dictator game, which can be interpreted as a measure of altruism, has been established by Ben-Ner and Kramer (2010).

This review is structured as follows. Section 2.2 describes our three data sets. In Section 2.3 we introduce our research strategy for investigating the link between personality and preferences. Section 2.4 presents evidence on the correlation between measures of personality and measures of preferences. In addition it contains an assessment of the explanatory power of preferences and personality in explaining important life outcomes. Section 2.5 concludes.

2.2 Data and Measures

In this section, we provide a description of the three complementary data sets that we employ for our analysis. Before we present our experimental and survey measures in detail, a few comments on identification are warranted. Economists typically try to infer preferences from choices, the so-called revealed preference approach. For example, one might surmise that a person who does not wear a safety belt an who invests in risky stocks has a preference for taking risks. It is, however, easy to show that the same behavioral pattern is compatible with very different risk preferences if other factors affect the person's decisions. For example, differences in beliefs about how risky driving without a safety-belt or investing in stocks actually is may affect decisions equally strong than underlying risk preferences. The problem is that the decision context is uncontrolled and person specific, rendering precise statements about preference parameters very difficult.[5] This is why economists run experiments to infer preferences. In a typical choice experiment subjects make decisions in a well-controlled decision environment. In risk experiments, for example, stakes and probabilities are fixed and the action space is identical for every subject. Observing subjects' decisions in a controlled experimental environment therefore rules out many potentially confounding factors, allowing a more precise identification of preferences. Even in an experiment, however, the identification of preferences is limited (see Manski (2002) for a thorough discussion on the identification of experimental outcomes). The same observed action can reflect different risk attitudes, for example, if the experimental subjects dispose of different wealth levels and the curvature of the utility function is not invariant to

[5]Conceptually identical problems apply to the identification of traits, such as ability, physical strength and personality characteristics from observed performance on tasks, when performance also depends on other unobserved factors such as time, energy and attention devoted to the task. An illuminating discussion of the identification problem is provided in section 3 of Almlund et al. (2011).

wealth levels. Despite these limitations experiments deliver much more precise behavioral outcomes than non-experimental observations. In strategic situations, which are relevant for measuring trust and reciprocity, we are able to elicit not just an action but a complete strategy. With field observations this is impossible. The relevance of eliciting a strategy is obvious: Suppose one observes a second mover who defects in a cooperation context, in response to a non-cooperative act of a first mover. This could reveal selfish preferences as well as reciprocal preferences. Disentangling the two requires knowledge about what the decision maker would have done, had the first mover cooperated. Eliciting a strategy instead of observing only actions does exactly this. Experimental observations have the additional advantage over survey responses that decisions have immediate monetary consequences. This is of obvious importance, for example, for identifying altruism. There is a big difference between simply stating altruistic preferences and revealing them in a costly manner.

2.2.1 Experimental Data

The first data set consists of decisions from laboratory experiments among university students. We ran a series of simple incentivized choice experiments to elicit preferences concerning risk taking, discounting, positive and negative reciprocity, and trust as well as altruism.[6] Table 2.1 presents an overview of the experiments and provides a short description of the elicitation methods and the obtained behavioral measures. Four important features about our experimental design are worth noting. First, subjects took part in two very similar experiments each for risk taking, discounting, trust and positive reciprocity. This allows us to average over both outcomes for each subject in order to minimize measurement error. Second, to reduce spillovers between different choices, we ran the experiments not in one single session but in two sessions, which were scheduled one week apart.[7] Third, to reduce possible income effects with respect to outcomes within a session, we gave feedback about experimental outcomes only at the end of an experimental session. Fourth, the vast majority of subjects in the experiments had never taken part in an experiment before. This eliminates possible confounds in behavior due to previous experiences in similar experiments. In total, 489 students from different majors from the University of Bonn par-

[6] For a detailed description of the experimental procedures see Falk et al. (2011).

[7] We reversed the order of the sessions for half of the subjects. Statistical tests reveal no significant order effects.

ticipated.[8] The experiments were run at the Laboratory for Experimental Economics at the University of Bonn (BonnEconLab). We used zTree (Fischbacher, 2007) as experimental software and recruited subjects using ORSEE (Greiner, 2004). Each session lasted about two hours, and average earnings were 64 Euros.

Table 2.1: Overview of the experimental measures in data set from laboratory experiments amon university students

Preference	**Experiment**	**Measure**
Time	Two lists of choices between an amount of money "today" and an amount of money "in 12 months".	Average switching point over both lists of choices from the early to the delayed amount.
Risk	Two lists of choices between a lottery and varying safe options.	Average switching point over both lists of choices from the lottery to the safe option.
Positive Reciprocity	Second-mover behavior in two versions of the trust game (strategy method).	Average amount sent back in both trust games.
Negative Reciprocity	Investment into punishment after unilateral defection of the opponent in a prisoner's dilemma (strategy method).	Amount invested into punishment.
Trust	First mover behavior in two versions of the trust game.	Average amount sent as a first mover in both trust games.
Altruism	First mover behavior in a dictator game with a charitable organization as recipient.	Size of donation.

Preference Measures

Risk Preferences To elicit risk attitudes we adapted the design from Dohmen et al. (2010). Subjects were shown a list of binary alternatives, a lottery and a (varying) safe option. The lottery was the same for each decision: If they chose the lottery participants could receive either 1000 points or zero points with 50 percent probability each. The safe option increased from row to row, starting from a value of (close to) zero, and increasing

[8]Out of these 489 students, 80 took part in a pretest of the study. Most of these 80 subjects had taken part in an experiment before. The pretest did not include the experiments on altruism and negative reciprocity.

up to a value of (close to) the maximum payoff of the lottery. To reduce measurement error subjects participated in two risk experiments. The choice list of the second experiment was simply a perturbed version of the first one. Perturbations were constructed such that a randomly drawn integer value between -5 and +5 was added to the safe option in every choice, corresponding to perturbations of maximally 5% of the step size of the increase in the safe option. The complete list of choices was shown to subjects on the first screen. Each choice situation was then presented on a separate screen, where subjects entered their respective choice. Subjects were informed that one choice in each list would be selected randomly and paid. Subjects with monotonic preferences should choose the lottery for lower safe options and switch to the safe option when the latter reaches or exceeds the level of their certainty equivalent. Thus switching points inform us about individual risk attitudes. The earlier a subject switches to the save option the less she is willing to take risks. For our analysis we constructed a risk preference measure using the average of the two switching points from the two experiments.[9]

Time Preferences To measure individuals' time preferences we implemented a procedure very similar to the one for risk attitudes. In the discounting experiments, subjects were given two lists of choices between an earlier amount of money ("today"), which was the same in all choices, and an increasing delayed amount of money ("in 12 months"). In the first row, the early amount was equal to the delayed amount. Delayed amounts increased from row to row by 2.5%. As for risk preferences subjects participated in a very similar second discounting experiment with small perturbations of delayed amounts between +0.5 and -0.5 percentage points. One choice in each of the two lists was randomly selected for payment. Payments resulting from the two experiments were sent to subjects via regular mail. If a subject chose the early amount, the payment was sent out on the day of the experimental session. If a subject chose the delayed amount, the payment was sent out with a delay of 12 months.[10] The switching point from early to delayed payment informs us about a subject's time preference. Subjects who switch later discount the future amount by more (i.e., are less patient) than subjects who switch earlier.[11] Our measure of individual discounting is

[9]If subjects switched between the lottery and the safe option more than once, we took the average switching row as an estimate of their certainty equivalent. This happened in 16 % of the cases in the first experiment on risk taking, and in 11 % of the cases in the second experiment.

[10]Keeping the payoff mode identical over both time horizons rules out credibility concerns.

[11]For subjects, who switched more than once, we took the average switching row as an estimate of their discount rate. This happened in 5 % of the cases in the first experiment on time discounting, and in 7 % of the cases in the second experiment.

the average switching row in both lists. To ease interpretation of the correlations reported below, we recode the measure, such that higher values imply earlier switching rows, i.e., a higher level of patience.

Trust We elicited trust from first-mover behavior in the so-called trust game (Berg, Dickhaut, and McCabe, 1995). We conducted two versions of the trust game. In one version, the amount sent by the first mover was doubled by the experimenter, whereas in the second version the amount was tripled. Every subject was in the role of the first and of the second mover twice.[12] Both trust games were incentivized, i.e., every (relevant) decision was paid. In the role of a first mover, subjects could choose to send any amount in $\{0, 50, 100, \ldots, 500\}$ points to the second mover. All interactions in the trust game as well as in all other social preference experiments were one-shot and anonymous (perfect stranger matching protocol). The average amount sent as a first mover in both trust games constitutes our experimental measure for trust: Subjects who send higher amounts of money are those who display higher levels of trust.

Positive Reciprocity To elicit positive reciprocal inclinations we measure subjects' second-mover behavior in the trust game (see above). We implemented the strategy method (Selten, 1967). This means that for every possible amount sent by the first mover, subjects were asked to indicate how much they wanted to send back. The actual decision of the first mover determined which of these decisions became payoff relevant. The average amount sent back as a second mover in both trust games was taken as individuals' willingness to reciprocate, such that higher values imply a higher willingness to reciprocate.

Negative Reciprocity To measure subjects' willingness to engage in costly punishment of unfair behavior, we conducted a prisoner's dilemma with a subsequent punishment stage.[13] In the punishment stage, subjects could choose to invest points in order to deduct points from their opponent. Punishment was costly. Again, we implemented the strategy method. Before taking their decisions in the first stage of the experiment (i.e., in the prisoner's dilemma) subjects were asked to indicate how many points they wanted to deduct from the other player in case he cooperated or defected, for both own cooperation and own defection. Then they played a simultaneous prisoner's dilemma. The outcome of the first stage determined which choice of the second stage became payoff relevant. The

[12] Overall, we therefore ran four trust games.

[13] The design of the experiment was adapted from Falk, Fehr, and Fischbacher (2005)

chosen investment into punishment after unilateral defection of the other player served as a measure of an individual's willingness to reciprocate negatively.

Altruism To measure altruistic behavior we had subjects take part in a modified dictator game in which the recipient was a charitable organization (adapted from Eckel and Grossmann, 1996). Subjects were endowed with 300 points and had to decide how much of this endowment to donate to a charitable organization.[14] This decision serves as our experimental measure of subjects' altruistic inclination.

Personality Measures

Big Five As part of the study, subjects were given a paper-and-pencil survey, which they were asked to fill out at home and return to us via mail.[15] Of the 489 subjects, 319 completed the survey and sent it back to us. The survey included the NEO-FFI version of the Big Five (Costa and McCrae, 1989). During the experimental sessions, all 489 subjects also answered a shorter version of the NEO-FFI: the BFI-S, a subset consisting of 15 items. The BFI-S has been developed by Gerlitz and Schupp (2005) and was also part of the 2005 and 2009 waves of the SOEP. Correlations between the long version and the short version of the Big Five differ between the five personality dimensions. The lowest correlation is $r = 0.48$ for openness, and the highest is $r = 0.71$ for conscientiousness (all p-values < 0.001). We constructed our Big Five measure in that we use data from the long version whenever available, while for the remaining subjects we refer to the short version. That way, we have measures of the Big Five domains for all 489 subjects.

Locus of Control The paper-and-pencil survey included 10 items that allows us to construct a measure of the locus of control for the 319 individuals who filled in the survey. These 10 items have been adapted from Rotter (1966) and they have also been implemented in the 2005 wave of the SOEP. The personality construct of locus of control assesses how much people believe they have control over their life outcomes, or how much their lives are determined by forces that are outside of their control, such as luck or faith. We constructed the measure such that higher values represent a more internal locus of control, i.e., the belief that the person can influence their life outcomes. Lower values represent a more

[14] Subjects could choose a charitable organization from a list, or name one themselves.

[15] We also handed out stamped envelopes with the address of our research institute, in order to minimize additional costs for returning the survey to us.

external locus of control.

2.2.2 Representative Experimental Data

The second data set we employ consists of experimental data for a representative sample of the German population.[16] This data set is used to assess whether the findings from the sample of university students can be corroborated in a representative sample. Subjects' risk and time preferences were elicited, and we again have information on participants' personality. The data used here stem from a study conducted in 2005 and contains information on 1012 individuals. For a detailed description of the study and its procedures see Dohmen et al. (2010).

Preference Measures The experiments on risk and time preferences were similar to the ones we used in the laboratory experiments. In both experiments subjects had to make multiple decisions in a list of choices. To elicit their risk preferences we had subjects choose between a lottery, which remained the same in all choices, and safe options, which increased in their value. As in the experiments discussed above, the switching point informs us about the individual's willingness to take risks. Similarly, to elicit individuals' time preferences we had all participants make a number of intertemporal choices. They had to decide between an amount "today" and a larger amount "12 months" later. The early amount remained the same in all choices. The first delayed amount presented to subjects was devised to imply a 2.5% return on the early amount assuming semi-annual compounding. In the subsequent choices the delayed payment was gradually increased and was calculated such that the implied rate of return rose in steps of 2.5 percentage points. Again, the switching points from the early to the delayed option inform us about the subjects' time preferences.

Personality Measures The five personality domains were assessed using the BFI-S (see Section 2.1.2 for a more detailed description).

2.2.3 Representative Panel Data

The third data set we use stems from the SOEP, a large panel data set that is representative of the adult population living in Germany (see Schupp and Wagner (2002) and Wagner,

[16] The same data set is used in Dohmen et al. (2010).

Frick, and Schupp (2007) for a detailed description of the SOEP). We use information from eight waves collected in the years between 2003 and 2009. In each of these waves more than 20,000 individuals were interviewed. The SOEP combines extensive sociodemographic information with various measures of attitudes, preferences and psychological traits. In particular, the SOEP includes survey items relating to all personality and preference measures that we discuss in the previous sections.

Personality and economic preference measures were elicited several times between 2003 and 2009. To construct a measure for each individual, we use the maximum available number of observations of a given measure. If several measures of personality and preferences are available, we take the average of the standardized measures of all years in which this measure was elicited. The resulting average is then standardized as well. In case a particular measure was elicited only in one wave (e.g., as it is the case for patience) we just take the standardized measure from that respective year. We restrict the sample to individuals for whom we have information about each personality and preference measure. This results in a sample size of 14,243 individuals.

Preference Measures As a measure for time preference we use answers to the following survey question: "How would you describe yourself: Are you generally an impatient person, or someone who always shows great patience?".[17] Participants gave an answer on an 11-point scale where zero means "very impatient" and 10 means "very patient". This survey question was implemented in the SOEP only in 2008. The risk preference question is worded in the same manner: "How do you see yourself: Are you generally a person who is fully prepared to take risks, or do you try to avoid taking risks?" Answers were given on an 11-point scale, where zero means "unwilling to take risks" and 10 means "fully prepared to take risks". This question was included in the 2004, 2006, 2008 and 2009 waves. The general risk question has been studied in various papers and has been validated using incentivized experiments in representative samples as well as through behavioral evidence in Dohmen et al. (2011a). In 2005 the SOEP contained six items to measure reciprocal inclinations, three items each on positive and negative reciprocity. Examples for positive and negative reciprocity are as follows: "If someone does me a favor, I am prepared to return it" and "If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the costs". Participants expressed how well these six statements apply to them on a seven-point Likert

[17]The behavioral validity of this question with respect to incentivized experiments is documented in Vischer et al. (2011).

scale. For a detailed description see Dohmen et al. (2009). Standard trust questions were included in the 2003 and 2008 waves, using three sub-statements about whether "one can trust people", whether "in these times one can't rely on anybody else" and whether "when dealing with strangers it is better to be cautious". Answers were given on a five-point scale ranging from "totally agree" to "totally disagree". Finally, our survey measure for altruism is the answer to the question of how important it is for the participant "to be there for others". Answers were given on a four-point scale. The altruism question was asked in the 2004 and 2008 waves.

Personality Measures The 2005 and 2009 waves of the SOEP contained the BFI-S questionnaire, developed by Gerlitz and Schupp (2005). The locus of control was elicited in 2005 using Rotter's (1966) locus of control scale. Both inventories were also used in our laboratory experimental data (see Section 2.1.2 for more details on the BFI-S and the locus of control scale).

2.3 Research Strategy

To answer the question of whether measures of personality and economic preferences are closely linked we first study the raw correlations between these measures. High correlations would indicate some degree of substitutability. Low correlations, conversely, would suggest that the two measurement systems are complementary concepts in explaining heterogeneity in behavior. Whether a correlation should be interpreted as "high" or "low" is of course always debatable. We therefore first look at statistical significance levels. Statistical significance, however, can also be found for correlations that are low in terms of effect size (Cohen, 1992). Following conventions in the social sciences we interpret effect sizes, i.e., correlations r, as rather "low" if r is between 0.1 and 0.3, as "medium" if r is between 0.3 and 0.5 and as "large" if r is larger than 0.5. Because the analysis of correlations is restricted to linear relations, we also check for potential non-linear associations by conducting non-parametric regressions. In particular, we look at kernel-weighted local linear polynomial regressions.

We then check to see whether measures of personality and preferences are substitutes or complements in terms of their explanatory power for life outcomes. In particular, we conduct linear regressions and assess the explanatory power of the two concepts by reporting levels of adjusted R^2. In these regressions, measures of personality and preferences are included individually as well as jointly. If the two measurement systems are substi-

tutes, adjusted R^2 in the combined regressions should not be distinctly higher than in regressions that include only one of the two concepts. The opposite should hold for complements. Additionally, we investigate model selection criteria in these regressions. We check for robustness using binary and ordered choice models as well as more comprehensive specifications including square terms and cross-products of all regressors.

2.4 Results

In this section we discuss our main findings. To ease comparison between data sets and measures, we standardized all experimental as well as all personality measures for the data analysis.

2.4.1 Correlation Structure

Experimental Data

Table 2.2 displays the 36 raw correlations of the personality and economic preference measures obtained from the laboratory experiments. A first inspection of Table 2.2 reveals that only 11 of these 36 correlations are statistically significant at the 5% or 1% level.[18] All correlation coefficients are smaller than 0.3 in absolute value. Hence there is no correlation with a "medium" effect size or larger. Moreover, of the 36 correlations only 11 exceed 0.1 in absolute value and only 1 slightly exceeds 0.2.[19]

Table 2.2 also shows that among all personality factors agreeableness exhibits the highest and statistically most significant correlations with measures of economic preferences. It is significantly correlated with measures for positive and negative reciprocity, trust and altruism (all p-values < 0.01) as well as with time preference (p-value < 0.05). Correlations with social preferences range between 0.1 and 0.3 in absolute value, indicating a small effect size according to the classification of Cohen (1988). The high frequency of significant correlations of agreeableness with social preferences is not surprising as the former is defined as "the tendency to act in a cooperative, unselfish manner,..." (see Table 2.5).

The finding of only moderate correlations between preference and personality measures does not necessarily indicate that these constructs are weakly connected; it indicates only

[18] Five additional correlations are weakly significant, i.e., significant at the 10% significance level.

[19] Results qualitatively stay the same when investigating Spearman correlations instead of Pearson correlations (see Table 2.6 in the appendix). Moreover, when looking at a potential linear mapping, i.e., linear regressions of either the Big Five on preferences or vice versa, R^2 is always below 10%.

Table 2.2: Pearson correlation structure experimental data set

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	LoC
Time	0.0370	0.0057	−0.0084	0.1026**	−0.0518	0.0847
Risk	−0.0379	−0.0611	0.0762*	0.0202	−0.1201***	0.0434
Pos. Reciprocity	0.1724***	0.0140	0.0211	0.2042***	0.0361	0.0152
Neg. Reciprocity	−0.0885*	−0.0393	0.0943*	−0.1451***	−0.0136	−0.1418**
Trust	0.1232***	−0.1300***	0.0004	0.1665***	−0.0134	−0.0140
Altruism	0.1242**	−0.0979*	0.0249	0.1911***	0.0847*	0.0480

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Correlations between economic preferences and the Big Five were calculated using 394 - 477 observations. Correlations between economic preferences and locus of control were calculated using between 254 - 315 observations. All measures are standardized.

that there are weak linear relations. For example, a perfect U-shaped relation between a personality factor and a preference would result in an insignificant linear correlation. To explore the possibility of non-linear relationships we therefore estimate kernel-weighted local linear polynomial regressions.[20] In each regression, we restrict the sample to a range of four standard deviations around the mean of each variable to circumvent an analysis biased by outliers. Therefore, the results are calculated using 70% to 97% of all observations. The predicted regressions are displayed in Figure 2.2. Although sometimes there are small deviations from linearity at the boundaries, the overall picture strongly suggests a linear relation in the vast majority of combinations.

Summarizing our analysis of the laboratory experimental data, we find that associations between preference and personality measures are linear and that the degree of association is rather low, suggesting a complementary relationship. We next turn to the question of whether the correlation patterns observed in student samples can be replicated in a sample that is representative of the adult population.

[20] We use the Epanechikov kernel and bandwidth is selected via the plugin estimator of the asymptotically optimal constant bandwidth.

Representative Experimental Data

Table 2.3: Pearson correlation structure representative experimental data

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Time	−0.0080	−0.0682	−0.0655	−0.0830*	−0.0602
Risk	0.1356***	−0.0720	0.0757	−0.0941**	−0.0290

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. All measures are standardized.

Table 2.3 shows the correlations between the outcomes from the risk and time experiments and the personality traits. As above, the measure for time is reversed so that higher values indicate higher patience. In terms of significance the pattern is similar to the one in the laboratory study. Only one correlation is significant at the 1%-level, one is significant at the 5%-level and one is significant at the 10%-level. In terms of effect size, only the coefficient of the association between openness and risk preferences exceeds the 0.1 benchmark to be classified as a small correlation (Cohen, 1988).[21] Interestingly, the sign is positive, in contrast to our laboratory data. The other two significant coefficients are even smaller. The analysis of representative data therefore confirms that the level of association between preference personality measures is rather small. However, we can draw this conclusion only with respect to time and risk preferences, as we do not have experimental data on trust and social preferences. We next analyze whether these findings also hold when looking at all preference measures in a large representative sample.

Representative Panel Data

In this section, we study whether our findings from the experiments generalize to a large representative sample using survey rather than experimental instruments for measuring economic preferences. Table 2.4 shows the raw correlations between personality measures and economic preferences using 14,243 observations from the SOEP. Given the large number of observations it is not surprising to find a large number of significant correlation coefficients (p-values < 0.05 for all correlation coefficients). In terms of effect size, however, only two correlations are of "medium" size, i.e., larger than 0.3. Of the reported 36 correlations, 18 can be classified as "small", whereas 16 correlations are even below 0.1. This confirms

[21] Results qualitatively stay the same when investigating Spearman correlations instead of Pearson correlations (see Table 2.7 in the appendix).

Table 2.4: Pearson correlation structure between personality measures and economic preferences from SOEP observations

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	LoC
Time	0.0183^{**}	0.1122^{***}	-0.0415^{***}	0.3122^{***}	-0.0584^{***}	0.0681^{***}
Risk	0.2793^{***}	-0.0400^{***}	0.2601^{***}	-0.1454^{***}	-0.0996^{***}	0.1521^{***}
Pos. Reciprocity	0.1814^{***}	0.2520^{***}	0.1473^{***}	0.1842^{***}	0.0872^{***}	0.0954^{***}
Neg. Reciprocity	-0.0522^{***}	-0.1558^{***}	-0.0264^{***}	-0.3756^{***}	0.0612^{***}	-0.2154^{***}
Trust	0.1272^{***}	-0.0680^{***}	0.0575^{***}	0.0945^{***}	-0.1919^{***}	0.2094^{***}
Altruism	0.1756^{***}	0.1495^{***}	0.1670^{***}	0.2557^{***}	0.0908^{***}	0.0874^{***}

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Correlations are calculated using 14,243 observations. All measures are standardized.

the overall picture that emerged from the analysis of the two experimental data sets.[22] A closer comparison of the SOEP survey measures with our experimental measures further reveals large similarities. As reported above, 11 correlations are significant at the 5% level in the experimental data. Ten of these correlations have the same sign and are significant at the 1% level using survey data. Moreover, as it is the case in the laboratory data set, the personality trait agreeableness exhibits the highest correlations with economic preferences, in particular social preferences. Although there are small differences in the results compared with the experimental data set (i.e., seven of the 36 correlation coefficients show a different sign), the general pattern emerging from the SOEP measures is consistent with our previous findings. Of the seven correlation coefficients only two are (weakly) significant in the experimental data set. Nevertheless, the inconsistency of signs brings into question the conjecture that correlations are universally identical (i.e., identical irrespective of age or other person characteristics). We return to this aspect in the final section.

