Department of Economics

On the Measurement of Welfare, Happiness and Inequality

Maja Rynko

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Florence, January 2012

EUROPEAN UNIVERSITY INSTITUTE
Department of Economics

On the Measurement of Welfare, Happiness and Inequality

Maja Rynko

Thesis submitted for assessment with a view to obtaining the degree of
Doctor of Economics of the European University Institute

Examining Board:

Professor Richard Spady, Johns Hopkins University (External Supervisor)
Professor Helmut Lütkepohl, European University Institute
Professor Stephen Pudney, University of Essex
Professor Ada Ferrer-i-Carbonell, Institut d'Anàlisi Econòmica

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

To those who are happy despite the econometrics predictions.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Acknowledgments

This thesis would not have been possible without the support of many people. I am especially grateful to my supervisor Richard Spady, for directing and inspiring my work. I would also like to thank my second advisor Helmut Luetkepohl, as well as those who gave me feedback on various occasions when I presented my work. I should include in this list, among others: Luigi Guiso, Alicia Perez-Alonso, Michal Markun, the participants of the "Well-being and Interpersonal Relations in the Economic Sphere" conference in Venice (especially Andrew Clark and Marina della Giusta), the participants of the "Warsaw International Economic Meeting" and of the "Well-Being and its Various Measurements" workshop at the European University Institute.

The datasets applied in the first and the second chapters of this thesis were kindly provided by CEPS-INSTEAD in Differdange, Luxembourg during my stay there as an IRISS Program visitor. I am very grateful for their assistance and hospitality. I thank also the European University Institute staff for being so helpful and sympathetic throughout my stay in Florence. The last, yet no less demanding, task of polishing the dissertation was conducted when I had already started my job at the Central Statistical Office in Warsaw. I very much appreciate the support I was given in the final stage of my work by my director Ireneusz Budzyński.

Finally, I am happy that the PhD program allowed me to make valuable friendships with many exceptional people. Without those people and my family, the time I spent at the European University Institute would not have left me with so many nice memories.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table of Contents

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

List of Figures

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

List of Tables

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Preface

This thesis addresses welfare measurement issues, with an emphasis on the measurement of happiness and inequality. It contributes to the economic literature in both methodological and empirical terms, with the empirical analysis employing the PACO/CHER, ECHP and GSS datasets.

Although human welfare is a multidimensional concept, a classical approach is to simply investigate the distribution of wealth and/or income. Our first chapter analyses income distribution in Poland, using comprehensive data from the year 2000. We use the concept of stochastic dominance to investigate the extent to which the income of certain subgroups (based largely on combinations of gender, education, and region) unambiguously exceeds that of others, and examine and formally assess hypotheses of stochastic dominance using recently developed statistical tests. The results of this approach are contrasted with simple scalar measures of inequality that are conventionally used. We find that males, the higher educated and those living in the urban areas are better off, while the regional dominance relationship are difficult to establish.

However, to a large extent human welfare draws on subjective feelings of happiness or similar subjective well-being concepts. While self-assessments of well-being can be elicited, the relation of such expressions to the underlying concept is intrinsically problematic. Consequently, in our second and third chapters we present a semiparametric framework that allows for the modeling of latent variables. This item response theory methodology is first applied to assess the differences in "happiness"across selected European states. A more detailed analysis suggests that the genesis of happiness is affected by relative social status; income is more important to high status individuals for example.

The third chapter concerns further challenges in happiness measurement in the presence of framing effects and/or differential item functioning ("DIF"). The impact of the ordering of questions on subjective well-being responses is studied under an extended item response theory model incorporating the DIF feature of the survey. Contrary to previous studies, the results indicate that individuals' happiness estimates are largely unbiased when the framing experiment is ignored. The methodology we develop allows for the assessment of framing and DIF effects and permits inter-subject comparison and analysis even when such effects are large.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

CHAPTER 1

Inequality in Poland at the Change of Millennium: The Stochastic Dominance Approach

Abstract

This paper investigates the income distribution in Poland in the year 2000 basing on the CHER / PACO data. The distribution of welfare is examined by means of stochastic dominance tests. The theory of the stochastic dominance concept, together with the suggested by literature testing methods are discussed. The empirical results relate to the gender, educational and spatial welfare differences in Poland. The obtained stochastic dominance orderings are compared among themselves using the *TPROB* measure. Moreover, the standard welfare measures are presented and contrasted to the main results of the paper.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

1.1 Introduction

The comparison of welfare across countries, regions, or some demographics-related groups has gained a lot of attention in both the methodological and empirical economic research. There exists a vast range of different approaches, usually drawing on some aggregate indices that provide a complete ordering of comparable populations.[1]

However, very often different indices introduce different rankings and the final conclusions may be very sensitive to the choice of the measure applied. This concern is expressed, for instance, in Davidson and Duclos (2000): "Since the influential work of Atkinson (1970), considerable effort has been devoted to making comparisons of welfare distributions more ethically robust, by making judgments only when all members of a wide class of inequality indices or social welfare functions lead to the same conclusion, rather than concentrating on some particular index."

An alternative to indices-based welfare comparison is the stochastic dominance analysis, which takes the whole distribution into account without imposing any information aggregation. If a stochastic dominance relation is established, it implies an unambiguous judgment on the investigated ordering. However, stochastic dominance tests may provide a partial order, i.e. the results may not allow for the comparison and ordering of distributions (in other words, it may be impossible to state any of the relations: $X \geq Y$ or $X \leq Y$).

The stochastic dominance concept is applied in many different fields. A rich literature of stochastic dominance relates to the theory of decision making under risk; the concept is also applied to queuing theory, reliability theory, statistical physics, epidemiology and insurance mathematics. It may also be employed for equilibrium pricing models, optimum choice of inputs in agriculture, analysis of the optimal firm's capital structure or for the impact investigation of a policy intervention or treatment effects on the outcome distributions of interest.

The application of stochastic dominance for measuring the inequality, poverty or for welfare comparison in general usually came after its implementation in decision making theory. However, there are some approaches developed in the field of welfare economics that were only later found to be equivalent to stochastic dominance relations.