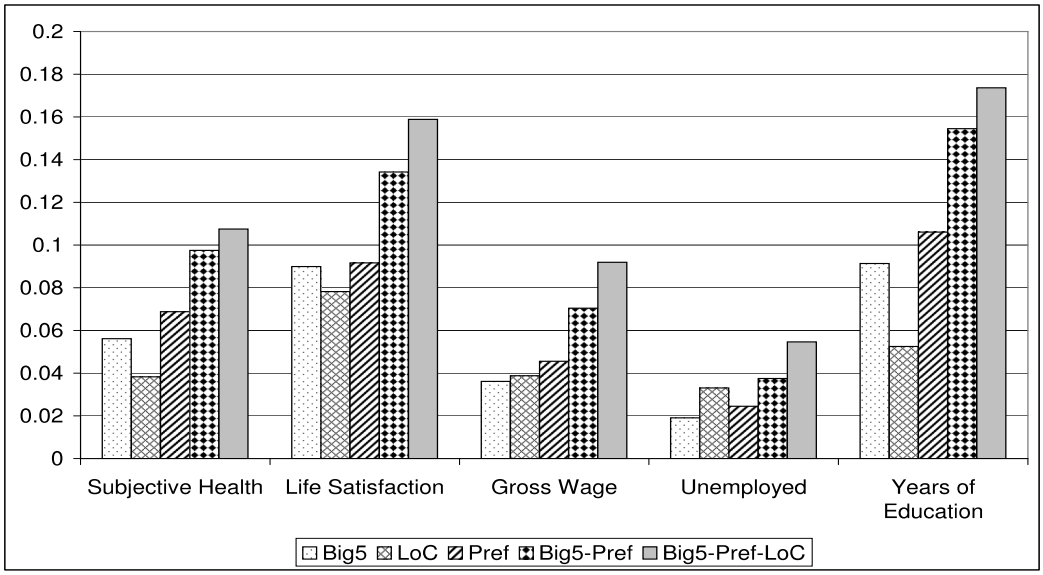
We conclude this section with an analysis of potential non-linearities between our SOEP preference and personality measures. As for the laboratory experimental data, we perform kernel-weighted local linear polynomial regressions restricting the sample in each regression to four standard deviations above and below the mean. The resulting subsamples represent 92% to 97% of the observations of the main sample. The predicted functions presented in Figure 2.3 show no particular non-linearities, except for some splines at the left ends of the considered range. Thus, analogous to the experimental data set, it is not the case that systematic non-linearities bias correlation coefficients.

2.4.2 Explanatory Power for Life Outcomes

All reported correlation structures indicate that personality and preference measures are far from perfectly substitutable. To determine whether they actually complement each other, we now analyze their explanatory power with respect to important life outcomes. To that end we again use data from the SOEP. In particular, we consider the following outcomes: subjective health, life satisfaction, gross wage, being unemployed and years of education. For each outcome we estimate linear regression models in which outcomes are regressed on the set of economic preferences, the Big Five and the locus of control, separately as well

[22] Results qualitatively stay the same when investigating Spearman correlations instead of Pearson correlations (see Table 2.8 in the appendix). Moreover, when looking at a potential linear mapping, i.e., linear regressions of either the Big Five on preferences or vice versa, R^2 is always around 15% with the exception of agreeableness, where R^2 reaches 28%.

Figure 2.1: Adjusted R^2 for Life Outcomes



Adjusted R^2's for linear regressions for life outcomes. The number of observations available varies for the different life outcomes: subjective health (14,218), life satisfaction (14,214), gross wage (7,199), unemployed (9,095), and years of education (13,768). Gross wage measures the gross hourly wage.

as jointly.[23] The idea is to assess the explanatory power of each concept in isolation and in combination. This enables us to check the extent to which explanatory power increases when combining the concepts and thus allows us to reach conclusions regarding the degree of their complementarity. The criterion used to compare differences in explanatory power is adjusted R^2.

All life outcomes we use come from the 2009 wave of the SOEP. Subjective health was measured on a five-point-scale, from "very good" to "bad". We reverse the answer scale such that higher values indicate a better subjective health status. Life satisfaction was elicited using the question "How satisfied are you with your life, all things considered?", which was answered on an 11-point-scale (with higher values indicating higher life satisfaction). Our measure for gross hourly wage is the gross monthly wage divided by monthly working hours.[24] Unemployment status is a binary variable equal to one if the person was unemployed at the time of the survey and zero otherwise. The variable years of education is created by adding up years of schooling and additional occupational training (including

[23]The corresponding regressions are shown in Table 2.9 in the appendix.

[24]Monthly working hours are calculated as the average weekly working hours multiplied by four.

university).[25]

Figure 2.1 shows adjusted R^2's for the different life outcomes. R^2 values for the three concepts – Big Five, Locus of Control and economic preferences – in isolation range from 1% to 10% and vary both between concepts and outcomes. Thus, they contribute to explaining heterogeneity in important life outcomes.[26] More important in light of our research question, however, is that the explanatory power is considerably larger when combining the Big Five, the locus of control and economic preferences compared to using each concept individually. Moreover, explanatory power is always maximized when all three concepts are included in the regression, hereafter referred to as the full model. In this case, resulting adjusted R^2 values reach levels of about 6% to 18%. This clearly indicates the existence of important complementarities among the different concepts.[27]

Because the question here is one of model selection, we also employ model selection criteria (in particular the Akaike and Bayesian information criterion) to check whether the full model is also chosen by model selection criteria. As can be seen in Table 2.10 in the appendix this is the case for all life outcomes considered, corroborating our previous results. We perform the same analysis using binary and ordered choice models when appropriate. Again, the full model is chosen by the model selection criteria in all cases. As another robustness check we consider more flexible models: Along with including each predictor linearly in our regressions we also include square terms and all possible cross-products (see Table 2.11 in the appendix). Again the full model obtains the highest adjusted R^2 measures when using ordinary-least-squares estimation and is also chosen by the information criteria in nearly all cases.[28] Results are again robust for employing binary and ordered choice models when appropriate. Moreover, in all models considered the joint hypothesis that all coefficients are equal to zero is always rejected at the 1% level (Tables 2.10 and 2.11 in the appendix). In summary, sizeable complementarities among the different concepts are corroborated in all robustness checks.

[25]For each school degree and occupational training (including university) official standard graduation times in years are used for the calculation.

[26]In the explanation of life outcomes such as gross wages, unemployment and years of education the preference for work versus leisure would probably play a key role. However, no question related to this preference was included in the survey.

[27]For an overview over the raw correlations between each preference and personality trait and life outcomes see Figure 2.4 and 2.5 in the Appendix to this chapter.

[28]Only the Bayesian information criterion chooses a model just including the locus of control when it comes to explaining gross wage and unemployment. However, this is not surprising given the number of regressors included and the tendency of Bayesian information criterion to choose parsimonious models.

2.5 Discussion

In this review we examine the relation between economic preferences and personality using three different data sets. We find no indication for a strong linear or a non-linear association between the two. Thus we conclude that the two concepts cannot substitute for each other. In fact, with regard to explaining heterogeneity in life outcomes, we find that the two concepts play complementary roles. Our findings imply that researchers in economics and psychology can benefit greatly from the respective disciplines when looking for potential sources of heterogeneity in life outcomes.

The finding of a rather low association between economic preferences and psychological measures of personality is perhaps not surprising. First, both concepts are constructed in very different ways. Whereas preferences are rooted in utility theory, derived in terms of specific functional forms of utility functions, the Big Five personality indicators originate in language analysis. Second, the Big Five measure rather broad aspects of personality. In particular, each dimension of the Big Five is by itself already an aggregation of different attitudes or subfacets. Thus, although our results show low associations between personality and economic preferences, we cannot exclude the possibility that there is a stronger degree of association between economic preferences and subfacets of the five personality traits. The trait extraversion, for example, comprises different attitudes, such as being "relatively outgoing, gregarious, sociable, and openly expressive" (see Table 2.5), measured by 12 different questions in the NEO-FFI or three different questions in the BFI-S. In other words, each personality measure is not only comprises multiple items, but more importantly captures distinct aspects of a character trait. Economic preferences, conversely, are defined more narrowly. For example, the concept of time preferences refers to the individual's willingness to abstain from something in the present in order to benefit from that decision in the future. Although this concept is applicable to different domains (e.g., to health outcomes or financial decision making) the underlying concept remains the same and is measured by standard incentivized experiments or survey items as employed in this study. In this sense, our preference measures might resemble the subordinate aspects of the five personality factors.

Third, the finding of strong complementarities between economic preferences and personality measures may simply reflect conceptual differences in the way economic and psychological models are constructed. The economic model explains heterogeneity in behavior in terms of three distinct components: preferences, beliefs and constraints, such as abili-

ties. In contrast, psychological measures such as the Big Five include notions of preferences as well as beliefs and constraints. In other words, in our analysis we correlate economic preferences at least partly with beliefs and constraints, which by construction should not necessarily be correlated. A good example is conscientiousness. Being able and willing to work hard and being organized comprises aspects of both, preferences and personal abilities. Likewise, emotional instability, which is part of the neuroticism facet, is related to personal inability rather than a preference. Even more extreme is the case of the locus of control, which is clearly a belief rather than a preference. This does not rule out the possibility that the two concepts are related, for example, because an external locus of control is conducive to the development of impatient behavior: if it does not pay off to invest because life circumstances are predominantly determined by circumstances beyond my control, the willingness to forgo current consumption and wait in order to earn a return in the future makes little sense. Yet, beliefs and preferences are two distinct concepts.

The main focus of this review is the rather weak association and complementary nature of economic and psychological measures of personality. We do not discuss the specific signs of the correlations or ways to integrate personality into the economic model. Important work in this direction has been done by Almlund et al., 2011a. Many signs of the correlations reported above are consistent across the three data sets, in particular those that are significant. For example, in all three data sets risk attitudes and extraversion are positively correlated, and risk and neuroticism are negatively correlated. There are important exceptions, however. In the student sample, for example, risk attitudes and openness are negatively correlated, whereas they are positively and significantly negatively correlated in the two representative data sets. These and other inconsistencies raise important questions. One possible reason for finding different signs is the use of different elicitation methods for economic preferences (experiments and survey responses). Another possibility is that the reported correlations vary over the life-cycle. If traits develop with different speed and at different points in life correlations should vary with age. This could explain differences between a relatively young student sample and the representative samples. Not much is known about how economic preferences develop over the life-cycle but at least for risk attitudes there seems to be a robust and large negative age effect on willingness to take risks (Dohmen et al., 2011a). Another possibility is that preferences and personality are generically differentially correlated between specific groups of the population (e.g., varying by gender, age, height or education). From an evolutionary perspective the co-evolution

of traits may serve different purposes depending on specific life circumstances. It may be "optimal" for one subgroup of the population to develop a positive correlation among particular traits, whereas for another subgroup it is adaptive to form a negative correlation. More work needs to be done to uncover potential group-specific correlations between personality and preferences.

The approach taken above is agnostic in the sense that we simply correlate existing and important measurement systems as they are. We think this is an important exercise but it can only be a first step. What is needed is the development of a comprehensive framework that combines insights from the approaches taken by economists and psychologists to capture sources of heterogeneity in behavior. It is surprising that the Big Five apparently misses important preferences such as attitudes towards risk and time. Similarly, the economic model is incomplete not only with respect to important preferences, but also with respect to heterogeneity in abilities and beliefs. In the standard economic framework, beliefs are assumed to be endogenous to the strategic situation and formed in a rational way. Perhaps, with the exception of interpersonal trust, beliefs are typically assumed to follow common prior assumptions and rational updating. The role of the locus of control in explaining fundamental life outcomes on top of preferences, however, reveals the importance of enduring and individual specific belief systems. Other examples include optimism, pessimism, religious beliefs and ideological beliefs. The stability of belief heterogeneity is not well understood. It probably originates in different priors inherited from parents, self-selection into peer groups and institutions with reinforcing belief characteristics and boundedly rational belief formation, such as selected perception, non-Bayesian updating and ego utility (Köszegi, 2006). Regardless of the precise channels that support enduring heterogeneous beliefs, economics would largely benefit from measuring and including them in explanations of economic outcomes. In addition, economists have started to model the fact that preferences and beliefs are intimately related and not separable as traditionally assumed. In fact, people often want to believe certain things, for example, in terms of being liked by others or being better than others (overconfidence). Finally, another important extension of the economic model would be the measurement of person-specific abilities. Whereas IQ has become a standard individual-specific characteristic to be included in outcome regressions, little work has acknowledged the importance of other competencies captured by Big Five traits, for example, the role of conscientiousness for educational or labor market outcomes.

A2 Appendix to Chapter 2

Table 2.5: Definitions of the Big Five Domains

Big Five Domain	**APA Dictionary Definition**
Openness	Individual differences in the tendency to be open to new aesthetic, cultural, and intellectual experiences.
Conscientiousness	The tendency to be organized, responsible, and hardworking; located at one end of a dimension of individual differences: conscientiousness vs. lack of direction.
Extraversion	An orientation of one's interests and energies toward the outer world of people and things rather than the inner world of subjective experience; includes the quality of being more outgoing, gregarious, sociable, and openly expressive.
Agreeableness	The tendency to act in a cooperative, unselfish manner; located at one end of a dimension of individual differences: agreeableness vs. disagreeableness.
Neuroticism	A chronic level of emotional instability and proneness to psychological distress.

This table is in parts reproduced from Borghans et al. (2008).

Table 2.7: Spearman correlation structure representative experimental data

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Time	−0.0199	−0.0737	−0.0764*	−0.0829*	−0.0598
Risk	0.1315*	−0.0744	0.0661	−0.0854*	−0.0261

*, **, and *** indicate significance at the 10%, 5%, and 1% level. All measures are standardized.

Table 2.6: Spearman correlation structure experimental data set

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	LoC
Time	0.0388	0.0162	−0.0114	0.1077**	−0.0684	0.1063*
Risk	0.0027	−0.0486	0.0786*	0.0206	−0.0995**	0.0485
Pos. Reciprocity	0.1606***	0.0078	0.0177	0.2029***	0.0152	0.0441
Neg. Reciprocity	−0.0967*	−0.0221	0.0462	−0.083*	−0.0165	−0.1376**
Trust	0.1354***	−0.1198***	0.002	0.1696***	−0.002	−0.0648
Altruism	0.0969*	−0.0804	0.0034	0.2000***	0.0879*	0.0418

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. Correlations between economic preferences and the Big Five were calculated using 394 - 477 observations. Correlations between economic preferences and Locus of Control were calculated using 254 - 315 observations. All measures are standardized.

Table 2.8: Spearman Correlation Structure SOEP

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	LoC
Time	0.0233	0.1192	−0.0342	0.3099	−0.0643	0.0709
Risk	0.2632	−0.0500	0.2452	−0.1496	−0.1049	0.1426
Pos. Reciprocity	0.1835	0.2622	0.1547	0.1947	0.0808	0.1041
Neg. Reciprocity	−0.0616	−0.1767	−0.0426	−0.3853	0.0572	−0.2257
Trust	0.1224	−0.0693	0.0523	0.0788	−0.1889	0.2012
Altruism	0.1693	0.1501	0.1602	0.2416	0.0860	0.0843

All correlations are significant at the 1% level and are calculated using 14,243 observations. All measures are standardized.

Table 2.9: Outcome Regressions: Representative Experimental Data

	(1)	(2)	(3)	(4)	(5)
Life Outcomes	Subj. Health	Life Satisf.	Gross Wage	Unemployed	Years of Educ.
Openness	0.043***	0.123***	0.989***	-0.018***	0.667***
	(0.009)	(0.017)	(0.162)	(0.004)	(0.027)
Conscientiousn.	0.038***	0.106***	0.565***	-0.014***	-0.182***
	(0.009)	(0.017)	(0.161)	(0.004)	(0.026)
Extraversion	0.026***	0.134***	-1.201***	0.006*	-0.309***
	(0.009)	(0.017)	(0.154)	(0.004)	(0.026)
Agreeableness	0.033***	0.139***	-1.288***	0.023***	-0.146***
	(0.010)	(0.018)	(0.165)	(0.004)	(0.028)
Neuroticism	-0.140***	-0.186***	-1.009***	0.018***	-0.272***
	(0.009)	(0.016)	(0.158)	(0.004)	(0.026)
LoC	0.105***	0.307***	1.899***	-0.043***	0.421***
	(0.008)	(0.015)	(0.145)	(0.003)	(0.024)
Patience	0.024***	0.129***	-0.343**	0.001	-0.151***
	(0.008)	(0.015)	(0.136)	(0.003)	(0.023)
Risk	0.131***	0.076***	0.415**	0.003	0.210***
	(0.009)	(0.017)	(0.166)	(0.004)	(0.027)
Pos. Recip.	-0.035***	0.006	0.388***	-0.002	0.005
	(0.008)	(0.015)	(0.140)	(0.003)	(0.023)
Neg. Recip.	0.064***	0.039**	-0.329**	0.006*	-0.137***
	(0.008)	(0.015)	(0.147)	(0.003)	(0.024)
Trust	0.122***	0.308***	1.763***	-0.035***	0.587***
	(0.009)	(0.015)	(0.145)	(0.003)	(0.024)
Altruism	0.070***	0.072***	-0.780***	0.005	0.084***
	(0.009)	(0.016)	(0.152)	(0.003)	(0.025)
Constant	3.300***	6.852***	16.100***	0.099***	12.346***
	(0.007)	(0.014)	(0.131)	(0.003)	(0.021)
Observations	14,218	14,214	7,199	9,095	13,768
Adj. R-squared	0.108	0.159	0.0919	0.0547	0.174

*, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively. All measures are standardized.

Figure 2.2: Kernel-weighted local linear polynomial regressions using experimental data

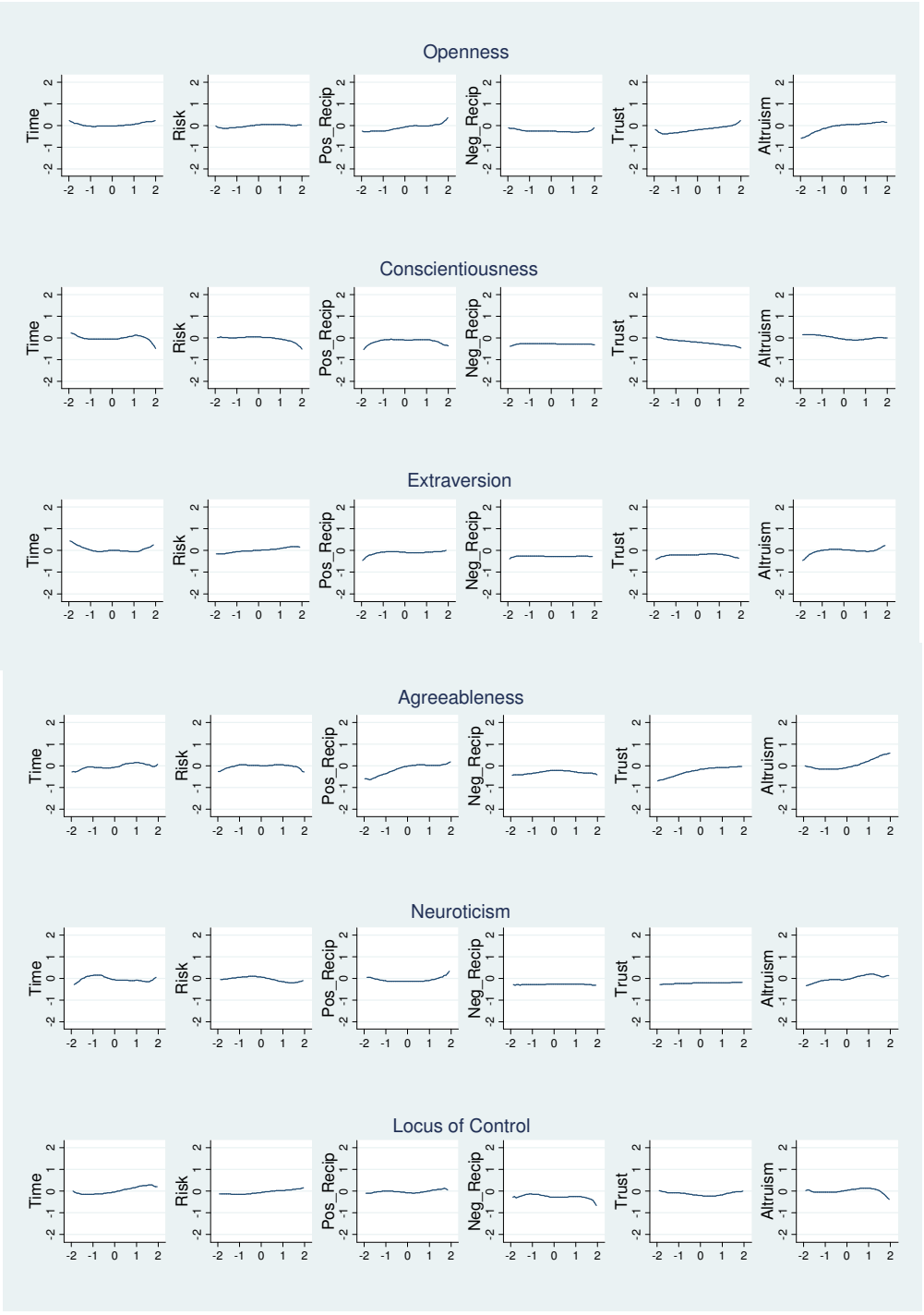


Figure 2.3: Kernel-weighted local linear polynomial regressions using SOEP data

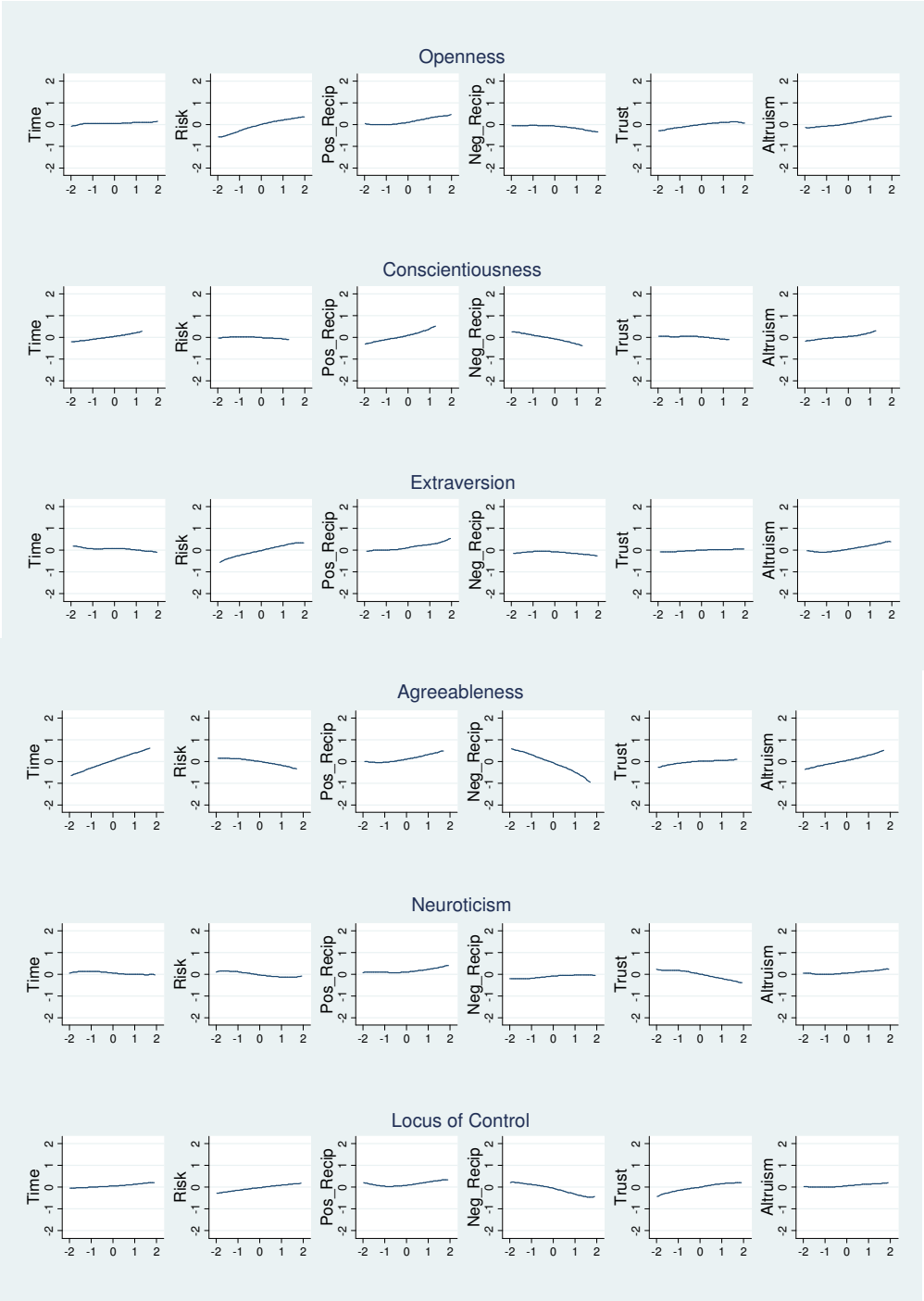
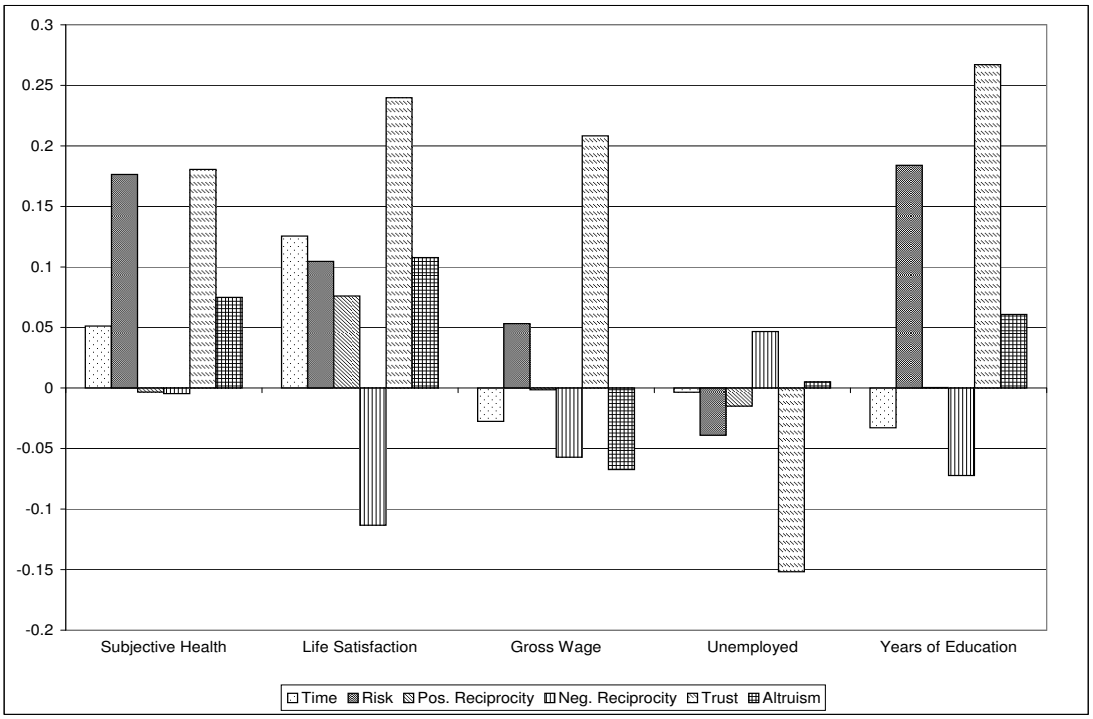
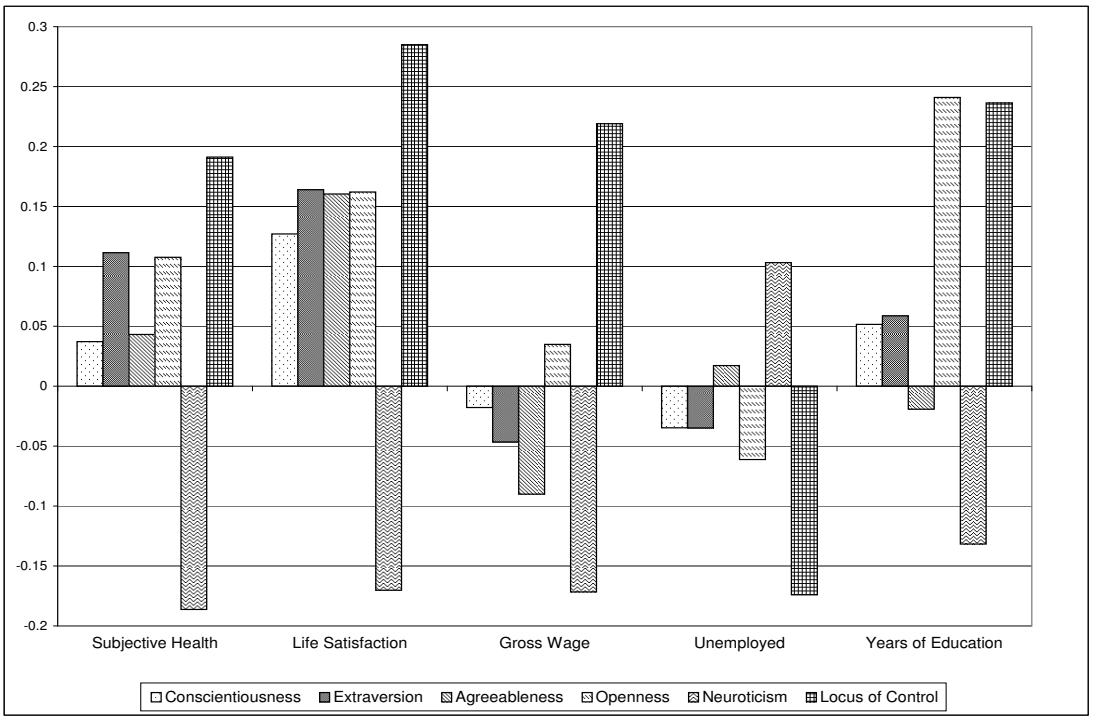


Figure 2.4: Correlation Coefficients Between Preference Measures and Life Outcomes Using SOEP Data



Pearson correlation coefficients between preference measures and life outcomes using SOEP data. Trust always shows the strongest association with life outcomes. More trust and a higher willingness to take risk are always related to better life outcomes, e.g. better health and greater life satisfaction, whereas negative reciprocity is associated with less life satisfaction and lower wages. The number of observations available varies for the different life outcomes: subjective health (14,218), life satisfaction (14,214), gross wage (7,199), unemployed (9,095), years of education (13,768). Gross wage measures the gross hourly wage.

Figure 2.5: Correlation Coefficients Between Personality Measures and Life Outcomes Using SOEP Data



Pearson correlation coefficients between personality measures and life outcomes using SOEP data. The locus of control and neuroticism show the strongest associations with life outcomes. A more internal locus of control is always related to better outcomes (e.g. better health or more life satisfaction), whereas a higher degree of neuroticism is associated with lower wages or a higher probability of being unemployed. The number of observations available varies for the different life outcomes: subjective health (14,218), life satisfaction (14,214), gross wage (7,199), unemployed (9,095), years of education (13,768). Gross wage measures the gross hourly wage.

Table 2.10: Linear representation of outcome regressions

	Subjective Health (OLS)					Subjective Health (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	0.0561	0.0383	0.0688	0.0975	0.1075	0.0220	0.0145	0.0268	0.0388	0.0429
F-Test/LR-Test	170.04	567.35	176.01	140.59	143.72	834.99	550.62	1016.47	1471.22	1627.11
AIC	37833	38094	37641	37201	37043	37139	37415	36960	36515	36361
BIC	37878	38109	37694	37292	37142	37207	37453	37035	36628	36482
	Life Satisfaction (OLS)					Life Satisfaction (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC					
adj. R^2/pseudo R^2	0.0899	0.0782	0.0917	0.1342	0.1588	0.0261	0.0219	0.0256	0.0390	0.0467
F-Test/LR-Test	281.88	1206.91	240.08	201.27	224.67	1406.38	1178.16	1376.73	2098.73	2513.61
AIC	55038	55216	55012	54335	53926	52448	52668	52480	51768	51355
BIC	55083	55231	55065	54426	54024	52561	52751	52601	51926	51521
	Gross Wage(OLS)									
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	-	-	-	-	-
adj. R^2/pseudo R^2	0.0361	0.0388	0.0456	0.0704	0.0919	-	-	-	-	-
F-Test/LR-Test	54.97	291.20	58.31	50.57	61.71	-	-	-	-	-
AIC	55088	55088	55042	54857	54690	-	-	-	-	-
BIC	55102	55102	55090	54940	54779	-	-	-	-	-
	Unemployed (OLS)					Unemployed (probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	0.0191	0.0331	0.0245	0.0375	0.0547	0.0322	0.0527	0.0412	0.0648	0.0926
F-Test/LR-Test	36.34	312.13	39.05	33.22	44.82	180.12	294.52	230.37	361.89	517.42
AIC	3067	2932	3017	2900	2738	5420	5298	5372	5250	5097
BIC	3110	2946	3067	2986	2830	5463	5312	5422	5336	5189
	Years of Education (OLS)					Years of Education (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC					
adj. R^2/pseudo R^2	0.0914	0.0525	0.1061	0.1545	0.1736	0.0209	0.0126	0.0241	0.0359	0.0415
F-Test/LR-Test	277.93	763.89	273.29	229.74	242.03	1355.80	817.10	1563.14	2329.14	2688.38
AIC	65506	66078	65282	64520	64206	63490	64021	63285	62529	62171
BIC	65551	66093	65335	64610	64304	63641	64141	63443	62724	62375

For the ordinary-least-squares (OLS) models we calculate R^2, whereas for the ordinal models we calculate pseudo R^2. The joint significance of all coefficients is tested using the F-test (OLS) and the LR-test (ordinal models). All F- and LR-tests are significant at the 1% level. With regard to the Akaike information criterion (AIC) and Bayesian information criterion (BIC), the smallest value for each outcome regression is underlined. Note that the full model (including the Big 5, locus of control and preferences) is always chosen by both information criteria. The number of observations available varies for the different life outcomes: subjective health (14,218), life satisfaction (14,214), gross wage (7,199), unemployed (9,095 obs.), and years of education (13,768). Gross wage measures the gross hourly wage.

Table 2.11: Outcome Regressions: Flexible Specification

	Subjective Health (OLS)					Subjective Health (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	.0632	.0388	.0714	.1054	.1165	.0251	.0146	.0282	.0435	.0483
F-Test/LR-Test	48.99	288.17	41.48	22.75	21.83	952.98	555.19	1068.56	1651.38	1834.03
AIC	37740	38088	37623	37142	36977	37051	37413	36949	36467	36310
BIC	37899	38110	37834	37732	37665	37232	37458	37184	37079	37021
	Life Satisfaction (OLS)					Life Satisfaction (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	.0948	.0783	.0948	.1397	.1659	.0278	.0219	.0273	.0422	.0505
F-Test/LR-Test	75.47	605.45	56.12	30.967	32.41	1493.78	1178.45	1470.26	2273.51	2715.76
AIC	54976	55214	54984	54311	53884	52391	52670	52428	51725	51309
BIC	55135	55237	55196	54901	54572	52617	52761	52708	52383	52065
	Gross Wage(OLS)									
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	-	-	-	-	-
adj. R^2/pseudo R^2	.0382	.0387	.0527	.0797	.1039	-	-	-	-	-
F-Test/LR-Test	15.30	145.74	15.84	9.092	10.27	-	-	-	-	-
AIC	55111	55090	55009	54851	54672	-	-	-	-	-
BIC	55256	55111	55202	55388	55298	-	-	-	-	-
	Unemployed (OLS)					Unemployed (probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	.0212	.0385	.0291	.0463	.0705	.0357	.0539	.0498	.0852	.1166
F-Test/LR-Test	10.87	183.13	11.11	6.73	8.66	199.54	301.02	278.38	475.96	651.83
AIC	3062	2882	2995	2882	2662	5431	5294	5366	5268	5118
BIC	3211	2903	3194	3437	3309	5580	5314	5565	5823	5766
	Years of Education (OLS)					Years of Education (o. probit)				
	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC	Big5	LoC	Pref	Big5-Pref	Big5-Pref-LoC
adj. R^2/pseudo R^2	.1043	.0525	.1200	.1771	.1982	.0243	.0126	.0281	.0433	.0497
F-Test/LR-Test	81.13	382.50	70.55	39.48	38.81	1575.60	817.25	1819.82	2808.59	3223.85
AIC	65324	66079	65087	64213	63869	63300	64023	63070	62181	61792
BIC	65482	66102	65297	64800	64554	63564	64151	63386	62874	62583

The outcome variables are regressed on the indicated personality and preference measures. The difference with regard to the linear specification is that the model includes squares of all variables as well as all cross-products. Cross-products are also calculated between concepts in case more than one concept is included, e.g., in the Big 5-preferences case, we also include the cross-term neuroticicsm*risk. For the ordinary-least-squares (OLS) models we calculate R^2, whereas for the ordinal models we calculate pseudo-R^2. The joint significance of all coefficients is tested using the F-test (OLS models) and the LR-test (ordinal models). All F- and LR-tests are significant at the 1% level. With regard to the Akaike information criterion (AIC) and Bayesian information criterion (BIC), the smallest value for each outcome regression is underlined. Note that the full model (including the Big 5, locus of control and preferences) is chosen by both information criteria in nearly all cases; only for gross wage and unemployment does the BIC indicate that the model with only LoC and LoC^2 included should be used. The number of observations available varies for the different life outcomes: subjective health (14,218), life satisfaction (14,214), gross wage (7,199), unemployed (9,095), and years of education (13,768). Gross wage measures the gross hourly wage.

Chapter 3

Nominal or Real? The Impact of Regional Price Levels on Satisfaction with Life

3.1 Introduction

Among the determinants of life satisfaction, income is of fundamental interest and importance to economists. Consequently, studies on the effect of income on life satisfaction are abundant. They range from cross-country studies on the relationship between gross national product and average reported life satisfaction to analyses of the effect of individual income on individual life satisfaction (for survey articles see, e.g., Oswald (1997), Frey and Stutzer (2002), Di Tella and MacCulloch (2006), Clark, Frijters, and Shields (2008), Dolan, Peasgood, and White (2008), and Stutzer and Frey (2010)).[1] Lacking adequate data on cross-sectional variation of prices, all research on individual life satisfaction conducted so far has basically used inflation adjusted nominal income as explanatory variable. According to microeconomic theory, however, individuals should derive satisfaction from consumption of goods that they can afford with their income rather than from nominal income. Hence, real income, i.e., nominal income adjusted for the local price level, is the appropriate concept to measure the effect of income on life satisfaction.

This paper therefore studies whether differences in local price levels affect individual satisfaction with life once we control for nominal income and local heterogeneity. To this end, we match two sources of data. The first is a novel and very comprehensive data set on local price levels in Germany, a price index covering each of Germany's 428 administrative

[1] Besides studying absolute income, the role of relative income (e.g., Clark and Oswald (1996), Luttmer (2005), Ferrer-i Carbonell (2005), Fliessbach et al. (2007)) and aspiration income (e.g., Stutzer (2004)) for individual life satisfaction has been explored.

districts. The price index reveals substantial price differences within Germany (up to 37%) and is, to our knowledge, unique at such a disaggregated level. Information used to construct the price index comprises more than 7 million data points. Having information on prices at a more aggregate administrative level (i.e., federal states) would not be sufficient for studying the effects of prices on life satisfaction. To illustrate, both the cheapest and the most expensive German district are geographically located in the same federal state. We match our price index data with data from the German Socio-Economic Panel (SOEP), a household panel survey, which is representative of the German population. It includes a question on individual life satisfaction, a wide range of control variables, and district identifiers. To identify the effect of the price level on life satisfaction, we estimate both pooled OLS and ordered probit models that include a comprehensive set of individual time-variant and time-invariant characteristics, among many others the 'Big Five' personality traits and economic preferences. Moreover, we control for district characteristics other than the price level that potentially influence life satisfaction such as local unemployment rate, local employment rate, average local household income, distance to the center of the closest large city, and guests-nights per capita, a proxy for attractiveness of the respective community.

Our main finding is a 'purchasing power effect'. For a given nominal income, a higher price level reduces satisfaction with life. The effect sizes are economically relevant. In our main specification, a 10% increase in the price level is predicted to decrease satisfaction with life by about 0.1 units, where satisfaction with life is measured on a scale from 0 to 10. This effect is roughly comparable to the decrease in life satisfaction caused by an increase in the distance travelled to work of about 100 kilometers. Being unemployed instead of full-time employed resembles the effect size of doubled prices. We perform various robustness checks and extend our analysis to two subdomains of well-being, in which the difference between nominal and real income is conceptually important: individual satisfaction with household income and individual satisfaction with standard of living. The results further confirm the purchasing power effect. For a given nominal income, higher local price levels reduce satisfaction with household income and satisfaction with standard of living at statistically and economically significant rates.

Our results show that not adjusting nationwide payments to regional price differences treats equals unequally in terms of individual life satisfaction. In this sense, our results provide an argument in favor of regional indexation of government transfer payments. They also question country-wide uniform public sector or minimum wages.

Beyond documenting the importance of local price levels for individual well-being, our study adds to uncovering how people perceive nominal and real quantities. From an economic policy perspective, perception of real versus nominal terms is, for example, important for determining optimal inflation rates to be targeted by central banks (Akerlof and Shiller, 2009). Economic theory usually assumes neutrality of money, i.e., that people think and act in terms of real quantities and are not guided by nominal quantities. In our case, neutrality of money implies that a price decrease should affect life satisfaction in the same way as an increase in nominal income that exactly offsets the price decrease in real income terms. In principle, deviations from neutrality of money could go in two directions. People could either overreact to changes in nominal income or to changes in prices.

An overreaction to nominal quantities is usually referred to as money illusion. Fisher (1928) was the first to suggest that people may exhibit money illusion.[2] In contrast, an overreaction to prices would imply that a decrease in prices increases life satisfaction more than a corresponding increase in disposable nominal income. An overreaction to prices is plausible if prices are more salient than nominal income. The importance of salience effects is documented in Chetty, Looney, and Kroft (2009), Blumkin, Ruffle, and Ganun (2010), and Finkelstein (2009) who provide evidence that consumers fail to sufficiently take into account less salient aspects in decision making.[3] Income is usually paid monthly and changes only infrequently. Furthermore, disposable income has many components that are not very salient such as taxes and government transfer payments. To the contrary, prices are experienced daily, at every instance of buying.

In contrast to most of the literature, our results on neutrality of money are based on yearly income data, i.e., large stakes for an individual. In favor of salience effects, our findings document that people tend to overreact to prices compared to nominal income. In our main specification, the estimated effect of a change of the price level on overall satisfaction with life is about 66% higher than the estimated effect of a corresponding change in nominal income. However, a formal test for neutrality of money, i.e., testing

[2]Money illusion was basically ignored in economic research until it was again studied by Shafir, Diamond, and Tversky (1997) who report evidence in favor of money illusion using questionnaire and experimental data. Weber et al. (2009) provide neuroeconomic evidence in favor of money illusion using functional magnetic resonance imaging. Using a laboratory experiment, Fehr and Tyran (2001) show that even a small extent of money illusion at the individual level may be sufficient to result in a large aggregate bias after a negative nominal shock.

[3]Chetty, Looney, and Kroft (2009) show that consumers underreact to less salient taxes, i.e., taxes that are not included in price tags. In a lab experiment, Blumkin, Ruffle, and Ganun (2010) find similar evidence. They show that less salient taxes distort the labor-leisure allocation. Finkelstein (2009) shows that drivers are less aware of tolls that are paid electronically and, as a consequence, driving is less elastic with respect to tolls that are paid electronically instead of manually.

whether the coefficients of the logarithm of nominal income and the logarithm of the price level differ significantly, does not reject neutrality of money.

The only other study on subjective well-being and price levels we are aware of is Boes, Lipp, and Winkelmann (2007). Their study differs from ours in many respects: the dependent variable, the available price level data, and methodology. They regress satisfaction with household income on price level data that was collected in 50 German cities, i.e., not in rural areas (Roos, 2006). Urban price levels are used to interpolate prices to the level of 13 out of 16 German federal states. Boes, Lipp, and Winkelmann (2007) test if people exhibit money illusion and do not find evidence for it. In contrast, we discuss and empirically document the effect of the local price level on overall satisfaction with life, a commonly used proxy for individual utility. Senik (2004) analyzes whether reference group income influences life satisfaction due to social comparisons or by providing information used to form expectations about one's own future income. She constructs 'real' income measures by using information on regional poverty lines of 38 Russian regions that are provided by the Russian longitudinal monitoring survey (RLMS) data set. Compared to our data, regional prices refer to much larger geographical units and are only available for comestible goods that account for about 9% of components of the price index we use. Luttmer (2005) also analyzes the influence of reference group income on individual well-being using average earnings in 'Public Use Microdata Areas' of the USA. To control for local characteristics that are both correlated with average local income and life satisfaction, he uses local housing prices and state fixed effects. Housing prices correspond to about one fifth of the information our price index contains. He finds that local housing prices are (insignificantly) negatively correlated with life satisfaction.

The remainder of the paper is organized as follows: section 3.2 describes both sources of data and section 3.3 explains our empirical strategy. Section 3.4 presents our results and several robustness checks. We discuss implications of our results and conclude in section 3.5.

3.2 Data

We use information on price levels of all 428 German districts ('Kreise'). Districts constitute administrative units comprising one or more cities and their surroundings. Districts are the smallest division of Germany for which it is feasible to collect detailed price data, because

in smaller units some of the products contained in the price index will not be available. The data on prices at district level have been collected by the German Administrative Office for Architecture and Comprehensive Regional Planning. Kawka et al. (2009) describe the data set, its collection and descriptive results on price levels in great detail.

The price index is based on the basket of commodities and the weights attached to each commodity that are used by the German Federal Statistical Office to calculate the German inflation rate. Table 3.1 lists the most important classes of goods that the basket of commodities contains. In terms of classes of goods, the price index covers 73.2% of this basket. In particular, more than 7 million data points on prices of 205 commodities have been collected at the district level. Commodities range from obvious candidates such as rental rates, electricity prices, or car prices to such detailed ones as dentist fees, prices for cinema tickets, costs for foreign language lessons, or entry fees for outdoor swimming pools.

Table 3.1: Main components of the basket of commodities

Commodity group	‰ of whole basket
Rent for dwellings (including rental value for owner-occupied dwelling)	203.30
Comestible goods	89.99
Goods and services for privately used vehicles	75.57
Electricity, gas, and other fuels	59.82
Clothing	39.42
Purchase of vehicles	37.50
Water supply and other dwelling related services	33.04
Food services	32.12
Leisure and cultural services	28.99
Telecommunication	27.12
Furniture, interior equipment, carpeting, and other floor coverings	26.50
Insurance services	24.88
Tobacco products	22.43
Personal hygiene	21.54
Leisure products, garden products, pets	21.53
Audiovisual, photographic, and information-processing devices and related equipment	19.01

Reproduced from the German Federal Statistical Office (2005) (see http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Statistiken/Preise/Verbraucherpreise/WarenkorbWaegungsschema/Waegungsschema,property=file.pdf). Displayed commodity groups account for about 750 ‰ of the whole basket of commodities.

With these data, a price index is constructed that provides an overall price level for each district. When constructing a price index, a weight needs to be attached to each individual commodity measuring its share of the whole basket of commodities. The price index is based on the weights that are used by the German Federal Statistical Office to construct the inflation rate. The weights are inferred from a household survey with 53,000 households that are asked about their income and consumption habits. With these weights, the price

index is constructed as an arithmetic mean. The weighting is the same for each individual and each district, i.e., it does not adjust for different consumption habits of rich and poor people, men and women, families and singles, young and old people or, more generally, for different individual or regional preferences for consumption. Such an approach certainly introduces some measurement error. Due to feasibility, it is, however, the standard approach in economics concerning price indices and also inflation rates. A clear advantage of this approach is that it allows for a direct comparison of different regional price levels and for a straightforward interpretation of the price index. Intuitively, we can ask what 'an average individual traveling through Germany' would need to pay for a given consumption bundle in each district. Since collecting such comprehensive data cannot be managed in a single year, the data were gathered in the years 2004 to 2009, with most of the data, roughly 85%, being collected from 2006 to 2008. The data are used to build a single time-invariant price level for each district.

The price index uses the district of the former German capital Bonn as baseline (100 points). The cheapest district is Tirschenreuth in the federal state of Bavaria with 83.37 points, while Munich with 114.40 points (also in Bavaria) is the most expensive district. Hence, the most expensive district is 37% more expensive than the cheapest, revealing a substantial price difference within Germany. Figure 3.1 in the Appendix shows a map of Germany indicating the price level of each district. Three observations are worth mentioning: price levels are generally lower in East than in West Germany and lower in Northern than in Southern Germany. Moreover, urban areas are more expensive than rural ones.