Usually, stochastic dominance analysis associates individual welfare with income, as is implemented in this paper. The analysis of income using stochastic dominance concepts may provide answers to different questions, starting from analysis of the direction of changes

[1] A complete ordering is an ability to compare any two different distributions / random variables and unambiguously assign them the statements "A is greater / smaller than B", understood in a broad sense.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

over time (e.g. is the distribution of income more equal than it was in the past? Has social welfare increased? e.g. Anderson 2003), going through international comparisons (e.g. are developing countries characterized by greater inequality than advanced countries? e.g. Bishop et al. 1993, Anderson 2004) and ending with the country's policy issues (e.g. do taxes lead to greater equality? Are there differences in income distribution between different groups within society - regarding gender, age, education, nationality? e.g. Maasoumi and Heshmati 2000). Even when the respective curves cross, the results obtained may be still informative, especially if dominance holds over some subsets of special interest to the analysis (e.g. the area below the poverty line).

This paper summarizes the literature of the stochastic dominance in the context of welfare investigation. A significant emphasis is laid on the stochastic dominance tests that were calculated in R software, translating the available GAUSS codes written by Garry Barrett, Stephen G. Donald and Ken X. Zhu.[2] Apart from the theory, empirical results are also presented, where the Polish income distribution in year 2000 is deeply investigated. The next section provides the general description of the stochastic dominance concept and section 3 presents the stochastic dominance tests. Section 4 provides the descriptive statistics of the dataset analysed, followed by a section covering the stochastic dominance empirical results, including the estimates of the *TPROB* measure. Section 6 presents the alternative welfare measures. Finally, section 7 concludes.

1.2 Stochastic Dominance: the concept and its application to welfare analysis

Stochastic dominance concepts allow for the ordering of distributions, applying the "being larger"ideas to random variables. In the context of welfare analysis, the general assumption behind the stochastic dominance investigation states that the statistical cumulative distribution functions for income contain sufficient information for ranking social states. Stochastic dominance relations may be defined at any order, however higher order relations lack an economic interpretation and these orderings become weaker for higher orders.[3] In the context of welfare analysis, the interest does not go beyond the third order of stochastic dominance relations.

The formal conditions for first, second and third order stochastic dominance are stated in the following definition.

[2]The R codes may be obtained from the author on request.

[3]The weakest stochastic order is the Laplace transform order for which the limit of order $j \to \infty$ is taken.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Definition 1.1 *Let X and Y be two random variables with continuous and monotonic cumulative distribution functions, F and G respectively. U_1 denotes the class of functions u such that $u' \geq 0$; U_2 denotes the class of all functions in U_1 for which $u'' \leq 0$ and U_3 denotes a subset of U_2 for which $u''' \geq 0$. $q_y(p)$ and $q_x(p)$ are the quantile functions, defined e.g. for X as $F[X \leq q_x(p)] = p$. The stochastic dominance relations are defined in a weak sense. The conditions (a), (b) and (c) are equivalent in each case.*

SD1: *Y first order stochastic dominates X if and only if:*

(a) $G(t) \leq F(t)$ *for all t in the support of X and Y*

(b) $E[u(Y)] \geq E(u(X)]$ *for all $u \in U_1$*

(c) $q_y(p) \geq q_x(p)$ *for all* $0 \leq p \leq 1$

SD2*: Y second order stochastic dominates X if and only if:*

(a) $\int_{-\infty}^{t} G(z)dz \leq \int_{-\infty}^{t} F(z)dz$ *for all t in the support of X and Y*

(b) $E[u(Y)] \geq E(u(X)]$ *for all $u \in U_2$*

(c) $\int_{0}^{p} q_y(z)dz \geq \int_{0}^{p} q_x(z)dz$ *for all* $0 \leq p \leq 1$

SD3*: Y third order stochastic dominates X if and only if:*

(a) $\int_{-\infty}^{t}\int_{-\infty}^{v} G(z)dzdv \leq \int_{-\infty}^{t}\int_{-\infty}^{v} F(z)dzdv$ *for all t and v in the support of X and Y with the end-point condition* $\int_{-\infty}^{+\infty} [G(t) - F(t)]dt \leq 0$.

(b) $E[u(Y)] \geq E(u(X)]$ *for all $u \in U_3$*

(c) $\int_{0}^{p}\int_{0}^{t} q_y(z)dzdt \geq \int_{0}^{p}\int_{0}^{t} q_x(z)dzdt$ *for all* $0 \leq p \leq 1$, *with* $\int_{0}^{1} q_y(z)dz \geq \int_{0}^{p} q_x(z)dz$

The end point conditions in the statements *(b)* and *(c)* ensure that the expected value of Y is greater/equal to that of X. The definition provided here relates to the weak dominance relation. The strict dominance may be easily obtained by adding the statement "and holds with the strict inequality for some t". However, no statistical tests can possibly differentiate between weak and strict (strong) relations.

The important characteristic of the stochastic dominance definition is its nested structure: a stochastic dominance relationship of order j implies the stochastic dominance relation of

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

order $j+1$, with the inverse not necessarily true. Therefore, having established e.g. SD1, there is no need to test for SD2.

The most straightforward way of analysing the stochastic dominance relates to the *(a)* points in the above definitions, where the cumulative distribution functions are analyzed. The definitions given in *(c)* constitute the so-called p-approach to dominance, referring to the quantile functions. The *(b)* definitions rely on the social welfare function properties, seen as an aggregation of individual utilities. The classes of functions required for the sequential orders of stochastic dominance have their interpretations in terms of social welfare functions. Usually, societies express preferences towards more equitable distributions and higher (real) income values. These properties are called the "equity" and "efficiency" preferences. The empirical investigations usually firstly compare the degree of inequality within each distribution and then introduce the information on the mean incomes. This is often done by referring to some selected indices, whereas the stochastic dominance approach allows to account simultaneously for different characteristics of the distributions.