To obtain a measure of prices that accounts for both cross-sectional variation of prices at the district level and variation of prices over time, we multiply district specific price levels with inflation rates using 2006 as baseline year. The smallest geographical unit for which regional inflation rates are available in Germany is at the level of the 16 federal states.[4]

We match the price index data and data from the SOEP using district identifiers.[5] The SOEP is a representative panel study of German households that started in 1984. We use five waves from 2004 to 2008.[6] In each wave, about 22,000 individuals in 12,000 households

[4]From 2004 to 2008, 13 out of a total of 16 federal states report inflation rates for each year. For the federal state of Bremen, only the value for 2004 is missing. The federal states Hamburg and Schleswig-Holstein do not report own inflation rates in any year. For all missings, we interpolate the state level inflation rates with the German wide inflation rate of the corresponding year.

[5]Due to data privacy protection rules, working with the SOEP data at district level requires a special mode of online access to the SOEP data, SOEPremote.

[6]We cannot comprehensively match the price data to SOEP data from 2009 onwards. In 2009, some district boundaries were restructured. The new district boundaries are only reflected in the SOEP data, but not in the price index data.

are interviewed. Data cover a wide range of topics such as individual attitudes, preferences, and personality, job characteristics, employment status and income, family characteristics, health status, and living conditions. Schupp and Wagner (2002) and Wagner, Frick, and Schupp (2007) provide an in-depth description of the SOEP.

Since the first wave in 1984 participants are asked about their satisfaction with life on an eleven point Likert scale, which constitutes our main dependent variable. The life satisfaction question reads: "How satisfied are you with your life, all things considered?". Life satisfaction is often used as a measure for individual welfare or utility.[7] It is also gaining importance as an evaluation tool for economic policy. For example, in 2008, French President Nicholas Sarkozy asked a commission of economists to develop better measures for economic performance and social progress than, for example, GDP. In their report, the so called 'Sarkozy commission' notes that "... the time is ripe for our measurement system to shift emphasis from measuring economic production to measuring people's well-being." (p.12, Stiglitz et al. (2009)).

As alternative dependent variables, we use individual satisfaction with household income and individual satisfaction with standard of living. They are elicited in the following SOEP questions: "How satisfied are you with your household income?" and "Overall, how satisfied are you with your standard of living?". Satisfaction with household income is available from 2004 to 2008, while satisfaction with standard of living is only available from 2004 to 2006. Both questions use an eleven point Likert scale. Compared to general satisfaction with life, satisfaction with household income or standard of living is smaller in scope and less apt as a proxy for overall individual utility. However, they are even more closely linked to real (as opposed to nominal) income. Thus, the two alternative dependent variables will be useful to provide further evidence on how regional price levels affect well-being.

Besides a district's price level, nominal income is the main explanatory variable. We measure nominal income by household disposable nominal income, i.e., after tax household income including all kinds of government transfer income.[8] Instead of calculating equivalence income, we control for the logarithm of persons living in the household.

Additionally, we use a very comprehensive and well-established set of control variables at both individual and district level. The time-varying control variables at the individual level are age, age squared, dummies for marital status (married, separated, divorced, widowed;

[7]For a detailed discussion on the relationship between satisfaction with life and utility see, for example, Clark, Frijters, and Shields (2008) and Oswald (2008).

[8]We exclude about 60 observations with incomes above 500,000 Euro to avoid results being influenced by extreme outliers. Including them does not change our results.

single as omitted category), dummies for employment status (employed full-time, employed part-time, maternity leave, non-participant; unemployed as omitted category), years of education, a binary variable indicating whether an individual is disabled, a continuous variable indicating the official level of disability, the number of children in the household, and the distance travelled to the workplace in kilometers.

Furthermore, we use a comprehensive set of individual specific, time-invariant control variables. We include dummies for gender, German nationality, whether an individual describes himself as religious, and information on the political orientation of a person, which was elicited in SOEP wave 2005 on a scale from 0 (extreme left wing) to 10 (extreme right wing). Most importantly, we control for an individual's personality, economic preferences, and beliefs. Becker et al. (2012) show that concepts from psychology and economics should be combined when modeling individual differences. Using this approach, a large fraction of the variance in outcomes such as life satisfaction can be explained. Building on research in personality psychology, our control variables encompass the so called "Big Five", which are five superordinated character traits into which most of the subordinated character traits can be mapped (Costa and McCrae, 1992). The Big Five are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.[9] For each trait, we use standardized questionnaire measures that were elicited in the 2005 wave of the SOEP. A further important personality trait is the so called locus of control (Rotter, 1966). Locus of control measures the extent to which people think they are in control of events in their life. Our measure of locus of control uses standardized questionnaire measures from the 2005 wave of the SOEP. In economics, individual differences are commonly modeled by differences in preferences and beliefs. Important preferences are the preference for risk and time as well as social preferences (altruism, positive and negative reciprocity). An important belief is trust. Except for time preferences, all preferences and beliefs mentioned above were elicited at least once in the SOEP between 2004 to 2008. Whenever we have multiple measures for a given concept, we use the average to reduce measurement error. All measures are standardized.

To model district characteristics other than the price level that could both influence satisfaction with life and be correlated with the price level, we also include control variables at district level. The time-varying control variables mainly encompass macroeconomic variables that capture the current economic situation at district level: the average unem-

[9]For a detailed description of the Big Five see, e.g., Borghans et al. (2008).

ployment rate, the average employment rate in jobs subject to social security contributions, and the logarithm of the average household income. The time-invariant variables include the district size in square kilometers, the distance to the center of the closest large city (measured at individual level in 2004), and the number of guest-nights per capita in 2007 that proxy local attractiveness in terms of natural beauty or cultural facilities.

3.3 Empirical Strategy

We estimate a pooled OLS model with error terms clustered at district level for individual i's satisfaction with life in district j and a given year t, H_{ijt}:

$$H_{ijt} = \beta_0 + \beta_1 ln(N_{it}) + \beta_2 ln(p_{jt}) + \beta_3 ln(s_{it}) + \mathbf{x}'_{it}\boldsymbol{\gamma}_1 + \mathbf{c}'_i\boldsymbol{\gamma}_2 + \mathbf{d}'_{jt}\boldsymbol{\gamma}_3 + \mathbf{d}'_j\boldsymbol{\gamma}_4 + \gamma_{5t}h_t + \epsilon_{ijt}$$

N_{it} is nominal income. p_{jt} is the price index that captures cross-sectional variation of prices across districts and variation of prices over time. s_{it} is the number of persons living in the household, $\mathbf{x}_{it}$ is a vector including individual specific, time-varying control variables, $\mathbf{c}_i$ is a vector of time-invariant individual characteristics. $\mathbf{d}_{jt}$ and $\mathbf{d}_j$ are vectors of time-variant and time-invariant control variables at district level, h_t is a year dummy, β_0 is a constant term, and ϵ_{ijt} the error term.

Our primary research question is whether, for a given nominal income, differences in regional price levels affect individual satisfaction with life, i.e., whether β_2 is significantly different from zero. In addition, the specification at hand allows for a direct test of neutrality of money by testing whether β_1 is significantly different from β_2 in absolute value. According to economic theory, real income, i.e., nominal income adjusted for the regional price level, as opposed to pure nominal income should be the relevant source of satisfaction with life. Consequently, the two coefficients β_1 and β_2 should be of the same size in absolute terms. Assuming that β_1 is positive and β_2 is negative, a β_1 that is larger than $|\beta_2|$ would indicate that people exhibit nominal illusion. If $|\beta_2|$ were larger than β_1, the average individual would overreact to prices compared to nominal income, i.e., would suffer more from a price increase than it would suffer from a corresponding decrease in nominal income.

With the data at hand, it is not feasible to identify how regional price differences affect satisfaction with life by estimating an individual and / or district fixed effects model with $ln(p_{jt})$ as key explanatory variable. Since $ln(p_{jt}) = ln(p_j) + ln(inflation_t)$, the price index consists of a time-invariant part, $ln(p_j)$, that contains cross-sectional price variation and

a time-variant part, the inflation rate. In a fixed effects regression using individual or district fixed effects, the time-invariant part of the price level, $ln(p_j)$, does not contribute to identifying the coefficient of $ln(p_{jt})$. The coefficient of $ln(p_{jt})$ would only be identified via the inflation rate. Thus, it would not contain any information on how regional price levels influence satisfaction with life.

This argument neglects that individuals who move from one district to another provide an alternative source of variation in local prices that could potentially be used to identify the effect of the regional price level on individual satisfaction with life. However, movers constitute only a very small group of our sample. Furthermore, movers are likely to be a peculiar subset of the population, experiencing particularly strong shocks to life satisfaction caused by shocks to unobserved heterogeneity, e.g., frequent reasons for moving are changing the job or moving to live together with the partner. Thus, we are reluctant to generalize results that are based on movers only to the population as a whole and exclude movers from our main specification. In fact, estimating a fixed effects specification that uses only observations on movers estimates the impact of income on happiness to be negative which is in stark contrast to all existing literature (see, e.g., the survey of Dolan, Peasgood, and White (2008)).

Since we cannot include individual or district fixed effects, we use a very comprehensive set of time-invariant individual and district characteristics as regressors to explicitly model time-invariant sources of heterogeneity in overall satisfaction with life as advocated by, e.g., Ferrer-i-Carbonell and Frijters (2004).

3.4 Results

We first present and discuss the effect of cross-sectional variation of prices on overall satisfaction with life, before studying how cross-sectional variation of prices affects individual satisfaction with household income and individual satisfaction with standard of living, the two alternative dependent variables we use.

3.4.1 Results for overall satisfaction with life

Table 3.2 displays the main estimation results. In all specifications, the logarithm of nominal income has a statistically significant, positive influence on satisfaction with life ($p < 0.01$). Moreover, all specifications document economies of scale at the household level as the

coefficient of the logarithm of household size ($p < 0.01$) is smaller than the coefficient of the logarithm of nominal income in absolute terms.

Table 3.2: Life Satisfaction

	(1) Pooled OLS no district characteristics	(2) Pooled OLS main specification	(3) Ordered Probit main specification	(4) Pooled OLS including movers	(5) Pooled OLS including East dummy
$ln(N)$	0.520***	0.485***	0.338***	0.458***	0.474***
	(0.028)	(0.029)	(0.019)	(0.026)	(0.029)
$ln(P)$	0.567**	−0.806**	−0.571**	−0.626*	−0.693*
	(0.281)	(0.398)	(0.290)	(0.369)	(0.399)
ln(persons in household)	−0.446***	−0.409***	−0.290***	−0.404***	−0.396***
	(0.048)	(0.048)	(0.031)	(0.046)	(0.048)
individual controls	yes	yes	yes	yes	yes
district controls	no	yes	yes	yes	yes
year dummies	yes	yes	yes	yes	yes
p-value of test ($\beta_1 = -\beta_2$)	0.000	0.417	0.418	0.647	0.581
R^2	0.2254	0.2298	–	0.2272	0.2313
# of observations	55,366	55,366	55,366	59,212	55,366

Dependent variable is individual life satisfaction. *, **, and *** indicate significance at the 10%, 5%, and 1% level. Standard errors, clustered at district level, are shown in parentheses. Time-varying individual controls are age, age squared, dummies for marital status (married, separated, divorced, widowed; single as omitted category), dummies for employment status (employed full-time, employed part-time, maternity leave, non-participant; unemployed as omitted category), years of education, a dummy for being disabled, a continuous variable indicating the official level of disability, the number of children in the household, and the distance travelled to the workplace in kilometers. Individual specific, time-invariant control variables are dummies for gender, German nationality, religiosity, a variable for political orientation, standardized measures of the Big Five (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism), locus of control, preference for risk, altruism, positive and negative reciprocity, trust. Control variables at district level include the average unemployment rate, the average employment rate, and the logarithm of the average household income. The time-invariant variables at district level are the district size in square kilometers, the distance to the center of the closest large city, and the number of guest-nights per capita. Finally, year dummies are included.

Column (1) shows the results of a regression including individual specific controls (time-varying and time-invariant), but no district characteristics other than the price level. In this specification, higher prices are associated with an increase in satisfaction with life ($p < 0.05$). At first sight, this result seems surprising. However, it is likely due to omitted variable bias: more 'attractive' districts have higher price levels. To control for the 'attractiveness' of a

given district, we proceed by adding district level control variables in column (2). First, the local unemployment rate, the employment rate, and the average district household income describe the current economic situation at district level. Second, the district's size and the distance to the center of the closest large city are proxies for how rural or urban a given district is and thus also for its infrastructure. Finally, the number of guest-nights per capita proxies local attractiveness in terms of natural beauty or cultural facilities.

Column (2) presents the results of our main specification.[10] There are two key insights. First, for a given nominal income, higher local prices decrease individual satisfaction with life ($p < 0.05$). A 10% increase in the price level is predicted to decrease satisfaction with life by 0.08 units, where satisfaction with life is measured at a 11 point Likert scale. To get a better intuition for the magnitude of the price level effect on life satisfaction, we compare the coefficient of the price level with coefficient of other explanatory variables. For example, an increase of the price level by around 8% decreases life satisfaction as much as an increase in the distance travelled to work of around 100 kilometers. Being unemployed instead of full-time employed resembles the effect size of a doubling of prices.

Second, our results do not reject neutrality of money. Testing whether the coefficient of nominal income, β_1, is significantly different from the coefficient of the price level, β_2, in absolute terms yields $p = 0.42$. However, in absolute terms, the coefficient of the logarithm of the price level is 66% larger than the coefficient of the logarithm of nominal income, indicating that people have the tendency to react stronger to changes in prices than to corresponding changes in nominal income. For example, while a 10% increase in the price level is predicted to decrease satisfaction with life by 0.08 units, a 10% decrease in nominal income is predicted to reduce satisfaction with life by only 0.05 units. Salience effects (Chetty, Looney, and Kroft (2009), Blumkin, Ruffle, and Ganun (2010), Finkelstein (2009)) offer a possible explanation for a larger impact of prices than of nominal income on satisfaction if prices are more salient than disposable income. This seems likely. Many components of disposable income might be less salient, e.g., taxes and government transfer payments, and, for most people, income changes are relatively rare events. In contrast, prices and price changes are experienced frequently, prices at every instance of buying.

We check the robustness of our main specification in various ways. First, in column (3), we take into account the ordinal nature of our dependent variable by estimating an

[10] Table 3.5 in the Appendix displays all estimated coefficients of the main specification. It documents that, in general, the estimated coefficients of our control variables are well in line with the existing literature. The time-invariant personality traits and economic preferences contribute significantly to explaining life satisfaction.

ordered probit model. Using the ordinal model, the coefficient of the price level remains significantly negative ($p < 0.05$). As a second robustness check, in column (4), we add observations from all movers to the sample. As noted before, movers constitute a peculiar subgroup that, when analyzed separately in a fixed-effects framework, show a negative relationship between nominal income and satisfaction with life. However, including movers in our sample, results stay qualitatively the same. For a given nominal income, a higher price level is still predicted to decrease satisfaction with life ($p < 0.1$). Again, we do not reject neutrality of money. Finally, we include an additional dummy variable indicating whether a district lies in East or West Germany in column (5). Frijters, Haisken-DeNew, and Shields (2004) document that life satisfaction in East Germany is generally lower than in West Germany. Our district level explanatory variables should already capture a large share of differences between East and West Germany that still exist and affect satisfaction with life, such as differences in economic conditions. Including an East / West dummy allows controlling for potential further differences between East and West Germany. Once more, our results are stable and document that, for a given nominal income, higher prices reduce satisfaction with life ($p < 0.1$). Again, we do not reject neutrality of money ($p = 0.58$).

3.4.2 Results for satisfaction with household income and satisfaction with standard of living

In order to obtain further evidence on how the local price level affects individual well-being, we investigate the influence of the local price level on satisfaction with household income and satisfaction with standard of living. Real income seems to be a driving force for both subdomaines of individual well-being. In contrast, it is a well-established result that income has a significant impact on overall satisfaction with life, but, compared to other explanatory variables such as unemployment or health, the role of income is relatively small. Consequently, we hypothesize that the coefficients of nominal income and the local price level are larger in those two domains than for overall satisfaction with life.

Table 3.3: Satisfaction with Household Income

	(1) Pooled OLS no district characteristics	(2) Pooled OLS main specification	(3) Ordered Probit main specification	(4) Pooled OLS including movers	(5) Pooled OLS including East dummy
$ln(N)$	1.622***	1.586***	0.906***	1.550***	1.569***
	(0.042)	(0.041)	(0.024)	(0.039)	(0.041)
$ln(P)$	−0.134	−1.394**	−0.858**	−1.213**	−1.220**
	(0.360)	(0.599)	(0.338)	(0.569)	(0.602)
ln(persons in household)	−1.239***	−1.209***	−0.698***	−1.218***	−1.190***
	(0.069)	(0.069)	(0.037)	(0.066)	(0.069)
individual controls	yes	yes	yes	yes	yes
district controls	no	yes	yes	yes	yes
year dummies	yes	yes	yes	yes	yes
p-value of test ($\beta_1 = -\beta_2$)	0.000	0.749	0.888	0.555	0.563
R^2	0.3068	0.3095	–	0.3077	0.3116
# of observations	54,921	54,921	54,921	58,721	54,921

Dependent variable is satisfaction with household income. *, **, and *** indicate significance at the 10%, 5%, and 1% level. Standard errors, clustered at district level, are shown in parentheses. The control variables are exactly the same as in Table 2.

Table 3.4: Satisfaction with Standard of Living

	(1) Pooled OLS no district characteristics	(2) Pooled OLS main specification	(3) Ordered Probit main specification	(4) Pooled OLS including movers	(5) Pooled OLS including East dummy
$ln(N)$	0.908***	0.880***	0.606***	0.867***	0.869***
	(0.036)	(0.036)	(0.024)	(0.034)	(0.036)
$ln(P)$	−0.363	−1.158***	−1.134***	−1.295**	−1.419***
	(0.329)	(0.535)	(0.357)	(0.511)	(0.541)
ln(persons in household)	−0.799***	−0.777***	−0.542***	−0.791***	−0.763***
	(0.062)	(0.069)	(0.039)	(0.059)	(0.060)
individual controls	yes	yes	yes	yes	yes
district controls	no	yes	yes	yes	yes
year dummies	yes	yes	yes	yes	yes
p-value of test ($\beta_1 = -\beta_2$)	0.093	0.234	0.139	0.404	0.311
R^2	0.2601	0.2633	–	0.2609	0.2645
# of observations	32,926	32,926	32,926	35,186	32,926

Dependent variable is satisfaction with standard of living. *, **, and *** indicate significance at the 10%, 5%, and 1% level. Standard errors, clustered at district level, are shown in parentheses. The control variables are exactly the same as in Table 2.

Tables 3.3 and 3.4 present the results for satisfaction with household income and satis-

faction with standard of living, respectively. Except for the dependent variable, they rely on exactly the same specifications as table 3.2. In all specifications, it is indeed the case that the coefficients of nominal income and the local price level are, in absolute terms, larger for satisfaction with household income and satisfaction with standard of living than for overall life satisfaction. Furthermore, our main results derived for overall satisfaction with life are replicated for the two new dependent variables: there is a significant positive relationship between nominal income and satisfaction, but a negative effect of the local price level on satisfaction with household income and standard of living once district level control variables are included. Furthermore, neutrality of money is not rejected in any specification.

A further interesting finding is that, when evaluating their satisfaction with standard of living, we again find that people react stronger to changes in prices than to changes in nominal income. This effect is, however, not significant at conventional levels ($p = 0.23$). In contrast, for satisfaction with household income, the coefficient of the price level is slightly smaller than the coefficient of nominal income. This difference is not significant either ($p = 0.75$). One plausible explanation could again be salience effects: if people are directly asked about their satisfaction with household income, nominal income might be particularly salient.

3.5 Discussion

We have used a novel and very comprehensive data set on local price levels in Germany to study whether cross-sectional variation in price levels affects satisfaction with life once nominal income is controlled for. Our results show that information on price levels matters when analyzing satisfaction with life. We find that people exhibit significantly lower life satisfaction when living in a more expensive region. The effect of an increase in the price level on life satisfaction is also economically significant: A 10% increase in the price level decreases satisfaction with life by 0.08 units on a scale ranging from 0 to 10. Moreover, although a marginal price decrease is estimated to have a 66% stronger impact on life satisfaction than a corresponding increase in nominal income, this discrepancy is not large enough to reject neutrality of money. The result that, for a given nominal income, a higher price level reduces individual well-being also extends to subdomains of well-being, in particular satisfaction with household income and satisfaction with standard of living.

Our results are of relevance for advising policy, in particular if policy aims at treating equals equally. In that sense, our findings call for a regional indexation of government transfer payments, such as the US Supplemental Security Income (SSI), unemployment benefits, or social welfare benefits. Our results also put country-wide uniform public sector or minimum wages into question. In all examples, not adjusting nationwide payments to regional price differences risks treating equals unequally in terms of individual satisfaction with life.[11]

We believe that the price index data employed in this paper offer lots of scope for future research. Relevant questions that require detailed information on local price levels comprise, e.g., the effect of the price level on whether wages are perceived as fair, how job search activity or investments in human capital depend on regional price differences, and whether local price levels affect migration within a country.

[11] Of course, the validity of these arguments rests on a ceteris paribus assumption, i.e., groups who get compensated for differences in the price level are assumed to be small enough for a change in their nominal income not to affect the local price level.

A3 Appendix to Chapter 3

Figure 3.1: Regional Price Index

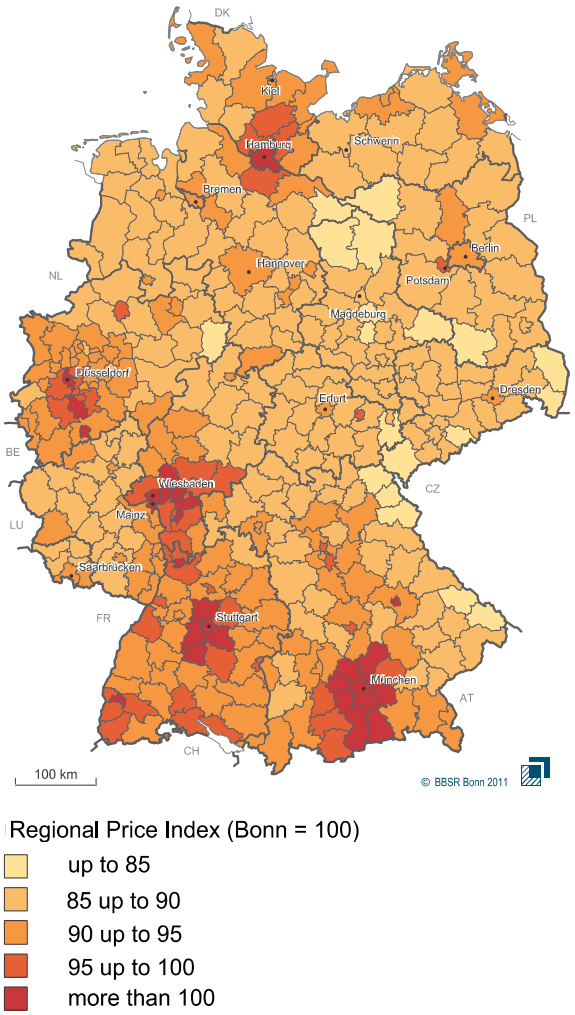


Figure from Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR), Raumordnungsbericht 2011, Bonn 2012. The colors display ranges of the originally scaled price index. Borders of the districts are marked by grey lines while borders of federal states are marked by dark grey lines.

Table 3.5: Detailed Results of Main Specifications (OLS)

	Life satisfaction	Satisfaction with household income	Satisfaction with standard of living
$ln(N)$	0.485***	1.586***	0.880***
	(0.029)	(0.041)	(0.036)
$ln(P)$	−0.806**	−1.394**	−1.518***
	(0.398)	(0.599)	(0.535)
ln(persons in household)	−0.409***	−1.209***	−0.777***
	(0.048)	(0.069)	(0.061)
dummy disabled	0.076	−0.066	0.118
	(0.097)	(0.096)	(0.098)
degree of disability	−0.012***	−0.004***	−0.007***
	(0.002)	(0.002)	(0.002)
male	0.008	−0.093***	−0.062**
	(0.021)	(0.028)	(0.025)
age	−0.039***	−0.069***	−0.060***
	(0.006)	(0.007)	(0.007)
age^2	0.038***	0.073***	0.063***
	(0.006)	(0.007)	(0.006)
years of education	−0.008*	−0.003	0.008
	(0.005)	(0.006)	(0.005)
number of children	0.087***	0.183***	0.123***
	(0.019)	(0.028)	(0.024)
dummy foreigner	0.049	−0.135	−0.250**
	(0.065)	(0.106)	(0.123)
married	0.152***	0.278***	0.216***
	(0.047)	(0.061)	(0.052)
separated	−0.502***	−0.380***	−0.444***
	(0.115)	(0.143)	(0.142)
divorced	−0.273***	−0.391***	−0.416***
	(0.062)	(0.085)	(0.079)
widowed	−0.127**	0.032	−0.111
	(0.064)	(0.083)	(0.071)
full-time employed	0.770***	1.111***	0.764***
	(0.054)	(0.065)	(0.067)
part-time employed	0.787***	1.046***	0.779***
	(0.055)	(0.066)	(0.066)
parental leave	1.156***	0.960***	0.941***
	(0.078)	(0.092)	(0.088)
out of the labor force	0.941***	1.335***	0.935***
	(0.055)	(0.069)	(0.070)
distance to work (in km)	−0.001**	0.000	0.000
	(0.000)	(0.000)	(0.000)
openness	0.087***	0.049**	0.070***
	(0.016)	(0.019)	(0.016)
conscientiousness	0.076***	0.080***	0.076***
	(0.016)	(0.019)	(0.019)
extraversion	0.042***	0.010	0.027*

Continued on next page

Table 3.5 – *Continued from previous page*

	(1)	(2)	(3)
	(0.015)	(0.017)	(0.016)
agreeableness	0.108***	0.069***	0.088***
	(0.015)	(0.019)	(0.015)
neuroticism	−0.083***	−0.086***	−0.054***
	(0.014)	(0.016)	(0.014)
locus of control	0.324***	0.323***	0.353***
	(0.015)	(0.019)	(0.019)
risk preference	0.097***	−0.047**	0.008
	(0.017)	(0.019)	(0.019)
positive reciprocity	0.029*	0.054***	0.055***
	(0.015)	(0.018)	(0.017)
negative reciprocity	0.032**	0.019	0.017
	(0.015)	(0.021)	(0.020)
trust	0.202***	0.227***	0.165***
	(0.015)	(0.019)	(0.017)
altruism	0.076***	0.052***	0.089***
	(0.015)	(0.019)	(0.019)
political orientation (large values indicate right wing)	0.023***	−0.001	0.011
	(0.007)	(0.009)	(0.009)
dummy religious	0.078**	0.173***	0.094**
	(0.032)	(0.040)	(0.036)
area of district (in 1000 km^2)	−0.024	−0.033	−0.021
	(0.031)	(0.041)	(0.036)
employment rate of district	−0.020***	−0.014**	−0.009*
	(0.005)	(0.007)	(0.005)
unemploymentrate of district	−0.031***	−0.029***	−0.020***
	(0.005)	(0.008)	(0.006)
log of average household income of district	0.262	0.275	0.381
	(0.234)	(0.352)	(0.333)
distance to next city center of district	−0.005	0.014	0.016
	(0.012)	(0.015)	(0.014)
yearly guest-nights per capita of district	0.000	−0.005	−0.006*
	(0.003)	(0.004)	(0.003)
year dummies	yes	yes	yes
p-value of test ($\beta_1 = -\beta_2$)	0.417	0.749	0.234
R^2	0.2298	0.3095	0.2633
# of observations	55,366	54,321	32,926

Dependent variable is individual life satisfaction. *, **, and *** indicate significance at the 10%, 5%, and 1% level. Standard errors, clustered at district level, are shown in parentheses. Section 2 contains a description the explanatory variables.

Chapter 4

How Does Socio-Economic Status Shape a Child's Personality?

4.1 Introduction

Both economic theory and empirical evidence have established a robust link between many important outcomes in life and economic preferences as well as IQ. Time preferences are a major determinant of making investments that will only pay off in the future. More patient individuals achieve higher levels of educational attainment, resulting in substantially higher earnings (Golsteyn, Grönqvist, and Lindahl, 2013; Shoda, Mischel, and Peake, 1990). Furthermore, they are more likely to exercise, to be a non-smoker, and less likely to be obese (Chabris et al., 2008; Sutter et al., 2010; Golsteyn, Grönqvist, and Lindahl, 2013). Risk preferences are another important predictor of both economic and health outcomes. A higher willingness to take risks is positively correlated with being self-employed, investing in stocks, with smoking, and taking exercises (Dohmen et al., 2011b). Social preferences that reflect an individual's degree of altruism are, e.g., related to overall satisfaction with life (Becker et al., 2012). Finally, higher levels of IQ are associated with higher levels of education (Heckman and Vytlacil, 2001), income (Hanushek and Woessmann, 2008), and job performance (Schmidt and Hunter, 2004). Outcomes like educational attainment, occupational choice, health related behavior, or satisfaction with life shape an individual's life. At the aggregate level, these outcomes are also important for societies as a whole, since they, for example, affect productivity or costs of the health care system. A better understanding of how these outcomes come about requires understanding how economic preferences and IQ form.

Conceptually, economic preferences and IQ are important facets of a person's personality (Almlund et al., 2011b; Borghans et al., 2008). Personality emerges in childhood and adolescence and is generally thought of as relatively stable afterwards.[1] This paper contributes to the understanding of the origins of economic preferences and IQ by documenting a systematic and strong relationship between parental socio-economic status (SES) and a child's economic preferences and IQ. We measure parental SES by the net household equivalence income and the mother's and father's average years of education. Previous work (Cunha and Heckman, 2007; Heckman, 2008) stresses the importance of parental investments in their children for shaping a child's personality. In the model suggested by Cunha and Heckman (2007) and Heckman (2008), parental SES is a prime candidate for shaping a child's personality since parental SES largely defines the monetary and cognitive resources available to parents for educating their child. These resources are a crucial prerequisite for parental investments.

We proceed in three steps. First, for each facet of personality under study, we document whether there is a significant relationship between parental SES and the respective facet of a child's personality. For that purpose, we regress the different personality traits on parental education and household income only. Our results document a strong relationship between parental socio-economic status and a child's risk preferences, time preferences, and IQ. An obvious question is whether differences in a child's personality are only driven by differences in household income and parental education or whether they are also shaped by other dimensions of a child's environment that differ by parental socio-economic status. We therefore, in a second step, investigate differences in family structure, initial conditions at birth, the personality profile of the child's mother, and different aspects of parental behavior such as parenting style, the time parents spend with their children, and the quality of time spent together. In our analysis, we show that these environmental dimensions systematically differ according to parental SES. In a final step, we investigate to which extent the relationship between parental SES and a child's personality is due to the documented differences in a child's environment. The additional environmental variables add explanatory power and reduce the coefficients of household income and parental education by about 20 to 40%. However, household income and parental education remain significant and important predictors of a child's personality.

[1] See Almlund et al. (2011b), p.117ff or Borghans et al. (2008) for a general discussion on the stability of personality. During their development process, children typically become more patient (Bettinger and Slonim, 2007), less risk seeking (Paulsen et al., 2011), and more altruistic (Fehr, Bernhard, and Rockenbach, 2008).

The dataset that we use in this study comprises measures of time preferences, risk preferences, and altruism that are based on incentivized experiments with 732 children. All 732 children also participated in fluid and crystallized IQ tests. Moreover, for all children we have detailed questionnaire measures (completed by their mothers) on their family's SES, the child's initial conditions at birth (such as weight at birth, the week of gestation at birth, and the number of older siblings at birth), and family structure (whether the child lives with a single parent, the age of the mother at birth, and the current number of siblings at home). Using information on the mother's IQ and the economic preferences of the mother, we partially control for genetic transmission of ability and a direct intergenerational transmission of economic preferences. Moreover, we have information on the parenting style, on how many hours per week the mother is the main carer of her child, and what parents actually do with their child when they spend time together.

Studying the relationship between parental SES and a child's personality is important in at least three respects. First, it enhances our understanding of the sources of heterogeneity in personality. Second, it may be helpful in explaining social immobility. It is well documented that individuals with different personality profiles are likely to end up with, e.g., different economic, social, and health outcomes in life (Chabris et al., 2008; Sutter et al., 2010; Golsteyn, Grönqvist, and Lindahl, 2013; Dohmen et al., 2011b; Heckman and Vytlacil, 2001; Hanushek and Woessmann, 2008). An individual's SES is one of those outcomes. Social immobility occurs if children from parents with high SES are more likely to develop a personality profile that is associated with outcomes that result in high SES. This is exactly the link we find.

Finally, the fact that children's preferences vary systematically with the SES of their parents is also important when explaining (later) life outcomes with differences in preferences in childhood or adolescence. An example is the seminal work of Mischel, Shoda, and Rodriguez (1989) who show that the amount of "self-imposed delay of gratification" at the age of four is significantly related to academic and social competence and verbal fluency ten years later. In the light of our findings, it is important to include information on parental SES when analyzing the influence of preferences on outcomes to avoid potential omitted variable bias. A similar line of reasoning applies to Sutter et al. (2010) who document that more impatient adolescents are more likely to spend money on alcohol and cigarettes, have a higher BMI, and are less likely to save money.

To the best of our knowledge, this is the first study to focus on analyzing the effect

of parental SES on children's time preferences or risk preferences measured in incentivized experiments. In the domain of social preferences, Bauer, Chytilová, and Pertold-Gebicka (2011) find that children from families with higher SES are more altruistic, i.e., more likely to donate to another child in a binary dictator game experiment. In contrast to their study, we do not find an effect of parental SES on children's altruism. While research on the relation of parental SES and children's economic preferences is still in its infancy, the relationship between parental SES and children's IQ is well established and, according to Bradley and Corwyn (2002), especially clear cut: Children from high SES families score significantly higher on IQ tests.

The remainder of the paper is organized as follows: First, we describe the composition of our sample and our measures of economic preferences and IQ. Section 4.3 outlines the estimation strategy. Section 4.4 contains the results and embeds our findings in the existing literature. In the final section, we discuss the main implications of our findings.

4.2 The data

4.2.1 The sample

Our sample consists of 732 children and their mothers who were recruited using official registry data.[2] Interviews took place in Bonn and Cologne (Germany) and were conducted by trained university students (mostly graduates) of psychology or education science. Children participated in a sequence of 7 experiments, 2 short intelligence tests for fluid and crystallized IQ, and answered a brief questionnaire. In total, the interviews lasted about one hour. Children were paid and incentivized using toys and a small amount of money with a total average value of about 9 Euro. We introduced an experimental currency called "stars". At the end of all experiments, children could exchange the amount of paper stars won into toys. Presents were arranged in four categories which visibly increased in value and subjective attractiveness to children (see figure 4.7 in the Appendix). During the experiments, children knew that more stars would result in the option to choose a present from a higher category. We ensured that each additional star that would not result in a higher category still had an extra value to the children by converting these additional stars

[2]We received a random selection of approximately 90% of the addresses of families living in Bonn and Cologne (Germany) who had children of age seven to nine. These families were informed about the study via postal mail. 12.5% (N=1874) of the contacted families were interested in participating. Mainly due to capacity contraints for the current study, we could not invite all families.

into Lego bricks.

During the time children participated in the experiments, their mothers[3] filled out a very comprehensive questionnaire with the following categories of topics:

- Basic information about the child, e.g., name, age, etc.
- Socio-economic background of the family
- Health status of the child and information about the early childhood environment
- Details about child care and parenting style
- Assessments of personality and attitudes of the child
- Personality, preferences, and attitudes of the mother

Table 4.1 shows some basic characteristics of the participating children and their mothers. The monthly net household equivalence income (hereafter referred to as income) is calculated by dividing the total monthly nominal household income (including any transfers, but after taxes) by a factor that takes into account household size. In particular, the factor takes on the value 1 for a single-person household. For each additional person aged 14 years or older 0.5 is added, for each person younger than 14 years 0.3 is added. This approach of calculating the equivalence income is suggested by the OECD. The rationale for calculating the equivalence income is to account for the number of persons living in a household, while also taking into account economies of scale of a household. On average, income equals 1265 Euro and the median is 1111 Euro. The average income in our sample corresponds to a household that is roughly positioned at the 40% quantile of the German income distribution.[4] The level of education is measured in years of education averaged over mothers and fathers. This variable is created by adding up numbers of years of schooling and occupational training (including university). On average, parents have 12.8 years of education which corresponds to having completed a standard apprenticeship after obtaining the secondary school level certificate. Roughly 52% of the children are boys. On average, the children have 1.18 siblings. The number of siblings ranges from 0 to 7 siblings with a

[3]Actually, 96% of the children were accompanied by their biological mother, 2% by their biological father, 3 children by a step or foster parent, one child by the new partner of a biological parent. We do not have unambiguous information on the accompanying person for about 1% of the children. Throughout the paper, we will use the term "mother" for the adult accompanying the child.

[4]This statement is based on own calculations using the self-reported income data from the SOEP in year 2009 and the cross-sectional weights provided in the SOEP data to make the data representative for the German population.

median of 1.

Table 4.1: Basic Characteristics of the Sample

Variable	Mean	Std. Dev.	Median	Min	Max
monthly net equivalence income (in 1000 Euro)	1.265	0.668	1.111	0.186	7.143
avg. parental years of education	12.81	2.79	12.25	7	18
male	0.52	–	–	–	–
age (in months)	93.39	6.29	92	84	113
number of siblings	1.18	1.05	1	0	7

4.2.2 Description of experiments and IQ tests

In the following, we explain the three incentivized experiments used to measure time preferences, risk preferences, and altruism of the children. We then present the IQ tests.

Time preferences: Piggybank experiment

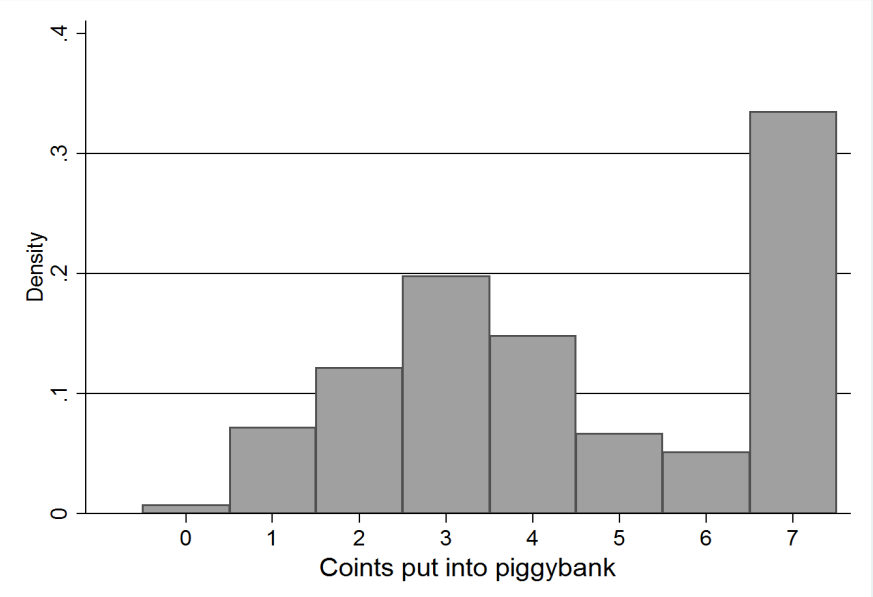
Children were endowed with seven 20 cent coins. They could choose how many coins to put in a piggybank and how many to take immediately. The amount put in the piggybank was doubled and sent to the children via postal mail one week after the interview. We took great care in ensuring that the amount of coins put into the piggybank was not influenced by children's trust in the saved money being indeed delivered to them: we explicitly addressed the letter to the children themselves, wrote addresses on the envelope, and put the saved amount of money in the envelope while the children were watching.[5] Understanding of the game was checked via a control question. The game only started after the children had fully understood its rules.

The amount of coins put into the piggybank is our observational measure for the child's discount rate. In particular, a higher number of coins put into the piggybank implies lower discounting of the future.[6] To see this, assume that the utility of consumption today and

[5]Moreover, detached from this experiment, we asked the children three questions concerning their general trust in other people. Using the answers to these questions, we build a standardized trust score. Neither Pearson nor Spearman correlations of the trust score with the number of saved coins are significantly different from zero at any conventional significance level. We infer that children's level of trust in other people does not influence their decision in the Piggybank experiment.

[6]This holds under the assumption that children do not take into account their current financial situation when evaluating the saving decision. Table 4.2 in section 4.4 provides some affirmative evidence for this assumption as net household equivalence income of the family is not significantly related to the decisions of the children.

Figure 4.1: Distribution of Saving Decisions (Histogram)



consumption in one week follows a twice differentiable utility function u with $u'(x) > 0$, $u''(x) < 0$, $u'(0) \to \infty$ and let future utility be depreciated by δ (discount factor), with $0 \leq \delta \leq 1$. Let a denote the number of coins put in the piggybank and let b denote the total number of coins available. Then, the child faces the following maximization problem:

$$\underset{a}{\text{maximize}} \quad u(b-a) + \delta u(a)$$

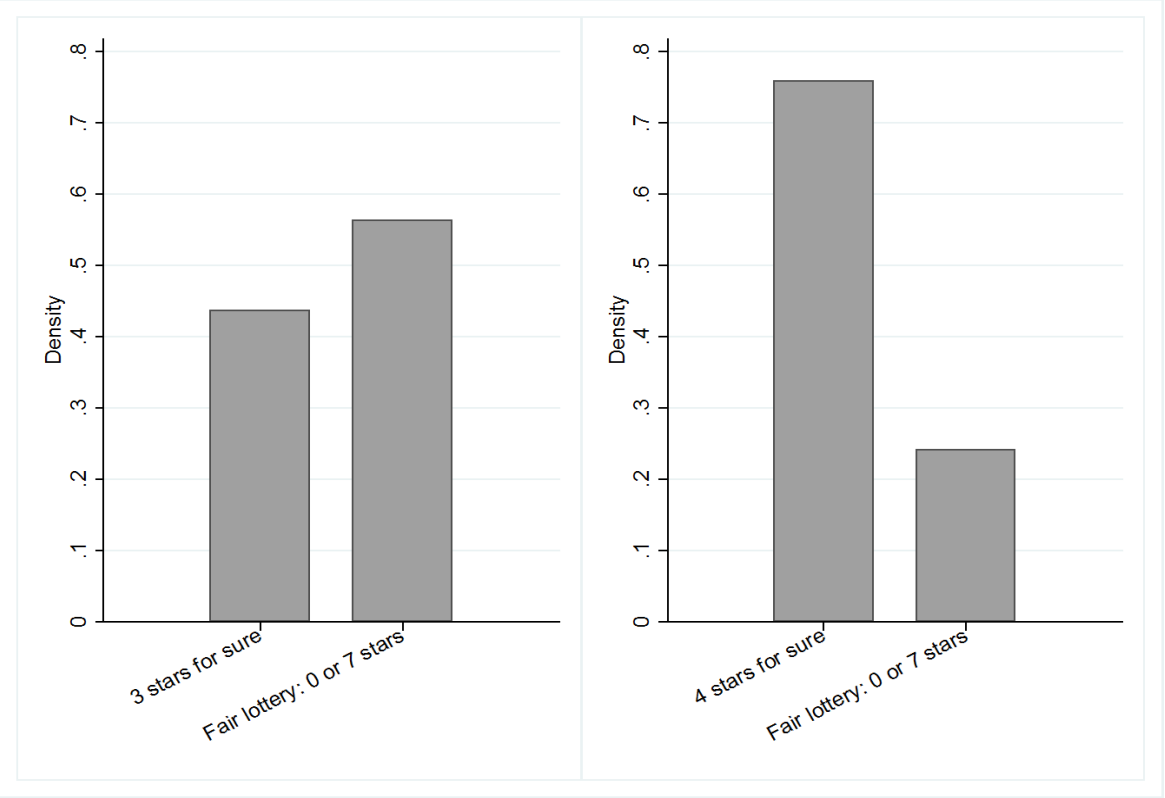
In the optimum, it holds

$$u'(b-a) = \delta u'(a).$$

This implies that larger values of δ result in larger values of a, i.e., the less the future is discounted, the more coins are put into the piggybank.

Figure 4.1 shows the distribution of saving decisions. About 35% of the children choose to "save" 7 coins. Overall, there is substantial variation in the saving choices. The average number of coins put in the piggybank is 4.49 with a standard deviation of 2.12 and a median value of 4.

Figure 4.2: Distribution of Risk Decisions



Risk preferences: Coin flipping experiment

To elicit risk preferences, the interviewer presented two coins. One of the coins had three stars printed on each side. The other coin had one side with seven stars and one side with zero stars. Children chose which coin should be tossed. The interviewer explained that choosing the coin with three stars on each side implies winning three stars for sure. Choosing the other coin, however, implies that the outcome (seven or zero stars) is determined by chance, with equal likelihood for the occurrence of each outcome. The fact that the safe amount (three stars) was also 'determined' by a coin toss ensures that children did not choose the risky option only for entertainment or game value. After the children had made their decision, but before actually tossing the chosen coin, the interviewer presented them two more coins in another color. Now, one coin had four stars on each side, while the other coin again had zero stars on one side and seven on the other. Children made their second decision and the interviewer tossed the two chosen coins. The order in which the two variations of the game were played was randomized.

The certainty equivalent of the "lottery coin" is 3.5. As such, only risk averse subjects

would choose the safe outcome of three stars over the lottery. Likewise, only risk seeking subjects would choose the lottery over the safe outcome of four stars. Thus, we have one situation in which we can identify risk averse subjects and one in which risk seeking subjects are identified. We classify a child as risk averse if he prefers three stars for sure over the lottery. A child is classified as risk seeking if he opts for the lottery instead of a safe amount of four. Finally, a child is risk neutral if he chooses the lottery instead of the safe amount of three and the safe amount of four instead of the lottery. Children who opt for the safe amount of three while choosing the lottery over the safe amount of four make an inconsistent choice and are excluded from the analysis.

Figure 4.2 depicts the frequencies of choices in the two lotteries excluding inconsistent choices. More children chose the lottery when the safe amount is lower. In particular, 56% of the children choose the lottery over the safe amount of three, while only 24% of the children choose the lottery in case the safe amount equals four. Overall, 39% of all children are risk averse, 29% are risk neutral, 22% are risk seeking, and 11% of the children make inconsistent choices.

Altruism: Dictator game experiment

Altruism was elicited using a binary sharing game (Fehr, Rützler, and Sutter, 2011). In the sharing game, subjects can decide between the allocations (2,0), i.e., two stars for themselves and no star for another child, or (1,1), i.e., one star for each child. Children were informed that the receiving child is of the about same age as they are, lives in the same city, but is unknown to them and has no relation to the interviewer.

Figure 4.3: Share of Altruistic Children

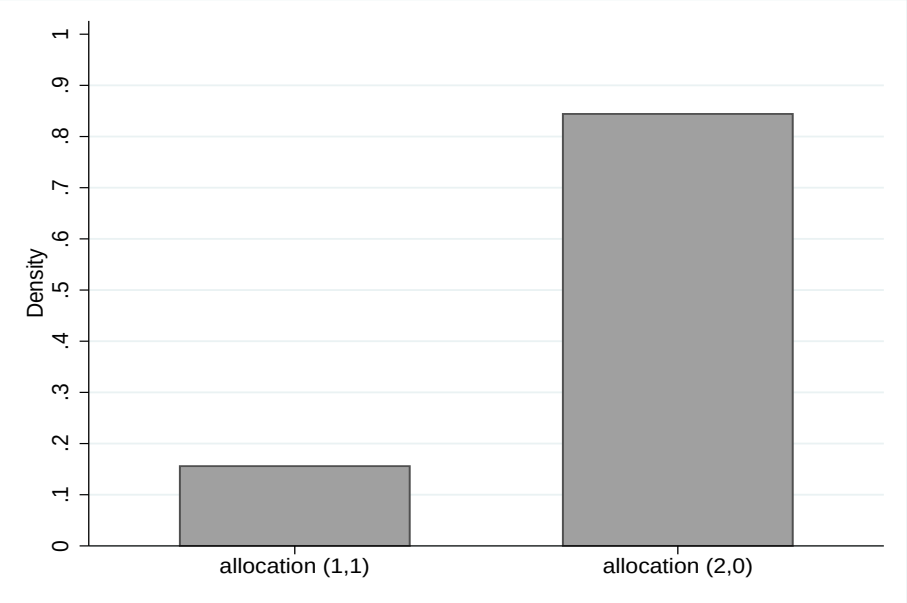


Figure 4.3 shows that about 15.6% of children behave altruistically, i.e., share the two stars equally, while 84.4% keep both stars for themselves.

IQ

We elicited two separate measures for crystallized and fluid IQ. Following the work of Cattell (1971), these two basic components form general intelligence or simply IQ. Fluid IQ measures the more hereditary part of the overall IQ that refers to general logical reasoning in new situations, intellectual capacity, or processing speed. Crystallized IQ is the part of overall IQ that broadly refers to knowledge that has been acquired in life, e.g., the vocabulary. Crystallized IQ is generally assumed to be the component of the overall IQ that is more malleable.