First order stochastic dominance of Y over X occurs when Y is more likely than X to take on large values. In other words, the distribution of income in population Y first order stochastically dominates population X if for any income level t the proportion of the population with income at or below t is lower in Y than in X. SD1 implies that for $n = 1, 3, 5, ...$ the relation $EX^n \leq EY^n$ holds (whenever the expectation exists), specifically implying the mean-dominance of population Y. If the analysis is restricted to the concern for poverty, the stochastic dominance relation may be considered only up to a selected poverty line, z, and interpreted in terms of the poverty line dominance. This is closely related to the headcount poverty measure, which in Y cannot exceed that in X regardless of the income cut-off, as long as it is below the level z.

SD1 corresponds to the requirement of function u being monotonically increasing in income with the social welfare function specification: $W(H) = \int u(z)dH(z)$, where H denotes the income distribution. The equivalent result of $W(Y) \geq W(X)$ states that social welfare is greater in population Y than X. First order stochastic dominance is often called a rank dominance, stochastic order, usual stochastic order or strong stochastic order and is a pure efficiency criterion.

Graphically, SD1 is equivalent to the situation where the $F_X(t)$ curve lies above $F_Y(t)$. For the quantile definition, this condition may be presented graphically by means of Q-Q plots, where quantiles for the investigated distributions are plotted against each other. If the Q-Q plot of F_X^{-1} against F_Y^{-1} lies below the 45^0 line, then Y first order stochastically dominates X. Analogically, the appropriate P-P plot should lie above the 45^0 line.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

First order stochastic dominance is a stochastic order that compares the size of random variables. The natural expansion of this approach occurs when the variability of random variables is included in the analysis, which leads to convex orders. This class of stochastic orders has a wide application in the decision making theory, where the variability is directly interpreted as the riskiness of an uncertain outcome. Second order stochastic dominance allows for the analysis of the size and the variability of the random variables simultaneously and in the literature is often referred to as increasing concave order.

The concavity assumption implies that the realizations of the random variables closer to the mean are associated with higher values of function u. It follows then that if the realizations of the random variables are less dispersed, then the expected value of $u(X)$ is higher than for more spread realizations. In other words, the dominated random variable X is here both "smaller"and more "variable". Therefore, second order stochastic dominance introduces the concern for inequality (or inequality aversion) directly into the social welfare function. Concave and increasing welfare functions register an increase in well-being (giving a higher weight to people with lower incomes) when there is an upward mobility in a community (i.e. when poor become richer). The concavity of the welfare function is sufficient to guarantee that the principle of transfers[4] holds and coincides with the requirement that the society favors more equitable income distributions.

Analogically, as in the SD1 case, the second order stochastic dominance relation may be analysed up to a selected poverty line z. The economic interpretation relates to the concept of poverty gap: for all poverty lines the average poverty gap in X is greater than that in Y up to the established poverty line z.[5]

Third order stochastic dominance goes further in capturing the preference for inequality reducing changes in the lower end of the distribution function ("transfer sensitivity"). The social welfare function exhibits increasing inequality aversion, that is the concern about inequality getting bigger when the general level of income increases. This is obtained by imposing the additional condition of the third derivative of u being positive. The third derivative measures the rate of change of curvature. The positive third derivative in the case of increasing concave functions means that the function becomes "less concave"for increasing values of X, that is the marginal slope is decreasing. The intuition behind this says that for smaller X, a unit increase is praised much higher than for greater X.

A summary of the welfare intuition behind the stochastic dominance relation may be illustrated by different types of plots. If the cdf of income is plotted, it gives the graphical

[4]The principle of transfers states that if a transfer d is made from a person with income y_1 to a person with lower income y_2 with the relation $y_2 \leq y_1 - d$, then the social welfare increases.

[5]An income-poverty gap is defined as a weighted sum of the income shortfalls of the poor.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

presentation of FSD and in the poverty context, this plot is called the poverty incidence curve. Each point at the poverty incidence curve gives the percentage of the population deemed poor if the respective coordinate on the horizontal axis is the poverty line. Plotting the area under the poverty incidence curve results in obtaining the poverty depth (or deficit) curve which corresponds to the SSD criterion. Each point of the poverty depth curve gives the aggregate poverty gap if, again, the respective coordinate on the horizontal axis is the poverty line. Finally, the plot of the area under the poverty depth curve is called the poverty severity curve and measures the squared poverty gap.[6]

Finally, the investigation of the stochastic dominance relation may go beyond the one-dimensional case. Instead of single random variables, the whole random vector and its relation to other random vectors may be of interest. The stochastic dominance concepts are also defined and analysed for a multivariate case and the economic literature provides empirical applications of multivariate welfare comparisons (e.g. McCaig and Yatchew (2007)).

1.3 Testing stochastic dominance

There are different approaches to statistically testing stochastic dominance relations and new testing methodologies regularly appear in the literature. The origin of the stochastic dominance tests may be found in McFadden (1989), where tests for first and second order stochastic dominance for independent samples were presented. However, the McFadden tests relied on the restrictions of equal numbers of observations across the tested samples, which may be seen as a serious limitation, especially in the case of welfare analysis.

Many stochastic dominance tests (including the one by McFadden) refer to Kolmogorov-Smirnov (*KS*) type tests, where the whole support of the distribution is taken into account. The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both the location and shape of the empirical cumulative distribution functions of the two samples. Thanks to the test characteristic of comparing all the points in the income range, the consistence of the KS tests can be guaranteed, together with satisfying the full set of restrictions implied by stochastic dominance. The critical values, or p-values, are usually obtained either through simulation methods or bootstrapping. Examples of the KS-type stochastic dominance tests are the *KS1*, *KS2*, *KSB1*, *KSB2* and *KSB3* tests described in the following subsection.

[6] A thorough discussion of the importance of the first, second and third order stochastic dominance relationship between income distributions for social welfare and poverty rankings of distributions may be found, among others, in Anderson (1996) and Davidson and Duclos (2000).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

An alternative approach relies on the multiple comparison tests where inference is based on the comparison of a fixed number of arbitrarily chosen distribution quantiles. Usually, the comparison points are chosen to be the deciles or quintiles of the empirical distributions. This method, although actually testing different hypotheses than the *KS*-type tests (that is, of a dominance at a limited number of points), is seen as a competitive one, with the computational easiness as its main advantage. However, test inconsistency is relatively more likely to appear in the multiple comparison framework and the tests may fail to examine all of the implications of stochastic dominance, being likely to lack power in some situations. The tests presented and applied in this study are the Wald test (W), maximal t-statistic test (MT), with its variant designed by Anderson, 1996 (MTA).