Figure 4.4: Distribution of Fluid IQ Scores (Histogram)

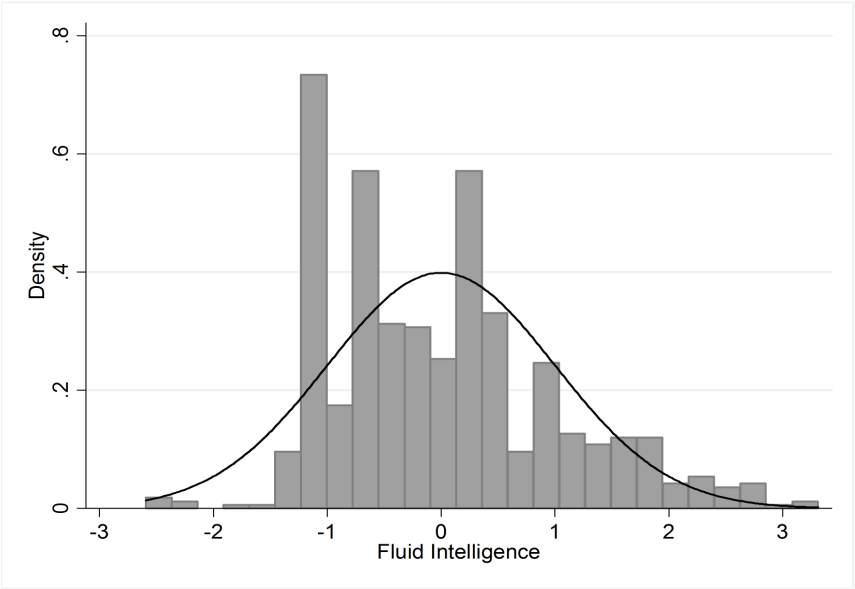
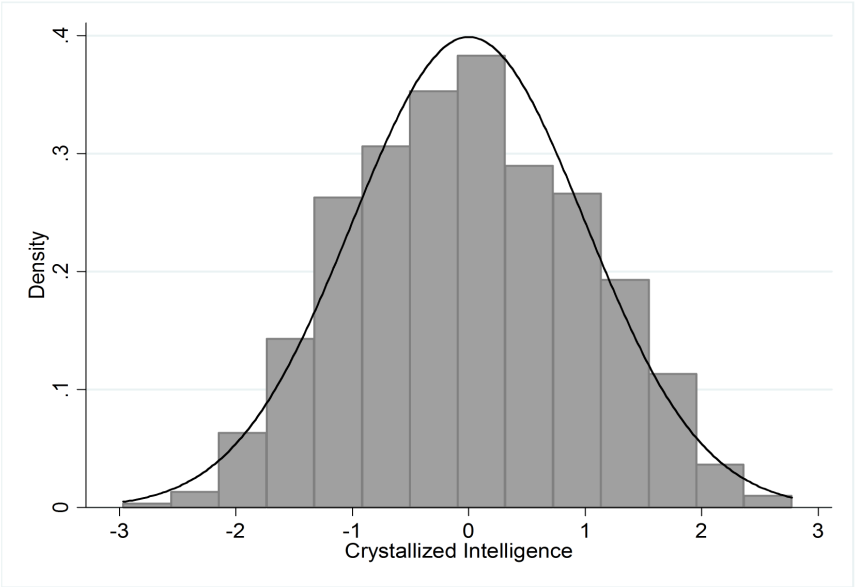


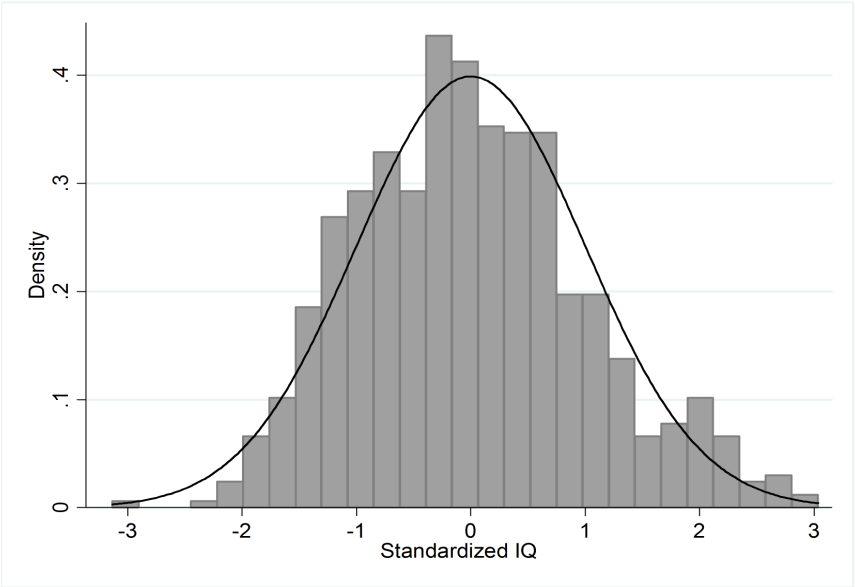
Figure 4.5: Distribution of Crystallized IQ Scores (Histogram)



We measured fluid IQ with the matrix test of the HAWIK IV, which is the German version of the well-established Wechsler IQ test for children (Petermann and Petermann, 2010). Children were presented up to 35 blocks or rows of pictures featuring different colors and forms. In each block or row one cell was missing. Each time, children had to choose

which of five pictures fitted best into the missing cell. The test contains a stopping rule which ends the test in case children produce four wrong answers in a row or in case four out of five answers in a row are wrong. The number of correct answers is our proxy for fluid IQ. Crystallized IQ was measured using 14 items of the German translation of the commonly used Peabody Picture Vocabulary Test Revised (PPVT-R) (Dunn and Dunn, 2007).[7] Here, the interviewer read out a word and showed the child four pictures. Children had to decide which picture fitted the word best. The number of correct answers is our measure for crystallized IQ.

Figure 4.6: Distribution of IQ Scores (Histogram)



We standardize both, the measure for fluid and the one for crystallized IQ. The distribution of fluid and crystallized IQ scores, which are positively correlated (correlation coefficient of 0.28), is shown in Figures 4.4 and 4.5. In each picture, the width of each bar is chosen such that each bar corresponds to one (discrete) value of the obtained IQ score. For comparison purposes, we also plot a standard normal distribution in the histograms. Moreover, we calculate the overall IQ as the sum of the two standardized variables which is then again standardized. The overall IQ scores, which are shown in Figure 4.6, lie in an interval of three standard deviations around the mean. Expressed on the typical IQ scale with mean 100 and standard deviation of 15 IQ points, we observe IQ's ranging from about

[7]Due to time constraints, we had to restrict the test to 14 items. We have chosen those 14 items that had the largest discriminatory power in the SOEP pretest data of the mother and child questionnaires "MukiIIIb" and "MukiIIIc" that were based on a 61 item version of the PPVT-R test.

55 to 145. This shows that our IQ tests sensitively differentiate between a wide range of possible IQ scores.

4.3 Estimation strategy

Our aim is to investigate how children's personality is influenced by the SES of their parents. For that purpose, we use measures of two crucial dimensions of parental SES: household income and parental education.[8] More precisely, we use average years of education of the parents and the logarithm of monthly disposable household income (in net equivalence terms) as explanatory variables. Using the logarithm allows for a decreasing marginal effect of income on our measures of personality traits.[9]

We provide three different sets of results. First, for each facet of personality under study, we document whether there is a significant relationship between parental SES and a child's personality. For that purpose, we regress the different personality traits on parental education and the logarithm of household income only (baseline specifications). In a next step, we add the child's age in months and a gender dummy as explanatory variables since previous research has documented their predictive power for a child's personality. Age and gender are clearly orthogonal to parental SES.

An obvious question that arises is how differences in SES affect a child's environment and how these environmental differences shape the formation of preferences. Are the documented differences in a child's personality only due to differences in household income and parental education, i.e., differences in the monetary and cognitive resources that parents have available for educating their child? Or are they also due to other dimensions of a child's everyday environment that differ by parental SES? The available information covers differences in family structure, initial conditions at birth, the personality profile of the child's mother, and different aspects of parental behavior such as parenting style, the time parents spent with their children, and the quality of time spent together. In the second part of the results section, we document that these variables differ significantly by parental

[8]There exist no universal consensus about how to measure SES. It is usually measured by some combination of income, education, and occupation. Bradley and Corwyn (2002) provide a brief discussion of the history and definition of the term SES. Our data also contain information on parental occupation (in 20 categories such as being self-employed, a blue-collar worker, a white collar-worker, or a civil servant). Still, we prefer focusing on parental income and education as measures of parental SES, because they are quantifiable in natural units and thus well apt for our empirical analysis. Furthermore, taken together, variation in educational attainment and income, largely captures variation in occupational status.

[9]In terms of significance, results for all dependent variables are robust to using a linear income term instead of the logarithm of income.

SES. In the last part of the results section, we add these variables as explanatory variables to our baseline specifications (full specifications).

The remainder of this section motivates and describes the variables that we use to characterize a child's environment, that could differ by parental SES and affect a child's personality. First, the work of Heckman and coauthors (Cunha and Heckman, 2007; Heckman, 2008) stresses the importance of parental investments in their children as well as children's initial conditions for the development of a child's personality. It seems plausible that parental investments and initial conditions differ by parental SES. For example, having a mother who works will increase household income and, at the same time, is likely to reduce the amount of time a mother spends with her child. To add information on some of the most important parental investments we include separate variables that capture the amount of time (in hours per week) that mothers spend with their children as the main carer and the quality of time spent together. The quality of time spent together is derived in a principal component analysis on questionnaire measures that ask for the kind and frequency of joint activities of mothers and their children (for details see Appendix "Additional Information on Explanatory Variables"). The principal component analysis results in four components that we include as additional regressors. The components capture the intensity of (i) everyday interaction with the child such as having joint meals, talking, or doing homework assignments together, (ii) highly interactive activities such as reading to a child, going in for sports or playing music with the child, (iii) low interaction activities that provide inputs to the child, but imply lower levels of interaction between parents and children such as meeting other families, shopping, going to theater, a museum, or the movies and (iv) joint media consumption, e.g., watching TV or playing computer games together. Furthermore, we control for initial conditions by including information on a child's weight at birth, the week of gestation in which a child was born, and the child's number of siblings at birth.

Second, we control for different parenting styles using six variables that reflect to which extent parenting styles are characterized by emotional warmth, negative communication, inconsistent parenting practices, monitoring, strict control, and psychological control (for details see Appendix "Additional Information on Explanatory Variables"). Doepke and Zilibotti (2012) present a theoretical model in which they argue that parenting style depends on the socio-economic environment a family lives in and that parenting style may affect children's preferences.

Third, we use information on the mother's IQ and economic preferences.[10] For adults, previous studies document intergenerational transmission of economic preferences (Dohmen et al., 2012; Kosse and Pfeiffer, 2012). For example, grown-up children whose parents are risk averse display a higher likelihood of being risk averse as well. Therefore, we include questionnaire measures of the mother's risk preferences, time preferences, and altruism as further control variables. These measures have been validated using incentivized experiments by Falk et al. (2012). Moreover, evidence for the transmission of cognitive ability from parents to their children is abundant. Besides measuring children's fluid and crystallized IQ, we have also elicited mothers' fluid IQ. Whenever an IQ score of the child is the dependent variable, we control for the mother's IQ. Whenever a child's economic preference is the dependent variable, we control for both the child's overall IQ score and the mother's IQ.

Finally, we include information on the family structure: a dummy for whether a child lives with a single parent, the mother's age at birth that could have affected the mother's level of education, and how many siblings currently live in the household.

Appendix "Additional Information on Explanatory Variables" documents how the additional control variables were elicited and Table 4.6 in the Appendix contains basic summary statistics of all explanatory variables.

4.4 Results

4.4.1 The relationship between parental socio-economic status and a child's personality

Tables 4.2 and 4.3 present the main results of our analysis. Columns (1.1), (2.1), (2.3), and (3.1) in both tables display the results of the baseline specifications. They suggest that all facets of a child's personality under study vary systematically with parental SES. The only exception is altruism. More precisely, we find that children of higher educated parents are significantly more patient (p-value < 0.01) and significantly less likely to make risk seeking choices (p-value < 0.01). Furthermore, crystallized and overall IQ are significantly higher

[10] Unfortunately, we do not have information on the father's economic preferences and IQ. However, Anger and Heineck (2010) show that mothers play a more important role in the transmission of cognitive abilities than fathers. Furthermore, Dohmen et al. (2012) document a strong positive correlation of preferences within married couples that is consistent with positive assortative mating, a prediction of the models of Bisin and Verdier (2000) and Bisin and Verdier (2001) on the cultural transmission of preferences.

Table 4.2: Baseline Specifications - Economic Preferences

Variables	(1.1) Time Preferences (Coins in Piggybank)	(1.2)	(2.1) Risk Neutrality (Coin 3-3 vs. 7-0)	(2.2)	(2.3) Risk Seeking (Coin 4-4 vs. 7-0)	(2.4)	(3.1) Altruism (Dictator Game)	(3.2)
Ln(income)	0.132	0.049	−0.013	−0.018	0.005	0.004	−0.019	−0.016
	(0.186)	(0.182)	(0.047)	(0.047)	(0.041)	(0.041)	(0.032)	(0.032)
Education	0.134***	0.152***	−0.003	−0.001	−0.021***	−0.020***	0.002	0.001
	(0.033)	(0.032)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)
Age child		0.050***		−0.001		−0.006**		0.005**
		(0.012)		(0.003)		(0.003)		(0.002)
Male		0.572***		0.099***		0.067**		−0.096***
		(0.152)		(0.038)		(0.033)		(0.027)
Constant	2.760***	−2.410**	0.177	0.341	0.006	1.629*	−1.110***	−2.757***
	(0.415)	(1.193)	(0.274)	(0.791)	(0.289)	(0.868)	(0.309)	(0.955)
Obs.	721	721	648	648	648	648	724	724
R-squared	0.035	0.072	0.005	0.013	0.030	0.042	0.001	0.029

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

We estimate the correlates of time preferences using OLS and the correlates of risk preferences and altruism with a probit model and display coefficients for OLS and average marginal effects for probit. The dependent variable for time preferences is the number of coins put into the piggybank. The binary variable risk neutrality equals 1 if a child chooses the coin with 7 stars on one and zero stars on the other side and 0 if a child chooses the coin with three stars on each side (risk aversion). The binary variable risk seeking equals 1 if a child chooses the coin with 7 stars on one and zero stars on the other side and 0 if a child chooses the coin with four stars on each side (risk aversion or risk neutrality). We exclude all inconsistent risk choices. In columns (2.1) to (2.4), we include one additional explanatory variable, a binary indicator of which coin toss decision was presented first. The binary variables altruism equals 1 if the child shares two stars equally and 0 if it keeps both stars. Income denotes monthly net household equivalence income in thousand Euro, years of education measures the mother's and father's average years of education. Children's age is measured in months, male is binary indicator that equals 1 for boys and zero for girls. R-squared displays Adjusted R-squared in OLS regressions and Pseudo R-squared in probit regressions.

for higher levels of household income and parental education (all p-values < 0.01). The same is true for fluid IQ (for income, $p < 0.05$, for parental education, $p < 0.1$).

Table 4.3: Baseline Specifications - IQ

	(1.1)	(1.2)	(2.1)	(2.2)	(3.1)	(3.2)
Variables	Overall IQ score		Crystallized IQ score		Fluid IQ score	
Ln(income)	0.348***	0.309***	0.390***	0.351***	0.168**	0.143*
	(0.084)	(0.080)	(0.082)	(0.079)	(0.085)	(0.083)
Education	0.071***	0.080***	0.084***	0.093***	0.029*	0.035**
	(0.015)	(0.014)	(0.015)	(0.015)	(0.015)	(0.015)
Age child		0.045***		0.038***		0.034***
		(0.005)		(0.005)		(0.006)
Male		0.081		0.133**		−0.003
		(0.067)		(0.066)		(0.072)
Constant	−0.944***	−5.277***	−1.120***	−4.844***	−0.392**	−3.602***
	(0.188)	(0.534)	(0.189)	(0.521)	(0.189)	(0.592)
Observations	731	731	731	731	731	731
R-squared	0.105	0.185	0.140	0.201	0.021	0.065
Robust standard errors in parentheses,*** p<0.01, ** p<0.05, * p<0.1						

We estimate the correlates of overall, crystallized, and fluid IQ scores using OLS. Crystallized IQ is the standardized outcome (mean 0, a standard deviation of 1) of the short version of the Peabody Picture Vocabulary Test Revised (PPVT-R). Fluid IQ is the standardized outcome (mean 0, a standard deviation of 1) of the matrix test of the HAWIK IV. Overall IQ is the standardized sum (mean 0, a standard deviation of 1) of our measures of crystallized and fluid IQ. Income denotes monthly net household equivalence income in thousand Euro, years of education measures the mother's and father's average years of education, children's age is measured in months, male is binary indicator that equals 1 for boys and zero for girls.

Columns (1.2), (2.2), (2.4), and (3.2) in Table 4.2 show that there is a statistically significant gender effect for each of the three key economic preferences. Boys are more patient and less altruistic than girls. They are more likely to be risk neutral instead of risk averse and more likely to be risk seeking than girls. The fact that boys are more likely to take risks is well-established in the literature (Moreira, Matsushitab, and Silva, 2010; Cárdenas et al., 2011; Sutter et al., 2010). The influence of gender on patience is less clear. While we find that boys are more patient than girls, Bettinger and Slonim (2007) and Castillo et al. (2011) find the opposite. Concerning altruism, our findings corroborate the results of Fehr, Rützler, and Sutter (2011) that are based on the same sharing game that we use. In line with our results, in the sharing game, girls of age 8 or 9 are significantly more likely than boys to choose the (1,1) option and hence are more altruistic.

Moreover, we find that older children are more patient, less likely to make risk seeking choices, and score higher in both crystallized and fluid IQ tests. The effects of age on preferences are well in line with the literature. Mischel and Metzner (1962) and Bettinger and Slonim (2007) also find that older children are more patient and, e.g., Slovic (1966) documents that older children are less willing to take risks. Also in line with our results in column (2.2) of Table 4.3, e.g., Horn and Cattell (1967) document that crystallized IQ increases with age.

4.4.2 How does a child's environment differ by parental socio-economic status?

Table 4 displays pairwise correlation coefficients and the corresponding significance levels between the logarithm of household income and parental education, our key measures of parental SES, and further variables that characterize a child's everyday environment. Most of the correlations are highly significant revealing that the environment of children from families with different SES differs significantly.

Table 4.4: Differences in a child's environment by parental SES

	Ln(income) Correlation	Significance	Education Correlation	Significance
Week of gestation	0.12	***	0.14	***
Weight at birth	0.11	***	0.13	***
N older siblings	-0.16	***	-0.16	***
Time child care	-0.13	***	-0.03	
Low interaction	-0.15	***	-0.18	***
High interaction	0.18	***	0.23	***
Media	-0.19	***	-0.25	***
Everyday	0.07	*	0.05	
Style warmth	0.12	***	0.02	
Style neg. comm.	-0.05		0.04	
Style inconsistent	-0.07	*	-0.12	***
Style strict	-0.10	**	-0.06	
Style monitor	0.08	*	0.06	
Style psycho	-0.20	***	-0.22	***
IQ child	0.28	***	0.29	***
IQ mother	0.34	***	0.37	***
Time preferences mother	0.04		0.13	***
Risk preferences mother	0.08	**	0.08	**
Altruism mother	-0.02		-0.02	
N siblings	-0.22	***	-0.10	**
Single parent	-0.03		0.03	
Age mother	0.29	***	0.32	***

Results from Table 4 show that initial conditions at birth differ significantly for children from families with different parental SES. Children from parents with higher income and higher educational attainment typically have a higher weight at birth and are born in a later week of gestation, two indicators that represent favorable initial conditions. Furthermore,

children with high SES background typically have fewer older siblings and, thus, are likely to receive more parental attention.

Also for quality of time, Table 4 documents strong differences according to parental SES. While parents with higher SES more often engage in highly interactive activities with their children such as reading to them, low SES parents more often engage in joint activities that involve lower levels of interaction such as meeting other families or shopping and spend more of the joint time with media consumption. Only in terms of joint everyday activities such as talking or having meals together differences are small. As one would expect, the correlation between household income and the time parents spend with child care is negative and significant. This correlation is likely to be due to working mothers who contribute to higher levels of household income and spend less time with their children as main care giver. In contrast, the correlation between parental education and time spent with child care is not significant.

With respect to parenting style, parents with higher SES are less likely to use inconsistent parenting practices and a parenting style that is characterized by psychological control. Inconsistent parenting practices consist of, e.g., threatening a child with a punishment without actually implementing it or the absence of consistent rules of behavior for a child. An example of a parent who is exerting psychological control is a parent who does not talk to his child for while because the child did something wrong. Moreover, in families with higher income, parenting styles that are less strict (e.g., rely less on punishment) and characterized by emotional warmth (e.g., praising a child or showing a child that parents love it) are significantly more likely to prevail.

Table 4 also documents significant correlations between SES and maternal personality. As their children, high SES mothers tend to have a higher fluid IQ score and to be more patient. Furthermore, high SES mothers tend to be less risk averse. The correlation between maternal altruism and SES, however, is not significant.

Finally, families with different SES also differ in family structure. On average, high SES families consist of significantly older mothers with fewer children. The correlation between parental SES and living in a single parent family is not significant. However, the single parent dummy is the only variable for which the correlation coefficient changes substantially when looking at the mother's years of education only instead of the average parental years of education. The correlation between maternal years of education and the single parent dummy is 0.19 and highly significant ($p < 0.01$), i.e., children of higher educated mothers,

but not children of higher educated parents in general, are more likely to live in a single parent family.

In sum, the correlations in Table 4 document that the childhood environment differs significantly between high and low SES households. This is true for almost all dimensions we have elicited. Building on these findings we now incorporate these environmental factors in the baseline specifications described in Tables 2 and 3.

4.4.3 Which differences in a child's environment translate into differences in a child's personality?

Table 5 displays results of the full specifications in which, for each facet of personality under study, we add information on the mother's IQ and economic preferences, maternal time investments, the quality of time mothers and children spend together, parenting style, initial conditions at birth, and family structure as explanatory variables. Except for the time the mother spends as the main carer of her children, all types of additional variables add explanatory power. Furthermore, in the full specifications, parental income and years of education remain significant predictors of many facets of a child's personality, in particular a child's patience, overall IQ, and crystallized IQ. However, the size of the coefficients of household income and average parental years of education is about 20 to 40% smaller than in the baseline specifications.

Table 4.5: Full specifications

Variables	Time Pref.	Risk Neutral	Risk Seeking	Altruism	Overall IQ	Crystal. IQ	Fluid IQ
Ln(income)	−0.176	0.014	0.012	−0.022	0.216**	0.224**	0.122
Education	0.099**	0.000	−0.012	0.000	0.057***	0.063***	0.028
Age child	0.038***	−0.002	−0.008***	0.002	0.040***	0.037***	0.028***
Male	0.530***	0.111***	0.047	−0.118***	0.082	0.137*	−0.006
IQ child	0.291***	0.031	0.032	0.010			
Dummy wave 2	0.853	0.241	0.342	−0.070	0.774	0.993	0.246
IQ mother	0.035	−0.037	−0.031	−0.008	0.077*	0.092**	0.031
Time pref. mother	−0.036	−0.003	0.008	−0.007	−0.014	−0.008	−0.015
Risk pref. mother	0.062*	0.018**	0.001	−0.002	−0.017	−0.004	−0.023
Altruism mother	0.087**	0.005	0.000	−0.003	−0.000	0.025	−0.025
Week gestation	0.006	0.005	0.004	−0.000	−0.001	−0.006	0.005
Weight at birth	0.004	0.004	−0.004	−0.004	0.017**	0.010	0.018**
# older siblings	−0.195*	0.042	0.048**	−0.029	−0.065	−0.094*	−0.010
Dummy time	−0.475	−0.021	0.060	−0.029	−0.040	−0.085	0.021
Time child care	0.001	0.001	0.000	−0.000	0.000	0.001	−0.001
Dummy quality	0.363	0.131**	0.002	0.046	−0.175*	−0.174*	−0.106
Low interaction	−0.021	0.025	0.023	−0.022	−0.033	−0.036	−0.017
Everyday	0.030	−0.015	−0.002	−0.017	0.067**	0.069***	0.038
Media	−0.018	−0.007	0.013	0.029**	−0.023	−0.063*	0.027
High interaction	−0.055	−0.034*	−0.044**	−0.003	−0.014	0.020	−0.042
Style warmth	−0.064	−0.038	−0.031	0.030	−0.055	−0.046	−0.042
Style neg. comm.	−0.237	−0.016	0.012	−0.027	−0.020	0.013	−0.044
Style inconsistent	0.193	0.051*	0.027	0.030	−0.089*	−0.017	−0.126**
Style strict	−0.024	0.002	0.020	−0.004	0.022	0.013	0.022
Style monitor	−0.037	−0.024	−0.053	0.014	0.019	−0.066	0.097
Style psycho	0.025	−0.008	−0.005	−0.044	−0.063	−0.169**	0.068
# siblings	0.014	0.020	−0.005	−0.013	−0.022	−0.011	−0.024
Single parent	−0.331*	0.051	−0.001	−0.012	0.199**	0.095	0.224**
Age mother	0.013	−0.001	−0.003	−0.004	0.002	0.008	−0.005
Lottery 4-4 first		0.082*	0.109***				
Constant	−1.757	−1.288	1.420	−0.139	−5.037***	−4.724***	−3.338***
Observations	629	563	563	631	638	638	638
R-squared	0.074	0.069	0.096	0.097	0.208	0.219	0.075
*** p<0.01, ** p<0.05, * p<0.1							

We estimate the correlates of time preferences, overall, crystallized, and fluid IQ scores using OLS and the correlates of altruism and risk preferences with a probit model. We display average marginal effects for probit. The dependent variables are exactly the same as in Tables 2 and 3. In columns (2.1) to (2.4), we additionally include a binary variable that equals 1 if the first lottery decision was 4-4 versus 7-0 and 0 otherwise. Income denotes monthly net household equivalence income in thousand Euro, years of education measures the mother's and father's average years of education, children's age is measured in months, male is binary indicator that equals 1 for boys and 0 for girls. All other explanatory variables are described in the Appendix. R-squared displays Adjusted R-squared in OLS regressions and Pseudo R-squared in probit regressions.

In the following, we discuss the results of the full specifications in more detail. Children of higher educated parents are significantly more patient ($p < 0.01$). The child's overall IQ score is a highly significant predictor of a child's patience ($p < 0.01$). Our results on the relationship between IQ and time preferences corroborate the results by Shamosh and Gray (2008). In a meta-study they find that cognitive skills and patience are positively correlated. Mischel and Metzner (1962) find a positive correlation of delayed gratification and IQ, while Bartling et al. (2010) find a positive correlation of patience with crystallized IQ only. Furthermore, children with (more) older siblings or a single parent are predicted to be less patient (both $p < 0.1$). Finally, more altruistic mothers have more patient children ($p < 0.05$). These latter three findings suggest that patience is a trait that needs to be trained under the guidance of an adult that dedicates all attention exclusively to the child.

The shape of the mother's risk preferences, the number of older siblings, and the quality of time parents and children spend together have explanatory power for a child's risk preferences. Our results provide evidence for a direct intergenerational transmission of risk preferences already in childhood. A child is more likely to be risk neutral instead of risk averse if its mother is less risk averse ($p < 0.05$). If mothers spend more high quality time with their children, their children are both less likely to be risk averse ($p < 0.1$) and less likely to be risk-seeking ($p < 0.05$), which, taken together, results in a higher likelihood that children are risk neutral. Finally, children who have (more) older siblings are more likely to be risk seeking ($p < 0.05$). For risk seeking, the coefficient of parental years of education is no longer significant ($p = 0.12$) in the full specification.

Both the baseline and the full specifications underline that a child's degree of altruism does not differ by parental SES. Despite the comprehensive set of control variables, it is actually hard to project a child's altruistic behavior at all. In the full specification, a child's gender remains the only variable with predictive power. As an aside, when adding children's amount of pocket money (i.e., children's "own income") as explanatory variable to the full specification, we find that higher amounts of pocket money are associated with lower levels of altruism ($p < 0.05$).[11]

The mother's IQ, the initial conditions at birth, the quality of parental time, parenting style, and family structure are all significant predictors of a child's IQ. On average, mothers with a higher fluid IQ have children with a higher overall and crystallized IQ ($p < 0.1$ and $p < 0.05$, respectively). More favorable initial conditions are positively correlated

[11]For all other personality traits under study, the amount of pocket money is not significant when added as explanatory variable.

with IQ: children with a higher weight at birth are predicted to have a higher overall and fluid IQ score ($p < 0.05$). On average, children with (more) older siblings have lower crystallized IQ scores ($p < 0.1$). Spending more time with talking to the child or having joint meals (i.e., in "everyday interactions") increases the child's crystallized and overall IQ ($p < 0.05$). In contrast, spending more time with joint media consumption is associated with a lower crystallized IQ ($p < 0.1$). Inconsistent parenting practices are predicted to lower overall and fluid IQ ($p < 0.1$ and $p < 0.05$, respectively), a parenting style that aims at psychological control is associated with a lower crystallized IQ ($p < 0.05$). Finally, children of single parents who, on average, have significantly better educated mothers score higher on overall and fluid IQ ($p < 0.05$). Despite the strong explanatory power of the control variables, children from high SES families are still predicted to have significantly higher overall and crystallized IQ scores (for income, p-values < 0.05, for parental education, p-values < 0.01). For fluid IQ, the coefficients of household income ($p = 0.22$) and parental education ($p = 0.15$) that were (marginally) significant in the baseline specification are no longer significant.

The effect of parental SES on a child's personality is not only statistically significant, but also relevant in terms of effect size. For example, in the full specifications, the patience of a child whose parents have four more years of education, e.g., a university degree as opposed to a standard apprenticeship, is as high as if the child was about one year older. On average, having parents who have one more year of education is associated with an increase in a child's patience to the level of a child who has a mother with a 5 point higher IQ score. The effect of having parents with one more year of education is predicted to increase a child's overall and crystallized IQ by about 0.06 standard deviations (i.e., about 1 point). A 10% increase in household income is predicted to increase both a child's overall and crystallized IQ by about 0.3 points.

4.4.4 Related literature

The literature on the relationship between a child's economic preferences and parental SES is basically nonexistent. We are not aware of any other study that investigates the effect of parental SES on children's time preferences. Delaney and Doyle (2012) is the study that comes closest to analyzing this relationship. They use parental answers to questions concerning psychological concepts such as hyperactivity, impulsivity, and persistence of three year old children and show that children from families with higher SES (measured by

household income and parental education) are less impulsive.

Concerning risk preferences, Alan et al. (2013) study the intergenerational transmission of risk attitudes and use maternal and paternal years of education as control variables that turn out not to be significant. With an average of 5.9 years of education for mothers and 7.4 years for fathers their sample seems to be largely restricted to low SES families. Furthermore, they use information on a family's belongings and monthly expenditures to construct four dummies that split their sample in SES quartiles. These dummies do not have predictive power for boys' risk attitudes, but girls from low SES families are less risk averse. In their data, mother's and child's risk attitudes are measured in a similar, incentivized task. However, their measure of risk attitudes cannot distinguish between risk neutrality and risk seeking, i.e., the range in which we document a significant relation between parental education and risk preferences in the baseline specification. For adults, there is evidence that parental education enhances the willingness to take risk (Dohmen et al., 2011b). At first sight, this finding seems to contradict our finding that higher parental education reduces the probability of being risk seeking as opposed to being risk neutral or risk averse. However, only a small fraction of adults is risk seeking (e.g., 9% in Dohmen et al. (2011b)), while 24% of the children in our study are risk seeking. Hence, an increased willingness to take risks for adults essentially means going from risk aversion in the direction of risk neutrality. That corresponds to what we find for children: On average, children of higher educated parents have risk preferences which are closer to risk neutrality. The only difference is that they move in the direction of risk neutrality from risk seeking instead of risk aversion.

In the domain of social preferences, Bauer, Chytilová, and Pertold-Gebicka (2011) is the only closely related study.[12] They assess children's social preferences in a series of choice tasks as in Fehr, Bernhard, and Rockenbach (2008) and Fehr, Rützler, and Sutter (2011) and find that children from higher educated parents are more likely to share two tokens equally with another child instead of keeping both for themselves in the binary sharing game. The educational attainment of the parents is proxied by a dummy variable indicating low education, i.e., that neither mother nor father has completed secondary school education with a leaving exam. In contrast to their study, we do not find an effect of parental SES on children's altruism. We can only speculate why results differ. Notable

[12]Benenson, Pascoe, and Radmore (2007) also present evidence that higher SES leads to higher levels of altruism. In their study, however, SES is only measured at school level using the fraction of children who receive a free lunch.

differences between both studies comprise sample size, the way parental SES is measured, and the kind of controls that are used.[13]

While research on the relation of parental SES and children's economic preferences is still in its infancy, the effect of parental SES on children's overall IQ is well established. Neff (1938) documented the positive correlation of IQ and parental SES. In a study on adopted children, Capron and Duyme (1989) use information on parental SES of both foster and biological parents of the same children to illustrate that parental SES is positively correlated with children's IQ even if the effect cannot work through genetic transmission. Rindermann, Flores-Mendoza, and Mansur-Alves (2010) and Turkheimer et al. (2003) are examples of studies that separately analyze the role of parental SES for crystallized and fluid IQ. Again, they document a significant and positive effect of higher parental SES on both components of IQ. Anger and Heineck (2010) and Rindermann, Flores-Mendoza, and Mansur-Alves (2010) point to a larger parental influence on crystallized IQ as opposed to fluid IQ that is supposed to have a stronger hereditary component than crystallized IQ. Similarly, Hackman and Farah (2009) provide evidence that the effect of poverty is especially strong for certain neurocognitive systems, including language ability that is reflected in crystallized IQ scores. Our findings are in line with all results on the relationship between parental SES and IQ that are described above.

4.5 Discussion

Our results show that parental SES is a systematic predictor of a child's personality. Children from families with higher SES are more patient, less likely to be risk seeking, and score higher on tests of crystallized, fluid, and overall IQ. In a first, purely correlational analysis, we have focused on how two core dimensions of parental SES, i.e., household income and parental education, relate to a child's personality. As a second step, we have documented that the family environment that children from families with different SES live in differs not only with respect to parental education and household income, but also in many other dimensions. In the final step of analysis, we included a broad set of further explanatory variables that characterize a child's environment and reflect systematic differences between families with different SES such as quantity and quality of time parents spend with their

[13]We use more detailed, continuous measures of parental SES capturing income and education, while Bauer, Chytilová, and Pertold-Gebicka (2011) focus on a binary indicator variable for education. As control variables, Bauer, Chytilová, and Pertold-Gebicka (2011) use a child's age, gender, and dummies indicating whether parents are separated and whether the mother is working full time.

children, parenting style, the mother's IQ score and economic preferences, initial conditions at birth, and family structure. While these additional variables have explanatory power for the shape of a child's personality, our results document an effect of parental education and income on top of these variables.

Our results enhance the understanding of the origins of heterogeneity in personality. In the following, we discuss further implications of our finding that children from families with different SES have different personality profiles.

First, we discuss our results in the light of the literature that relates preferences and cognitive ability to many important outcomes in life. For patience and IQ, the claim that higher levels tend to be more favorable for many life outcomes is largely uncontroversial. For example, higher levels of IQ are associated with higher levels of education (Heckman and Vytlacil, 2001) and income (Hanushek and Woessmann, 2008). Also higher levels of patience predict a wide range of positive outcomes in later life such as higher educational attainment (Shoda, Mischel, and Peake, 1990), substantially higher earnings, less use of welfare and fewer days of unemployment (Golsteyn, Grönqvist, and Lindahl, 2013), and better health outcomes (Chabris et al., 2008; Sutter et al., 2010; Golsteyn, Grönqvist, and Lindahl, 2013; Shoda, Mischel, and Peake, 1990). In that sense, our results suggest that, on average, children from families with lower SES are disadvantaged already when they are about seven to nine years old. Of course, this line of reasoning assumes that disparities in personality of children from different socio-economic backgrounds persist or even increase as children grow older. The literature on the formation of cognitive and non-cognitive skills does indeed provide evidence for growing disparities that are due to the self-reinforcing and cross-fertilizing character of skills (Cunha and Heckman, 2007).

In contrast to patience and IQ, there is no obvious optimal degree of risk aversion that is independent from the environment an individual lives in. Doepke and Zilibotti (2012) introduce the distinction between endogenous and exogenous risk that individuals are exposed to. While exogenous risks cannot be avoided, taking an endogenous risk is a deliberate decision that depends on the individual risk attitude. If families with low SES are more strongly exposed to exogenous risks such as, for example, street crime in their neighborhood or the risk of becoming unemployed, low SES parents have lower incentives to intervene their children's risk seeking behavior. Being less risk averse might be helpful to successfully cope in a relatively high risk environment. Coping to environments with different degrees of exposure to exogenous risks could explain the pattern that we observe

in our data that children from families with lower SES are more likely to be risk seeking. Also with respect to endogenous risks, it is hard to claim that there is a unique optimal level of risk attitude. For example, Dohmen et al. (2011b) document that a higher willingness to take risks is both associated with outcomes that are typically thought of as detrimental (e.g., smoking) or supportive to good health (e.g., taking exercises).

Second, our results provide insights when it comes to explaining social immobility, i.e., the fact that, as adults, children from high (low) SES families tend to have higher (lower) SES themselves. One possible explanation is that, for children in high SES families, there are more resources available, which can be invested into forming personality traits that are promising for obtaining a higher educational attainment and a higher income. For the case of time preferences, this idea is, for example, formulated by Becker and Mulligan (1997). Our findings support their hypothesis: Children from families with higher SES have lower discount rates. Hence, their time preferences will induce them to make decisions which are more forward looking and therefore more profitable in the long-run. Available resources also significantly influence a child's ability as measured by IQ. Together, the effects of parental SES on a child's personality result in a tendency to favor a similar outcome in terms of SES, i.e., in social immobility.

Finally, our results also deliver new insights for studies that focus on explaining life outcomes by different preference profiles in childhood. Consider, for example, the seminal work by Mischel and co-authors (Mischel, Shoda, and Peake, 1988; Mischel, Shoda, and Rodriguez, 1989; Shoda, Mischel, and Peake, 1990). In a series of experiments, they measure children's patience at the age of four in the so-called Marshmallow task. In this task, children were presented two marshmallows. If they were able to abstain from eating the first marshmallow for about 15 minutes, they also received the second marshmallow. The amount of "self-imposed delay of gratification" at the age of four is significantly related to, e.g., academic and social competence, verbal fluency, and the skill level ten years later (Mischel, Shoda, and Rodriguez, 1989). The results of Mischel, Shoda, and Rodriguez (1989) are obtained without controlling for parental SES. We show that children from low SES families exhibit lower levels of patience, the economic concept that is most closely related to delay of gratification. Hence, studies that investigate the effect of time preferences on outcomes without controlling for parental SES are likely to overestimate the effect of time preferences due to omitted variable bias. Sutter et al. (2010) investigate the role of time and risk preferences of adolescents for their behavior. Controlling for age, gender,

the number of siblings, the amount of pocket money and the German and math grades, they find that more impatient adolescents are more likely to spend money on alcohol and cigarettes, have a higher BMI, and are less likely to save money. In contrast, risk preferences are only a weak predictor of behavior. Since Sutter et al. (2010) do not control for parental SES, the coefficient of time preferences could potentially pick up the influence of parental SES on behavior. While the work of Mischel, Shoda, and Rodriguez (1989) and Sutter et al. (2010) is highly relevant, our results highlight the need for future research estimating the relationship between economic preferences and life outcomes or behavior using a rich set of control variables, among them parental SES. For example, advocating childhood interventions aimed at increasing children's patience may be a useful policy advice if the effects documented in Mischel, Shoda, and Rodriguez (1989) and Sutter et al. (2010) are indeed driven by preferences. If, however, less favorable health outcomes and behaviors are due to lower levels of parental monetary or cognitive resources instead of lower levels of patience addressing patience would not lead to a change health outcomes and behaviors.

For future research, it is interesting to investigate how the personality profiles of children from families with different SES develop over time. Do the differences in personality stay constant, do they tend to converge during adolescence or do they further diverge? Is it true that, e.g., low levels of patience in childhood imply low levels of patience in adulthood or is the rank order of preferences not that stable? With a test-retest reliability of about 0.7, IQ is known to be quite rank-order stable already after age 6-10 (Hopkins and Bracht, 1975; Schuerger and Witt, 1989). In contrast, we are not aware of any study that presents evidence on the stability of economic preferences that is based on longitudinal data. Moreover, it seems important to analyze how the childhood environment beyond the family influences the development of personality in childhood. For example, interventions in the childhood environment might be able to loosen the link between parental SES and a child's personality.

A4 Appendix to Chapter 4

B4 Additional Figures

Figure 4.7: Arrangement of Presents

C4 Additional Tables

Table 4.6: Summary statistics

Variables	Observations	Mean	Standard Deviation	Minimum	Maximum
Ln(income)	731	1.27	0.67	0.19	7.14
Education	732	12.81	2.79	7	18
Male	732	0.52	0.50	0	1
Age child	732	93.39	6.29	84	113
# siblings	732	1.18	1.05	0	7
Single parent	732	0.36	0.48	0	1
Age mother	701	30.78	6.04	14.67	49.25
Dummy wave 2	732	0.83	0.38	0	1
IQ mother	590	-3.18e-09	1	-3.25	2.65
Time pref. mother	711	7.57	2.15	0	10
Risk pref. mother	713	4.93	2.66	0	10
Altruism mother	715	7.85	1.92	0	10
Style warmth	595	4.38	0.52	2	5
Style neg. comm.	595	2.06	0.68	1	4.5
Style inconsistent	595	2.30	0.80	1	5
Style strict	593	2.69	0.77	1	5
Style monitor	595	4.74	0.43	2	5
Style psycho	594	1.44	0.58	1	4.5
Dummy time	732	0.87	0.33	0	1
Time child care	640	59.81	39.12	0	168
Dummy quality	732	0.83	0.37	0	1
Low interaction	610	-2.20e-09	1.42	-3.22	5.89
Everyday	610	-1.36e-10	1.33	-10.22	2.47
Media	610	9.50e-10	1.34	-2.88	4.47
High interaction	610	2.13e-10	1.38	-3.75	5.42
Week gestation	712	38.64	2.73	23	47
Weight at birth	710	33.11	6.39	11	53.5
# older siblings	724	1.80	1.08	1	11

D4 Additional Information on Explanatory Variables

All additional control variables are based on mothers' answers to questions of the mother questionnaire.

Single parent – dummy variable that equals 1 if a parent is living together with a child only (and not with a husband, wife, or partner) and 0 otherwise

Age mother – age of the mother at birth of the child (in years)

Dummy wave 2 – dummy variable indicating whether information from wave 2 is available for a particular individual. The dummy is also used in an interaction with mothers' IQ scores and information on parenting styles (all other variables stem from the first wave of data collection).

IQ mother – IQ score of the mother is based on a 10 item subset of the Standard Progressive Matrices Plus (SPM Plus) test. We have chosen the 10 item subset to obtain maximal discriminatory power across individuals according to own pretests. The variable corresponds to the standardized number of right answers. In the full specifications, we use the interaction of the variable "IQ mother" and "Dummy wave 2".

All information on parenting style was elicited in the questionnaire of wave 2. Consequently, in the full specifications, we use six interaction terms of the parenting style variables listed below and the "Dummy wave 2" as control variables. Each of the six parenting style variables is based on two (out of originally three) items of the parental questionnaire for seven to eight year old children in the SOEP (Bioage08a and Bioage08b). For each dimension of parenting style, we have chosen those two items that had the highest corrected-item-total-correlation in the SOEP waves from 2010 and 2011. All items have a common scale ranging from 1 (never) to 5 (always). To assign a single value to each style, we sum the scores of the two items and divide the sum by two if both items are available. If information on one item is missing, we use the available information from the other item as the value of the style. The introductory question was "How often do the following things happen?". Below, we report the wording of the two items used for each style.

Style warmth – I show my child with words and gestures that I like him/her. I praise my child.

Style neg. comm. – I yell at my child because he/she did something wrong. I scold my child because I am angry at him/her.

Style inconsistent – I threat my child with a punishment but do not actually follow through. I find it hard to set and keep consistent rules for my child.

Style strict – If my child does something against my will, I punish him/her. I make it clear to my child that he/she is not to break the rules or question my decisions.

Style monitor – When my child goes out, I know exactly where he/she is. When my child goes out, I ask what he/she did and experienced.

Style psycho – I think my child is ungrateful when he/she does not obey me. I do not talk to my child for a while because he/she did something wrong.

Time pref. mother – Standardized answer to the question: How would you describe yourself: Are you generally an impatient person, or someone who always shows great patience? Please tick a box on the scale, where the value 0 means "very impatient" and the value 10 means "very patient" (source: SOEP).

Risk pref. mother – Standardized answer to the question: How do you see yourself: Are you generally willing to take risks (risk-prone), or do you try to avoid risks (risk-averse)? Please answer on a scale from 0 to 10, where 0 means risk-averse and 10 means risk-prone (source: SOEP).

Altruism mother – Standardized answer to the question: How would you assess your willingness to share with others without expecting anything in return, for example your willingness to give to charity? Please use a scale from 0 to 10, where 0 means you are "completely unwilling to share" and a 10 means you are "very willing to share". You can also use the values in-between to indicate where you fall on the scale.

Dummy time – dummy variable that equals 1 if information on the variable "time child care" is available and 0 otherwise; the dummy is used in an interaction with the information on time spent with child care.

Time child care – answer to the question: "Please consider a typical week: How many hours per week are you the main care giver of your children?". In the full specifications, we use the interaction of the variable "time child care" and "Dummy time".

Dummy quality – dummy variable that equals 1 if information on all four variables "Low interaction", "Everyday", "Media", and "High interaction" is available and 0 otherwise.

Low interaction, Everyday, Media, High interaction – The four variables containing information on the quality of time mothers and children spent together are derived in a rotated Principal Component Analysis on the following 16 items that results in four principal components. Most of the items are taken from the German version of the child questionnaire 5-6 years old from the SOEP wave 2008. The introductory question was: "How many times during the last 14 days did you or the main care giver engage in the following activities with your child?". Answers were given on a four item scale: daily – several times per week – at least once per week – never. List of items: (1) Do homework assignments with the child, (2) talk to each other, (3) have a joint meal (lunch / dinner), (4) have a joint snack (e.g., eat cake), (5) outdoor activities (take a walk etc.), (6) go shopping with the child, (7) visit other families with children, (8) painting or doing arts and crafts, (9) playing cards/game of dice, (10) watching television or videos with the child, (11) playing PC or internet games together, (12) going to theater for children, circus, museum etc., (13) reading/telling German stories, (14) go in for sports with the child, (15) go to music lessons or play music together, (16) go to the movies.

Roughly speaking, the variable "Low interaction" loads high on the factors (6), (7), (12),

and (16), the variable "high interaction" on (13), (14), and (15), the variable "media" on (10) and (11), and the variable "everyday" on (3), (4), (1), (2), and (5).

Week gestation – indicates the week of gestation in which the child was born

Weight at birth – indicates the weight of the child at birth (in 10 grams)

 # older siblings at birth – indicates the number of siblings at birth

Bibliography

Abramovich, Felix, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. 2006. "Adapting to Unknown Sparsity by Controlling the False Discovery Rate." *The Annals of Statistics* 34 (2):584–653.

Acosta-González, Eduardo and Fernando Fernández-Rodríguez. 2007. "Model selection via genetic algorithms illustrated with cross-country growth data." *Empirical Economics* 33:313–337.

Akerlof, G.A. and R.J. Shiller. 2009. *Animal Spirits: How Human Psychology Drives the Economy, and Why it Matters for Global Capitalism.* Princeton: Princeton University Press.

Alan, Sule, Nazli Baydar, Teodora Boneva, Thomas F. Crossley, and Seda Ertac. 2013. "Parental Socialization Effort and the Intergenerational Transmission of Risk Preferences." *Working Paper* .

Allport, G. W. and H. S. Odbert. 1936. "Traitnames. A psycho-lexical study." *Psychological Monographs* 47 (211):171.

Almlund, M., A.L. Duckworth, J. Heckman, and T. Kautz. 2011a. "Personality Psychology and Economics." *IZA DP No. 5500* .

———. 2011b. "Personality Psychology and Economics." *Handbook of the Economics of Education, Volume 4* :1–181.

Anderson, J., S. Burks, C. DeYoung, and A. Rustichini. 2011. "Toward the Integration of Personality Theory and Decision Theory in the Explanation of Economic Behavior." *mimeo* .

Anger, Silke and Guido Heineck. 2010. "Do Smart Parents Raise Smart Children? The Intergenerational Transmission of Cognitive Abilities." *Journal of Population Economics* 23:1255–1282.

Barro, R. J. and X. X. Sala-i-Martin. 1995. *Economic Growth*. McGraw Hill.

Barro, Robert J. 1991. "Economic Growth in a Cross-Section of Countries." *Quarterly Journal of Economics* 106 (2):407–443.

Bartling, Bj́orn, Ernst Fehr, Barbara Fischer, Fabian Kosse, Michel Maréchal, Friedhelm Pfeiffer, Daniel Schunk, J́urgen Schupp, C. Katharina Spieß, and Gert G. Wagner. 2010. "Determinanten kindlicher Geduld – Ergebnisse einer Experimentalstudie im Haushaltskontext." *Journal of Applied Social Science Studies - Schmollers Jahrbuch* 130:297–323.