The literature provides a range of ways in which the hypotheses tested may be formulated. The KS-type tests, discussed below, set the null of weak dominance, whereas Anderson (1996) assumes the common underlying distributions under the null. There exist some other examples in the literature where the alternative hypothesis, *H1,* states strong dominance and the null its converse, which may result in a situation in which a distribution dominates another almost everywhere, yet the null is not rejected. In many cases a rejection of the null may be seen as an inconclusive outcome, since it fails to rank the two populations and often, no intuition concerning the cause of refection is provided. Also, in the absence of information on the power of the test, non-rejection of dominance may not enable one to accept dominance.

1.3.1 Barrett and Donald (2003) KS-type tests

Barrett and Donald (2003) presents five KS-type tests allowing for different sample sizes and investigation of dominance relationships at different orders. As signaled above, the whole support of the compared income distributions is taken into account. The objects being compared are multiple partial integrals of the underlying income distribution and are compared at all points in the income range. A variety of simulation and bootstrap methods are applied to estimate the asymptotic p-value.

The assumptions required for the *KS*-type tests are listed below.

Assumption 1 ***The KS-type tests:***

1. *F and G have common support on $[0, \overline{z}]$, where $\overline{z} < \infty$,*

2. *F and G are continuous on $[0, \overline{z}]$,*

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

3. *$\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^M$ are independent random samples from distributions with cdf's F and G respectively,*

4. *The sampling scheme is such that as $N, M \to \infty$, $\frac{N}{N+M} \to \lambda$ where $0 < \lambda < 1$.*

Assumption 1 gives a natural restriction for the income distribution zero lower bound. However, it may be set to any finite number. On the other hand, changes of the upper bound value are also possible. Setting $\bar{z}$ equal to a specific value from the income support results in the poverty comparison and allows for the testing of the stochastic dominance relationship up to an arbitrarily set poverty line level (truncated dominance method).

For notational reasons, following Barrett and Donald (2003), we introduce the integral operator $\Im_j(\cdot, G)$, integrating the function G to order $j-1$, which is stated formally as:

$$\Im_1(z, G) = G(z); \quad \Im_2(z, G) = \int_0^z \Im_1(t; G)dt; \quad \Im_3(z, G) = \int_0^z \Im_2(t; G)dt. \tag{1.1}$$

The integral operator may also be stated recursively: $\Im_j(z; G) = \int_0^z \Im_{j-1}(t; G)dt$, or generally expressed as:

$$\Im_j(z; G) = \frac{1}{(j-1)!} \int_0^z (z-t)^{j-1} dG(t), \tag{1.2}$$

which is implemented empirically as:

$$\Im_j(z; \widehat{G}_M) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{(j-1)!} \mathbf{1}(Y_i \leq z)(z - Y_i)^{j-1}. \tag{1.3}$$

The hypotheses tested are stated as:

H_0^j**:** $\Im_j(z, G) \leq \Im_j(z, F)$ for all $z \in [0, \bar{z}]$,

H_1^j**:** $\Im_j(z, G) > \Im_j(z, F)$ for some $z \in [0, \bar{z}]$,

That is, the null states the weak dominance of G over F, also including the case where the distributions are equal everywhere. The KS test statistic of stochastic dominance at order j is given by:

$$\widehat{S}_j = \left(\frac{NM}{N+M}\right)^{1/2} \sup_z \left[\Im_j(z; \widehat{G}_M) - \Im_j(z; \widehat{F}_N)\right], \tag{1.4}$$

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

where $\widehat{G}_M$ and $\widehat{F}_N$ denote the empirical distributions constructed as:

$$\hat{G}_M(z) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}(Y_i \leq z), \tag{1.5}$$

and for $\hat{F}_N$ analogically. The decision rule applied here is of the form:

$$\text{reject} \quad H_0^j \quad \text{if} \quad \widehat{S}_j > c_j \quad (\text{or if} \quad \widehat{p}_j < \alpha).$$

However, it only in the *SD1* case that an analytic asymptotic distribution of the test statistic exists and the critical value is then calculated as: $c_1(\alpha) = \sqrt{-0.5\log\alpha}$, or equivalently the p-value is given by: $\alpha = \exp(-2(\widehat{S}_1)^2)$. Since the test statistic distributions of higher order dominance depend on the underlying distributions F and G, it follows that the appropriate rejection regions do not have analytical solutions and simulation methods are required. The characterization of the limiting distributions uses the facts that $\sqrt{N}(\widehat{F}_N - F) \Longrightarrow \mathcal{B}_F \circ F$ and $\sqrt{M}(\widehat{G}_M - G) \Longrightarrow \mathcal{B}_G \circ G$, with $\mathcal{B}$ denoting the Brownian Bridge.

Two types of *KS* tests, *KS1* and *KS2,* introduce the simulated process $\mathcal{B}_F^* \circ \widehat{F}_N$ and $\mathcal{B}_G^* \circ \widehat{G}_M$, which evaluated for distribution F at a point z is implemented as:

$$\mathcal{B}_F^*(z;\hat{F}_N) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\mathbf{1}(X_i \leq z) - \hat{F}_N(z))U_i^F \tag{1.6}$$

with U_i^F denoting a sequence of i.i.d. $N(0,1)$ random variables that are independent of the samples. The simulated version of the Brownian Bridge corresponding to G is analogical. The simulated p-values are obtained respectively for *KS1* and *KS2* type tests according to the formulas:

$$\widehat{p}_j^F = P_U(\sup_z \Im_j(z;\mathcal{B}_F^* \circ \widehat{F}_N) > \widehat{S}_j) \tag{1.7}$$

$$\widehat{p}_j^{F,G} = P_U(\sup_z(\sqrt{\widehat{\lambda}}\,\Im_j(z;\mathcal{B}_G^* \circ \widehat{G}_M) - \sqrt{1-\widehat{\lambda}}\,\Im_j(z;\mathcal{B}_F^* \circ \widehat{F}_N)) > \widehat{S}_j) \tag{1.8}$$

where $\hat{\lambda} = \frac{N}{N+M}$, N and M denote the respective sample sizes of F and G and $P_U(.)$, the latter being the probability associated with the normal random variables U_i^F and U_i^G and is conditional on the realized sample. The computational implementation of the p-value formulas firstly requires a calculation for each replication:

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

$$\bar{S}^F_{j,r} = \max_{t_k} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (\Im_j(t_k; 1_{X_i}) - \Im_j(t_k; \hat{F}_n)) U^F_{i,r} \tag{1.9}$$

$$\bar{S}^{F,G}_{j,r} = \max_{t_k} \sqrt{\frac{NM}{N+M}} \sum_{i=1}^{N} \left[(\Im_j(t_k; 1_{Y_i}) - \Im_j(t_k; \hat{G}_n)) U^G_{i,r} - (\Im_j(t_k; 1_{X_i}) - \Im_j(t_k; \hat{F}_n)) U^F_{i,r} \right] \tag{1.10}$$

with $r = 1, ..., R$, and R denoting the number of replications in the Monte Carlo simulation. Finally, the p-values of interest are calculated as:

$$\hat{p}^F_j \approx \frac{1}{R} \sum_{r=1}^{R} 1(\bar{S}^F_{j,r} > \hat{S}_j) \tag{1.11}$$

and analogically for $\bar{S}^{F,G}_{j,r}$.

The bootstrap methods, often applicable in more complicated situations when the p-value simulation may not be sufficient, are referred to by *KSB1*, *KSB2* and *KSB3* and described by the following formulas:

$$\bar{S}^F_j = \sqrt{N} \sup_z \left[\Im_j(z; \hat{F}^*_N) - \Im_j(z; \hat{F}_N) \right] \tag{1.12}$$

$$\bar{S}^{F,G}_{j,1} = \sqrt{\frac{NM}{N+M}} \sup_z \left[\Im_j(z; \hat{G}^*_M) - \Im_j(z; \hat{F}^*_N) \right] \tag{1.13}$$

$$\bar{S}^{F,G}_{j,2} = \sqrt{\frac{NM}{N+M}} \sup_z \left[(\Im_j(z; \hat{G}^*_M) - \Im_j(z; \hat{G}_M)) - (\Im_j(z; \hat{F}^*_N) - \Im_j(z; \hat{F}_N)) \right] \tag{1.14}$$

with $\hat{F}^*_N(z) = \frac{1}{N} \sum 1(X^*_i \leq z)$ for a random sample X^*_i drawn from $\aleph = \{X_1, ...X_N\}$ in the case of *KSB1*. For *KSB2* and *KSB3*, $\hat{F}^*_N(z)$ and $\hat{G}^*_N(z)$ are the empirical cdfs of random samples of sizes N and M respectively, drawn from $\Re = \{X_1, ...X_N, Y_1, ...Y_M\}$. The random variables corresponding to $\bar{S}^F_j$, $\bar{S}^{F,G}_{j,1}$, $\bar{S}^{F,G}_{j,2}$ are simulated and the probability that each exceeds the value of the statistic, given the appropriate sample, is approximated by the Monte Carlo simulation. A detailed description of this testing methodology and the associated proofs may be found in Barrett and Donald (2003).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

1.3.2 The Wald and MT tests

Davidson and Duclos (2000) describes and derives the asymptotic sampling distributions for a range of stochastic dominance tests based on the multiple comparison. This paper was not only mostly interested in estimating the thresholds up to which one population stochastically dominates another, at a given order,[7] but also in how the relation on the whole support of distributions may be investigated. The tests are designed to verify the hypothesis:

$H_0^j: \Delta_j(z_l) \leq 0$ for all $l \in \{1, ..., k\}$

$H_1^j: \Delta_j(z_l) > 0$ for some $l \in \{1, ..., k\}$

where j indicates the order of stochastic dominance being tested and l denotes the evaluation points (usually the quintiles or deciles of the pooled X and Y distribution), whereas $\Delta_j(z_l) = \Im_j(z_l; G) - \Im_j(z_l; F)$.

Davidson and Duclos (2000) basically considers two types of tests. The first is strongly related to the Wald test, with the test statistic defined as:

$$\hat{W}_j = min_{\Delta \in R_+^k} \left\{ (\hat{\Delta}_j - \Delta)' \hat{\Omega}_j^{-1} (\hat{\Delta}_j - \Delta) \right\} \tag{1.15}$$

where $\hat{\Omega}_j$ is the estimate of the variance-covariance matrix of $\hat{\Delta}_j$. The Wald statistic has an asymptotic distribution that is a mixture of chi-squared random variables. The critical values (or corresponding p-values) are usually simulated (which is implemented in the empirical analysis), unless k is sufficiently small.

The alternative approach is that of the Maximal t-statistic and refers to the t-statistics calculated for each $\Delta_j(z_l)$, which are tested to identify if they are equal to zero against the alternative that they are larger than zero. The individual t-statistic is given by:

$$\hat{t}_j(z_l) = \frac{\hat{\Delta}_j(z_l)}{\sqrt{\hat{\Omega}_{j,ll}}} \tag{1.16}$$

The test statistic is constructed as $\hat{S}_j^{MT} = \max_l \hat{t}_j^{z_l}$. A simplified test could be performed by rejecting the null if the largest t-statistic is large enough. This statistic has, however, a nonstandard distribution. The suggested procedure for simulating the p-value in the maximal t-statistic framework is:

[7] In the case of first order stochastic dominance, the estimated threshold is equivalent to the maximum common poverty line

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

$$\hat{p}_j^{MT} = \frac{1}{R}\sum_{s=1}^{R} \mathbf{1}(max\left\{\hat{\Gamma}_j^{1/2} Z_s\right\} > \hat{S}_j^{MT}) \tag{1.17}$$

where $\hat{\Gamma}_j^{1/2}$ is the Cholesky decomposition of a consistent estimate of Γ_j (the correlation matrix corresponding to Ω_j) and Z_s are random numbers drawn from multivariate standard normal distribution.