Bauer, Michal, Julie Chytilová, and Barbara Pertold-Gebicka. 2011. "Effects of Parental Background on Other-Regarding Preferences in Children." *IZA Discussion Papers 6026, Institute for the Study of Labor (IZA)* .

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. 2012. "The Relationship Between Economic Preferences and Psychological Personality Measures." *Annual Review of Economics* 4 (1):453–478.

Becker, Gary S. and Casey B. Mulligan. 1997. "The Endogenous Determination of Time Preference." *The Quarterly Journal of Economics* 112:729–758.

Ben-Ner, A. and A. Kramer. 2010. "Personality and Altruism in the Dictator Game: Relationship to Giving to Kin, Collaborators, Competitors, and Neutrals." *Personality and Individual Differences* 51 (3):216–221.

Benenson, Joyce F., Joanna Pascoe, and Nicola Radmore. 2007. "Children's Altruistic Behavior in the Dictator Game." *Evolution and Human Behavior* 28:168–175.

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B* 57 (1):289–300.

Benjamini, Y., A. M. Krieger, and D. Yekutieli. 2006. "Adaptive linear step-up procedure that control the false discovery rate." *Biometrika* 93 (3):491–507.

Benjamini, Y. and D. Yekutieli. 2001. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics* 29 (4):1165–1188.

Benjamini, Yoav and Yulia Gavrilov. 2009. "A Simple Forward Selection Procedure Based on False Discovery Rate Control." *The Annals of Applied Statistics* 3 (1):179–198.

Berg, J., J. Dickhaut, and K. McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10 (1):122–142.

Bettinger, Eric and Robert Slonim. 2007. "Patience among Children." *Journal of Public Economics* 91:343–363.

Bibby, P.A. and E. Ferguson. 2010. "The Ability to Process Emotional Information Predicts Loss Aversion." *Personality and Individual Differences* 51 (3):263–266.

Blumkin, T., B.J. Ruffle, and Y. Ganun. 2010. "Are Income and Consumption Taxes Ever Really Equivalent? Evidence from a Real-Effort Experiment with Real Goods." *IZA Disussion Paper No.5145* .

Boes, Stefan, Markus Lipp, and Rainer Winkelmann. 2007. "Money Illusion under Test." *Economics Letters* 94 (3):332–337.

Borghans, L., B.H.H. Golsteyn, J.J. Heckman, and H. Meijers. 2009. "Gender Differences in Risk Aversion and Ambiguity Aversion." *Journal of the European Economic Association* 7 (2–3):649–658.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43 (4):972–1059.

Bradley, Robert H. and Robert F. Corwyn. 2002. "Socioeconomic Status and Child Development." *Annual Review of Psychology* 53:371–399.

Brock, W. A. and S.N. Durlauf. 2001. "Growth Empirics and Reality." *The World Bank Economic Review* 15:229–272.

Capron, Christiane and Michel Duyme. 1989. "Assessment of Effects of Socio-Economic Status on IQ in a Full Cross-Fostering Study." *Nature* 340:552–554.

Cárdenas, Juan-Camilo, Anna Dreber, Emma von Essen, and Eva Ranehill. 2011. "Gender Differences in Competitiveness and Risk Taking: Comparing Children in Colombia and Sweden." *Journal of Economic Behavior and Organization* 83 (1):11–23.

Castillo, Marco, Paul J. Ferraro, Jeffrey L. Jordan, and Ragan Petrie. 2011. "The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children." *Journal of Public Economics* 95:1377–1385.

Cattell, Raymond Bernard. 1971. *Abilities: Their Structure, Growth, and Action.* New York: Houghton Mifflin.

Chabris, Christopher F., David Laibson, Carrie L. Morris, Jonathon P. Schuldt, and Dmitry Taubinsky. 2008. "Individual Laboratory-Measured Discount Rates Predict Field Behavior." *Journal of Risk and Uncertainty* 37 (2):237–269.

Chetty, R., A. Looney, and K. Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4):1145–1177.

Clark, Andrew E., Paul Frijters, and Michael A. Shields. 2008. "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles." *Journal of Economic Literature* 46 (1):95–144.

Clark, Andrew E. and Andrew J. Oswald. 1996. "Satisfaction and comparison income." *Journal of Public Economics* 61:359–381.

Cohen, Jacob. 1988. *Statistical power analysis for the social sciences (2nd ed.).* Hillsdale, NJ: Erlbaum.

———. 1992. "A Power Primer." *Psychological Bulletin* 112 (1):155–159.

Costa, P.T. and R.R. McCrae. 1989. *NEO-PI/FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

———. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI).* Odessa, FL, Psychological Assessment Resources.

Crespo Cuaresma, Jesus. 2011. "How Different Is Africa? A Comment on Masanjala and Papageorgiou." *Journal of Applied Econometrics* 26:1041–1047.

Cunha, Flavio and James J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2):31–47.

Daly, M., L. Delaney, and C.P. Harmon. 2009. "Psychological and Biological Foundations of Time Preferences." *Journal of the European Economic Association* 7 (2–3):659–669.

Davidson, Russell and James G. MacKinnon. 1981. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." *Econometrica* 49 (3):781–793.

Deckers, Thomas and Christoph Hanck. 2013. "Multiple Testing for Output Convergence." *Macroeconomic Dynamics* Forthcoming, available at: dx.doi.org/10.1017/S1365100512000338.

———. forthcoming. "Variable Selection in Cross-Section Regressions: Comparisons and Extensions." *Oxford Bulletin of Economics and Statistics* .

Delaney, Liam and Orla Doyle. 2012. "Socioeconomic Differences in Early Childhood Time Preferences." *Journal of Economic Psychology* 33:237–247.

Di Tella, Rafael and Robert MacCulloch. 2006. "Some uses of happiness data in economics." *The Journal of Economic Perspectives* 20 (1):25–46.

Doepke, Matthias and Fabrizio Zilibotti. 2012. "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission." *Discussion Paper* .

Dohmen, T., A. Falk, D. Huffman, and U. Sunde. 2008. "Representative Trust and Reciprocity: Prevalence and Determinants." *Economic Inquiry* 46 (1):84–90.

———. 2009. "Homo Reciprocans: Survey Evidence on Behavioral Outcomes." *Economic Journal* 119 (536):592–612.

———. 2010. "Are Risk Aversion and Impatience Related to Cognitive Ability?" *American Economic Review* 100 (3):1238–1260.

———. 2011a. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3):522–550.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2012. "The Intergenerational Transmission of Risk and Trust Attitudes." *Review of Economic Studies* 79 (2):645–677.

Dohmen, Thomas, David Huffman, Uwe Sunde, JÃ¼rgen Schupp, and Gert G. Wagner. 2011b. "Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences." *Journal of the European Economic Association* 9 (3):522–550.

Dolan, Paul, Tessa Peasgood, and Methew White. 2008. "Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being." *Journal of Economic Psychology* 29 (1):94–122.

Doppelhofer, Gernot and Melvyn Weeks. 2009. "Jointness of Growth Determinants." *Journal of Applied Econometrics* 24:209–244.

Dunn, Douglas M. and Lloyd M. Dunn. 2007. *Peabody Picture Vocabulary Test, Fourth Edition, Manual*. Minneapolis, MN: NCS Pearson, Inc.

Durlauf, S. N, A. Kourtellos, and C. M. Tan. 2008. "Are any growth theories robust?" *The Economic Journal* 118:329–346.

Eckel, C.C. and P.J. Grossmann. 1996. "Altruism in Anonymous Dictator Games." *Games and Economic Behavior* 16 (2):181–191.

———. 2002. "Sex Differences and Statistical Stereotyping in Attitudes Toward Financial Risk." *Evolution and Human Behavior* 23 (4):281–295.

Eicher, Theo S., Chris Papageorgiou, and Adrian E. Raftery. 2011. "Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants." *Journal of Applied Econometrics* 26 (1):30–55.

Falk, A., T. Dohmen, D. Huffman, U. Sunde, and A. Becker. 2011. "A Validated Preference Module." *mimeo* .

Falk, A., E. Fehr, and U. Fischbacher. 2005. "Driving Forces behind Informal Sanctions." *Econometrica* 73 (6):2017–2030.

Falk, A. and U. Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior* 54 (2):293–315.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2012. "An Experimentally Validated Preference Survey Module." *University of Bonn: mimeo* .

Fehr, E. 2009. "On the Economics and Biology of Trust." *Journal of the European Economic Association* 7 (2–3):235–266.

Fehr, Ernst, Helen Bernhard, and Bettina Rockenbach. 2008. "Egalitarianism in Young Children." *Nature* 454:1079–1084.

Fehr, Ernst, Daniela Rützler, and Matthias Sutter. 2011. "The Development of Egalitarianism, Altruism, Spite and Parochialism in Childhood and Adolescence." *IZA Discussion Papers 5530, Institute for the Study of Labor (IZA), forthcoming in the AER* .

Fehr, Ernst and Jean-Robert Tyran. 2001. "Does Money Illusion Matter?" *The American Economic Review* 91 (5):1239–1262.

Feldkircher, Martin and Stefan Zeugner. 2009. "Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging." Discussion Paper No. 202, IMF.

Fernandez, Carmen, Eduardo Ley, and Mark F.J. Steel. 2001. "Model uncertainty in cross-country growth regressions." *Journal of Applied Econometrics* 16 (5):563–576.

Ferrer-i Carbonell, Ada. 2005. "Income and Well-Being: An Empirical Analysis of the Comparison Income Effect." *Journal of Public Economics* 89 (5–6):997–1019.

Ferrer-i-Carbonell, Ada and Paul Frijters. 2004. "How important is methodology for the estimates of the determinants of happiness?" *The Economic Journal* 114:641–659.

Finkelstein, A. 2009. "E-ZTax: Tax Salience and Tax Rates." *The Quarterly Journal of Economics* 124 (3):969—1010.

Fischbacher, U. 2007. "zTree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10 (2):171–178.

Fisher, Irvine. 1928. *The Money Illusion*. New York: Adelphi.

Fliessbach, K., B. Weber, P. Trautner, T. Dohmen, U. Sunde, C. E. Elger, and A. Falk. 2007. "Social Comparison Affects Reward-Related Brain Activity in the Human Ventral Striatum." *Science* 318:1305—1308.

Frederick, S., G.F. Loewenstein, and T. O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature* 40 (2):351–401.

Frey, Bruno S. and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40 (2):402–435.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1).

Frijters, Paul, John P. Haisken-DeNew, and Michael A. Shields. 2004. "Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification." *American Economic Review* 94 (3):730–740.

Gerlitz, J.-Y. and J. Schupp. 2005. "Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP." *DIW Berlin, Research Notes* .

Goeman, Jelle J., Sara A. van de Geer, and Hans C. van Houwelingen. 2006. "Testing against a High Dimensional Alternative." *Journal of the Royal Statistical Society. Series B* 68 (3):477–493.

Gollier, C. 2001. *The Economics of Risk and Time*. Cambridge, MA: MIT Press.

Golsteyn, Bart H.H., Hans Grönqvist, and Lena Lindahl. 2013. "Time Preferences and Life Outcomes." *IZA DP No. 7165* .

Graybill, Franklin A. 1983. *Matrices with Applications in Statistics*. Belmont, CA: Wadsworth.

Greiner, B. 2004. "An online recruitment system for economic experiments." *Forschung und wissenschaftliches Rechnen* 63:79–93.

Grier, Kevin B. and Gordon Tullock. 1989. "An empirical analysis of cross-national economic growth, 1951-1980." *Journal of Monetary Economics* 24 (2):259–276.

Hackman, Daniel A. and Martha J. Farah. 2009. "Socioeconomic Status and the Developing Brain." *Treds in Cognitive Sciences* 13:65–73.

Hanck, Christoph. 2009. "For Which Countries did PPP hold? A Multiple Testing Approach." *Empirical Economics* 37:93–103.

Hanushek, Eric A. and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46 (3):607–668.

Heckman, James and Edward Vytlacil. 2001. "Identifying the Role of Cognitive Ability in Explaining the Level of and Change in the Return to Schooling." *Review of Economics and Statistics* 83 (1):1–12.

Heckman, James J. 2008. "Schools, Skills, and Synapses." *Economic Inquiry* 46 (3):289–324.

Hendry, D. H. and H.-M. Krolzig. 2004. "We Ran One Regression." *Oxford Bulletin of Economics and Statistics* 66 (5):799–810.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (1):65–70.

Hoover, Kevin D. and Stephen J. Perez. 1999. "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search." *Econometrics Journal* 2:167–191.

———. 2004. "Truth and Robustness in Cross-Country Growth Regressions." *Oxford Bulletin of Economics and Statistics* 66 (5):765–798.

Hopkins, Kenneth D. and Glenn H. Bracht. 1975. "Ten-Year Stability of Verbal and Nonverbal IQ Scores." *American Educational Research Journal* 12:469–477.

Horn, John L. and Raymond Bernard Cattell. 1967. "Age Differences in Fluid and Crystallized Intelligence." *Acta Psychologica* 26:107–129.

Huang, Jian, Joel L. Horowitz, and Fengrong Wei. 2010. "Variable Selection in Nonparametric Additive Models." *The Annals of Statistics* 38 (4):2282–2313.

Jeffreys, H. 1961. *Theory of Probability*. Oxford, UK: Oxford University Press.

Jensen, Peter S. and Allan H. Würtz. 2012. "Estimating the Effect of a Variable in a High-Dimensional Linear Model." *Econometrics Journal* 15 (2):325–357.

Jensen, Peter Sandholt. 2010. "Testing the null of a low dimensional growth model." *Empirical Economics* 38:193–215.

Kawka, Rupert, Sabine Beisswenger, Gabriele Costa, Heike Kemmerling, Sarah Müller, Thomas Pütz, Helena Schmidt, Stefan Schmidt, and Marisa Trimborn. 2009. *Regionaler Preisindex, Berichte Band 30*. Bundesinstitut für Bau-, Stadt- und Raumforschung.

Kormendi, Roger C. and Philip G. Meguire. 1985. "Macroeconomic determinants of growth: Cross-country evidence." *Journal of Monetary Economics* 16 (2):141–163.

Kosse, Fabian and Friedhelm Pfeiffer. 2012. "Impatience among Preschool Children and their Mothers." *Economics Letters* 115:493–495.

Köszegi, B. 2006. "Ego Utility, Overconfidence, and Task Choice." *Journal of the European Economic Association* 4 (4):673–707.

Krolzig, Hans-Martin and David F. Hendry. 2001. "Computer Automation of General-to-Specific Model Selection Procedures." *Journal of Economic Dynamics and Control* 25:831–836.

Leeb, Hannes and Benedikt Pötscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21:21–59.

Lehmann, E. L. and Joseph P. Romano. 2005. "Generalizations of the familywise error rate." *Annals of Statistics* 33:1138–1154.

Leng, C., Y. Lin, and G. Wahba. 2006. "A Note on the Lasso and Related Procedures in Model Selection." *Statistica Sinica* 16:1273–1284.

Levine, R. and D. Renelt. 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review* 82:942–963.

Ley, Eduardo and Mark F.J. Steel. 2009. "On the Effect of Prior Assumptions in BAyesian Model Averaging with Applications to Growth Regression." *Journal of Applied Econometrics* 24:651–674.

Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and Jim O. Berger. 2008. "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association* 103 (481):410–423.

Lovell, Michael C. 1983. "Data Mining." *The Review of Economics and Statistics* 65 (1):1–12.

Luttmer, Erzo F. P. 2005. "Neighbors as negatives: Relative earnings and well-being." *The Quarterly Journal of Economics* 120 (3):963–1002.

MacKinnon, James G. and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29:305–325.

Magnus, Jan R., Owen Powell, and Patricia Prüfer. 2010. "A Comparison of Two Model Averaging Techniques with an Application to Growth Empirics." *Journal of Econometrics* 154:139–153.

Mankiw, N. Gregory, David Romer, and David N. Weil. 1992. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107 (2):407–437.

Manski, Charles F. 2002. "Identification of decision rules in experiments on simple games of proposal and response." *European Economic Reviewy* 46:880–891.

Masanjala, W. and C. Papageorgiou. 2006. "Initial conditions, European colonialism and Africas growth." Discussion Paper No. 1, Department of Economics, Louisiana State University.

Mischel, Walter and Ralph Metzner. 1962. "Preference for Delayed Reward as a Function of Age, Intelligence, and Length of Delay Interval." *The Journal of Abnormal and Social Psychology* 64 (6):425–431.

Mischel, Walter, Yuichi Shoda, and Philip K. Peake. 1988. "The Natures of Adolescent Competencies Predicted by Preschool Delay of Gratification." *Journal of Personality and Social Psychology* 54:687–696.

Mischel, Walter, Yuichi Shoda, and Monica L. Rodriguez. 1989. "Delay of Gratification in Children." *Science* 244 (4907):933–938.

Mittelhammer, Ron C., George G. Judge, and Douglas J. Miller. 2000. *Econometric Foundations*. Cambridge, UK: Cambridge University Press.

Moon, Hyungsik R. and Benoit Perron. 2012. "Beyond Panel Unit Root Tests: Using Multiple Testing to Determine the Nonstationarity Properties of Individual Series in a Panel." *Journal of Econometrics* 169:29–33.

Moreira, Bruno, Raul Matsushitab, and Sergio Da Silva. 2010. "Risk Seeking Behavior of Preschool Children in a Gambling Task." *Journal of Economic Psychology* 31 (5).

Neff, Walter S. 1938. "Socioeconomic Status and Intelligence: A Critical Survey." *Psychological Bulletin* 35:727–757.

Oswald, Andrew. 1997. "Happiness and economic performance." *The Economic Journal* 107:1815–1831.

———. 2008. "On the curvature of the reporting function from objective reality to subjective feelings." *Economic Letters* 100:369–372.

Paulsen, Platt, Huettel, and Brannon. 2011. "Decision-Making under Risk in Children, Adolescents, and Young Adults." *Frontiers in Psychology* 2.

Petermann, Franz and Ulrike Petermann. 2010. *HAWIK-IV. Hamburg-Wechsler-Intelligenztest für Kinder - IV*. Bern: Huber.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Rindermann, Heiner, Carmen Flores-Mendoza, and Marcela Mansur-Alves. 2010. "Reciprocal Effects between Fluid and Crystallized Intelligence and their Dependence on Parents' Socioeconomic Status and Education." *Learning and Individual Differences* 20:544–548.

Roberts, Brent W. 2006. *Personality Development and Organizational Behavior*. In B. M. Staw (Ed.). Research on Organizational Behavior. Elsevier Science/JAI Press, 1–41.

———. 2009. "Back to the Future: Personality and Assessment and Personality Development." *Journal of Research in Personality* 43:137–145.

Romano, Joseph P. and Azeem M. Shaikh. 2006. "On stepdown control of the false discovery proportion." *IMS Lecture Notes - Monograph Series 2nd Lehmann Symposium - Optimality* 49:33–50.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2008a. "Control of the false discovery rate under dependence using the bootstrap and subsampling." *Test* 17:417–442.

———. 2008b. "Formalized Data Snooping Based on Generalized Error Rates." *Econometric Theory* 24 (2):404–447.

Romano, Joseph P. and Michael Wolf. 2010. "Balanced Control of Generalized Error Rates." *The Annals of Statistics* 38 (1):598–633.

Roos, Michael W. M. 2006. "Regional Price Levels in Germany." *Applied Economics* 38:1553–1566.

Rotter, J. 1966. "Generalized expectancies for internal versus external control of reinforcement." *Psychological Monographs* 80 (1):1–28.

Sala-i-Martin, X. X., G. Doppelhofer, and R. Miller. 2004. "Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach." *American Economic Review* 94 (4):813–835.

Sala-i-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87 (2):178–183.

Samuelson, P.A. 1937. "A Note on Measurement of Utility." *Review of Economic Studies* 4 (2):155–161.

Sarkar, Sanat K. 2006. "False Discovery and False Nondiscovery Rates in Single-Step Multiple Testing Procedures." *The Annals of Statistics* 34 (1):394–415.

Schmidt, Frank L. and John E. Hunter. 2004. "General Mental Ability in the World of Work: Occupational Attainment and Job Performance." *Journal of Personality and Social Psychology* 86 (1):162–73.

Schneider, Ulrike and Martin Wagner. 2012. "Catching Growth Determinants with the Adaptive Lasso." *German Economic Review* 13:71–85.

Schuerger, James M. and Anita C. Witt. 1989. "The Temporal Stability of Individually Tested Intelligence." *Journal of Clinical Psychology* 45:294–302.

Schupp, J. and G. G. Wagner. 2002. "Maintenance of and Innovation in Long-Term Panel Studies The Case of the German Socio-Economic Panel (GSOEP)." *Allgemeines Statistisches Archiv* 86 (2):163–175.

Selten, R. 1967. *Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Sauermann, H. (Ed.), Beiträge zur experimentellen Wirtschaftsforschung.* J.C.B. Mohr (Paul Siebeck), Tübingen.

Senik, Claudia. 2004. "When Information Dominates Comparison: Learning from Russian Subjective Panel Data." *Journal of Public Economics* 88 (9–10):2099–2123.

Shafir, Eldar, Peter Diamond, and Amos Tversky. 1997. "Money illsuion." *The Quarterly Journal Of Economics* 112 (2):341–374.

Shamosh, Noah A. and Jeremy R. Gray. 2008. "Delay Discounting and Intelligence: A Meta-Analysis." *Intelligence* 36:289–305.

Shoda, Yuichi, Walter Mischel, and Philip K. Peake. 1990. "Predicting Adolescent Cognitive and Self-Regulatory Competencies from Preschool Delay of Gratification: Identifying Diagnostic Conditions." *Developmental Psychology* 26 (6):978–986.

Slovic, Paul. 1966. "Risk-Taking in Children: Age and Sex Differences." *Child Development* 37 (1):169–176.

Stiglitz, J., A. Sen, J.-P. Fitoussi, Bina Agarwal, Kenneth J. Arrow, Anthony B. Atkinson, FranÃ§ois Bourguignon, Jean-Philippe Cotis, Angus S. Deaton, Kemal Dervis, Marc Fleurbaey, Nancy Folbre, Jean Gadrey, Enrico Giovannini, Roger Guesnerie, James J. Heckman, Geoffrey Heal, Claude Henry, Daniel Kahneman, Alan B. Krueger, Andrew J. Oswald, Robert D. Putnam, Nick Stern, Cass Sunstein, and Philippe Weil. 2009. *Report by the Comission on the Measurement of Economic Performance and Social Progress.* www.stiglitz-sen-fitoussi.fr.

Storey, J. D., J. E. Taylor, and D. Siegmund. 2004. "Strong control, conservative point estimation and siumltaneous conservative consistency of false discovery rates: A unified approach." *Journal of the Royal Statistical Society. Series B* 66:187–205.

Storey, John D. 2002. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society, Series B* 64 (3):479–498.

Stutzer, Alois. 2004. "The role of income aspirations in individual happiness." *Journal of Economic Behavior and Organization* 54:89–109.

Stutzer, Alois and Bruno S. Frey. 2010. "Recent Advances in the Economics of Individual Subjective Well-Being." *Social research: An International Quarterly* 77 (2):679–714.

Sutter, Matthias, Martin G. Kocher, Daniela Rützler, and Stefan T. Trautmann. 2010. "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior." *IZA Discussion Papers 5404, Institute for the Study of Labor (IZA), forthcoming in the AER* .

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

Turkheimer, Eric, Andreana Haley, Mary Waldron, Brian D'Onofrio, and Irving I. Gottesman. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14:623–628.

Vischer, T., T. Dohmen, A. Falk, D. Huffman, J. Schupp, U. Sunde, and G.G. Wagner. 2011. "Measuring Time Preferences in Representative Samples by Combining Experimental Elicitation and Survey Measures."
mimeo .

Wagner, Gert G., Joachim R. Frick, and Juergen Schupp. 2007. "The German Socio-Economic Panel Study (SOEP) - Evolution, Scope and Enhancements." *SOEPpapers 1, DIW Berlin, The German Socio-Economic Panel (SOEP)* .

Weber, Bernd, Antonio Rangel, Matthias Wibral, and Armin Falk. 2009. "The medial prefrontal cortex exhibits money illusion." *PNAS* 106:5025–5028.

White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5):1097–1126.

Wu, C. F. J. 1986. "Jackknife, bootstrap and other resampling methods in regression analysis." *Annals of Statistics* 14:1261–1295.

Zou, Hui. 2006. "The Adaptive Lasso and its Oracle Properties." *Journal of the American Statistical Association* 101 (476):1418–1429.

Zuckerman, M. 1994. *Behavioral Expressions and Biosocial Bases of Sensation Seeking.* New York, Cambridge University Press.