1.3.3 Maximal t-test by Anderson (1996)

The testing framework of Anderson (1996) is similar to the multiple comparison approach given above. There are, however, a few differences. Firstly, Anderson (1996) estimates the variance under the assumption that the $\Delta_1(z_l)$ are all zero. Secondly, the integrals defining $\Delta_j(z_l)$ for $j = 2, 3$ are approximated by a trapezoidal rule, whereas other presented methods are based on the direct integration of empirical results. The trapezoidal rule may produce inconsistent results.

The basic idea of the *MTA* is presented for the set of quintiles used as the evaluation points. The empirical example in the next sections involves the test evaluation for the deciles.

The rangespace of the two distributions is partitioned into five mutually exclusive and exhaustive intervals, with respective relative frequency vectors p_X and p_Y. In other words, if d_j is defined to be the jth interval length, then, with known F_X and F_Y, the probabilities of falling in the jth category would be given by:

$p_j = F(y^j) - F(y^{j-1})$, where $y^h = \sum_h^{i=1}$ and $F(y^0) = 0$

The SD2 and SD3 tests require evaluating the respective integrals as given in Definition 1. Using the trapezoidal rule of integrals approximation, we obtain the following formula for SD2:

$$C(y^j) = \int_0^{y^j} F(z)dz \approx 0.5\left[F(y^j)d_j + \sum_{i=1}^{j-1}(d_i + d_{i+1})F(y^i)\right] \tag{1.18}$$

The SD3 test involves the computation of:

$$\int_0^{y^j} C(z)dz \approx 0.5\left[C(y^j)d_j + \sum_{i=1}^{j-1}(d_i + d_{i+1})C(y^i)\right] \tag{1.19}$$

Defining the auxiliary matrices:

Defining the auxiliary matrices:

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

$$I_f = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and

$$I_F = 0.5 \begin{bmatrix} d_1 & 0 & 0 & 0 & 0 \\ d_1+d_2 & d_2 & 0 & 0 & 0 \\ d_1+d_2 & d_2+d_3 & d_3 & 0 & 0 \\ d_1+d_2 & d_2+d_3 & d_3+d_4 & d_4 & 0 \\ d_1+d_2 & d_2+d_3 & d_3+d_4 & d_4+d_5 & d_5 \end{bmatrix}$$

we can write the tested hypotheses as:

(SD1) H_0: $I_f(p^X - p^Y) = 0$ versus H_1: $I_f(p^X - p^Y) \leq 0$;

(SD2) H_0: $I_F I_f(p^X - p^Y) = 0$ versus H_1: $I_F I_f(p^X - p^Y) \leq 0$;

(SD3) H_0: $I_F I_F I_f(p^X - p^Y) = 0$ versus H_1: $I_F I_F I_f(p^X - p^Y) \leq 0$.

Note that the null states the non-dominance, i.e. it assumes the common underlying distribution. Under the null of a common population distribution and the assumption of independence of the samples, it is shown that the vector $v = p^X - p^Y$ is asymptotically distributed as $N(0, m\Omega)$, where $m = n^{-1}\frac{N+M}{NM}$ and the variance-covariance matrix is given by:

$$n^{-1}\Omega = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & -p_1p_4 & -p_1p_5 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & -p_2p_4 & -p_2p_5 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) & -p_3p_4 & -p_3p_5 \\ -p_4p_1 & -p_4p_2 & -p_4p_3 & p_4(1-p_4) & -p_4p_5 \\ -p_5p_1 & -p_5p_2 & -p_5p_3 & -p_5p_4 & p_5(1-p_5) \end{bmatrix}$$

where each p_i is the relative frequency corresponding to the hypothesized distribution. In practice, there is usually no pre-specified common null distribution and it is assumed that the each p corresponds to the relative frequency from the pooled sample. Stating the dominance of distribution Y over X requires that no element of the appropriate vector v be significantly greater than 0, whilst at least one is significantly less. The test is perfectly symmetric, so if the dominance of Y over X is not established, the relation $X \geq Y$ may be tested. However, the sequential testing may bring an indeterminate result, arising when $I_f(p^X - p^Y) \not\leq 0 \wedge \not\geq 0$

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

for SD1 and analogically for SD2 and SD3.

1.4 Data

The emphasis in this paper is laid on the empirical investigation of the income distribution in Poland in the year 2000. The data comes from the CHER / PACO database, which is administered by CEPS/INSTEAD, Luxembourg. The abbreviation PACO stands for the Panel Comparability project that aimed to construct an international comparative database integrating micro-data from various national household panels over a large number of years, both in Europe and America. The range of data for Poland covers the periods 1987-1990 and 1994-1996.

The CHER (Consortium of Household Panels for European Socio-Economic Research) was established as a feasibility study for developing and enhancing a comparative database of longitudinal household studies across countries. It can be thought of as the next step in development from PACO. It aims to harmonize and integrate micro datasets from a large variety of independent national panels and from the European Community Household Panel. The data available relates to, among other things, family structures, education, labour force participation, income distribution, poverty and problems of the elderly. There are seven National Panels as the main constitutive parts of the CHER micro database, including the GSOEP for Germany, BHPS for the United Kingdom, PSELL for Luxembourg, HBS for Poland, HHS for Hungary, PSBH for Belgium, SHP for Switzerland and PSID for the USA.[8] The files have been harmonized on major variables by the CEPS team together with national experts, producing a data file with records on different years and countries. The time coverage of the CHER dataset varies from country to country: for Germany the period 1990 to 2000 is included but data for Poland covers only the period 1994-2000. The last available year for the Polish dataset is analysed here.

In this study we look at individual net income from employment (including self-employment), obtained from the personal file in the dataset. This overcomes the problem of choosing the equivalence scale, which is necessary when the household's income is analysed. There are 3061 respondents considered in this study, born between 1941 and 1984, declaring themselves as normally working (more than 15 hours a week) and reporting positive income (but smaller than 100000).[9] Some descriptive statistics are presented in Table 1.1. The individuals work-

[8]The abbreviations stand for: German Socio-Economic Panel, British Household Panel Survey, Panel Socio-Economique Liewen zu Letzebuerg, Household Budgets Survey, Hungarian Household Survey, Panel Study on Belgian Households, Swiss Household Panel and Panel Study of Income Dynamics.

[9]After excluding from the sample the zero-income respondents, i.e. 382 people, the respondents with

ing part-time are included in the sample; they constitute a very small fraction and therefore should not bias the results significantly. Generally, part time work does not play a big role in the Polish labour market. Usually, in Western European countries these kinds of jobs are designed for women, allowing them to combine professional and family life. However, when we look at the gender division in Poland, around 73% of the respondents declaring part-time work are men. Nevertheless, considering the statistics on the number of hours actually worked in the last week, it is unambiguously the men who work more: the male average of 42.1 hours versus 36.6 hours for women.

Table 1.1. Sample descriptive statistics

Variable	**Frequency**	**Percentage**
Gender		
Male	1707	55.77
Female	1354	44.23
Education		
Primary education*	1734	56.65
Secondary education	1060	34.63
Third level education	267	8.72
Character of work		
Full-time	2947	96.28
Part-time	114	3.72
Urbanization		
Urban area	1619	52.89
Rural area	1442	47.11
	Mean	**Standard dev.**
Age	39.15	9.75
Hours worked**	39.67	13.62
Number of observations	3061	100

*Less than second stage of secondary education
**Number of hours actually worked last week

reported income above 100000 constitute around 1% of the sample. The values of income above the 100000 threshold are considered as outliers, which results in dropping an extra 54 people from the analysis.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

The sample is slightly overrepresentative of males. As far as the level of education is concerned, this variable appears in the further part of the analysis as a 2-category variable, which is due to merging the 2nd and 3rd categories, resulting in a "higher education" versus "lower education" comparison. This is necessary because of the very low frequency of respondents reporting a tertiary level of education[10] and the potential problems of significance for results. As far as the urban/rural area of residence is concerned, the inhabitants of rural areas are overrepresented, constituting almost 47% of the sample, whereas according to the Census 2002 data they accounted for approximately 36% of the population aged 20-59. However, in this study we are mostly interested in the distributions of income across differently specified groups and the relations between them. Therefore, no correction for the groups frequencies is applied.

Table 1.2 presents some descriptive statistics referring to individual incomes, weighting the observations by the personal weights included in the database, i.e., attempting to reveal the population (and not the sample) relations.

From these results, we can already make conclusions regarding mean-variance dominance, which requires the dominant random vector (random variable) to have a mean not smaller and variance not greater than that of the dominated one. This relation is observed here for education: the distribution of income for "higher education" individuals is characterized by a higher mean and a lower variance than that of those who have not completed the secondary level of education. The same refers to the urban-rural division: those living in urban areas are better off under these criteria. If the stochastic dominance conclusions were to be drawn for this case, based only on three comparison points (10th percentile, median and 90th percentile), then the orderings established through the mean-variance dominance would already be confirmed at the first order of stochastic dominance. However, the results of the formal tests taking into account the whole support of the distribution are presented in the next section.

It is difficult to draw conclusions about the welfare ordering concerning gender and regional divisions. For instance, in the case of a male-female comparison, the income distribution of males is characterized by a higher mean, but at the same time by greater dispersion. However, the selected percentiles already suggest that first order stochastic dominance rela-

[10]This percentage might seem to be low in comparison to Western European averages, however, there are big differences across age cohorts. A clear increase in university graduates is observed after the fall of Communism. According to the data from the National Census 2002 (*http://www.stat.gov.pl*), age cohorts of individuals above 40 years are characterized by percentages of individuals with higher levels of education not exceeding 9%. Naturally, it is also very low in the first two age cohorts (up to 24-year-old individuals). However, already 15.77% of Poles from the group aged 25-30 and 12.50% from the group 30-34 are classified as graduates in 2002.

tionship may appear.

Table 1.2. Summary statistics of income variable

	Mean	Sd. dev.	10th per.	Median	90th per.
Gender					
Male	15102.11	11699.62	5908.56	12080.74	26652.19
Female	10860.75	8080.61	4429.49	8953.85	18335.50
Education					
Lower	12360.88	10406.25	4109.03	9820.72	22203.59
Higher	14584.80	10371.68	6396.12	12021.44	24330.16
Urbanization					
Urban	13604.55	9218.64	6146.42	11330.59	23418.67
Rural	12946.07	11762.90	3948.72	9812.26	23309.51
East-west					
West	12936.92	9903.34	5603.55	10664.41	22324.56
East	13744.53	11253.80	4913.54	10762.39	25199.38
"Metro"-regions					
no	12991.30	9681.88	5408.24	10785.59	22272.24
yes	13594.37	11309.46	5017.26	10664.08	24655.63

The last two sections of the table refer to different regions, based either on the East-West division or if the region comprises a big city that could be classified as a "metropolis", driving up the economic development and the welfare of the region's inhabitants. A detailed discussion of the construction of these groups is given at a later stage in this paper. From the results presented in Table 1.2, it can be seen that the dominance relationships for the regional analysis are generally not clear at this stage.

1.5 Empirical results

The stochastic dominance tests above described are applied to investigate the relationships between differently defined subpopulations in Poland. Computationally, the tests were conducted in R. The p-values of the stochastic dominance tests for each subsection are presented in the tables in the Appendix of this chapter. In each table the header "Y versus X"should be understood as the hypothesis tested is "Y stochastically dominates X"(that is, the cdf of

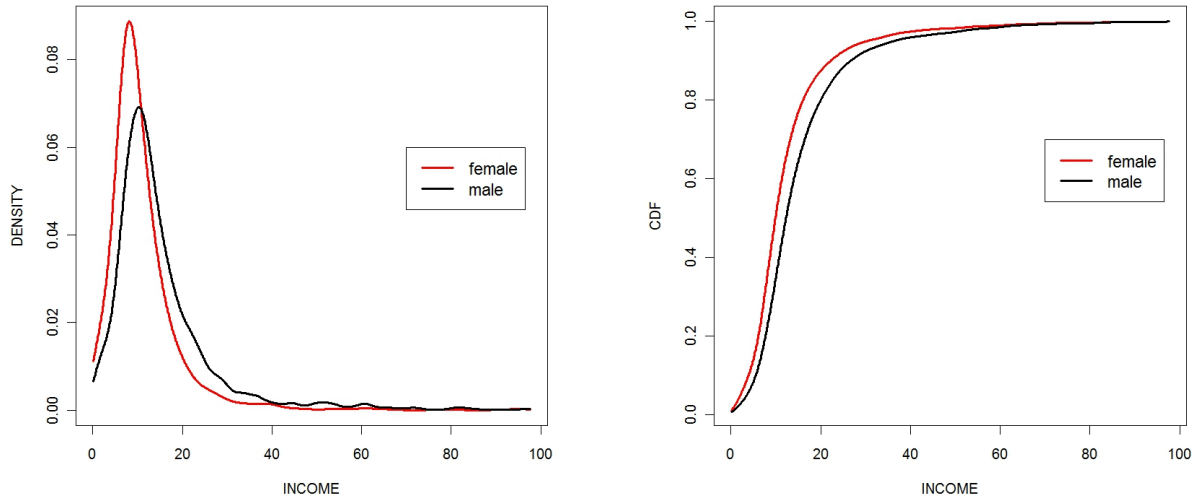
Y is less than or equal to cdf of X), apart from the *MTA* test, where Y versus X corresponds to H0 of distributional equality against H1 of stochastic dominance of X over Y.

1.5.1 Gender analysis

Under Communism, Polish women were encouraged to actively participate in the job market and the government created the conditions to reconcile family and work duties. High female participation rates continue to exist in post-Soviet countries, although the provision of state services has been reduced substantially. This, however, has not influenced labour market participation but the demographic processes instead, being reflected in decreasing fertility rates.

Nevertheless, despite high female labour market participation rates and even higher educational attainment than for men, there exist considerable differences in the distribution of earnings across genders. The plots below already reveal a first order stochastic dominance relationship, with the cdf for males lying unambiguously below the cdf of income for females.

Figure 1.1. Kernel density and cdf plots for men and women



Note: the plots were obtained using the R *np* package with gaussian kernel and kernel bandwidth set to 1.91 (for female) and 1.48 (male)

The eight different stochastic dominance tests (Table 1.6 in the Appendix) clearly state that we cannot reject the hypothesis of male net income distribution stochastically dominating female income distribution at the first order. Simultaneously, we reject the hypothesis of

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

any degree of stochastic dominance of female versus male net income distribution. Based on these results, it can be stated that the choice of any monotonic welfare function will confirm that men are better off than women in Poland with respect to income.

1.5.2 Education

The second analysis concerns the differences in earnings related to education. We expect that higher levels of education should increase the probability of a higher income. The summary statistics of educational attainment in Poland are presented in the table below.

Table 1.3. Education - frequencies for different subgroups

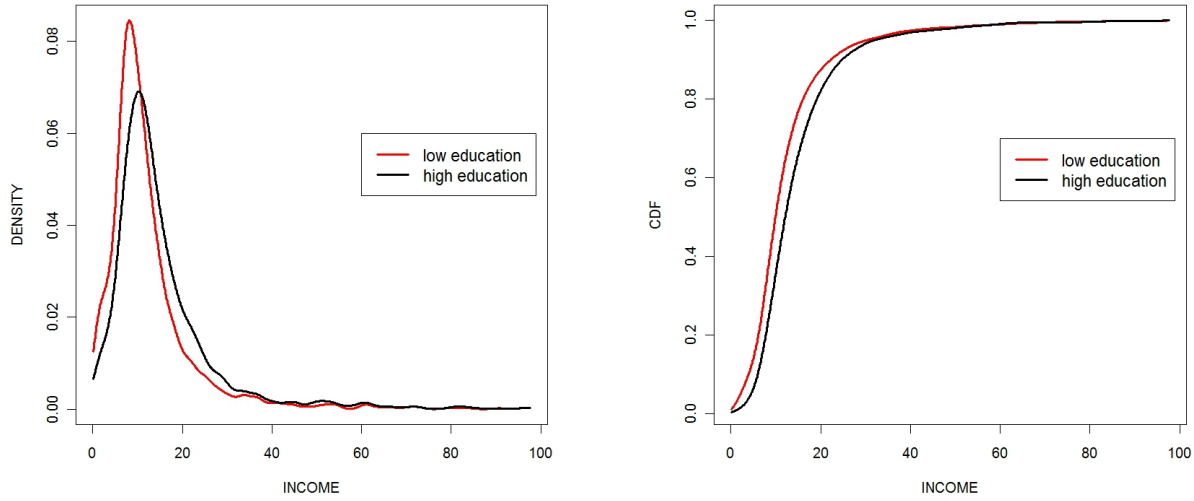
	1st stage		**2nd stage**		**3rd stage**		**Total**	
	Freq.	%	Freq.	%	Freq.	%	Freq.	%
All	1734	56.65%	1060	34.63%	267	8.72%	3061	100%
Gender								
Male	1158	67.84%	450	26.36%	99	5.80%	1707	55.77%
Female	576	42.54%	610	45.05%	168	12.41%	1354	44.23%
Urbanization								
Urban	739	45.65%	681	42.06%	199	12.29%	1619	52.89%
Rural	995	69.00%	379	26.28%	68	4.72%	1442	47.11%

Table 1.3 confirms that the difference between the male and female earnings cannot by explained by differences in education levels. On the contrary, a greater percentage of women have reached higher levels of education. The next subsection deals with the urbanizational differences presenting the results that urban areas are better off. In the latter case, the educational differences could constitute the reason for existing inequalities.

As the frequencies in the "3rd stage of education"row are very low, the second and third category were merged to ensure the computational stability. Therefore, the stochastic dominance relationship is concerned only for the group of people possessing less than the second stage of secondary education (called hereafter low education) versus those with at least finished secondary education (high education).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 1.2. Kernel density and cdf plots for different levels of education



Note: the plots were obtained using the R *np* package with gaussian kernel and kernel bandwidth set to 1.23 (for low education) and 1.60 (high education)

The cdf plot and the stochastic dominance results (Table 1.7 in the Appendix) again state unambiguously that the net income distribution of those possessing high education stochastically dominates the low education cdf even at the first order. Naturally, this result is seen as positive.

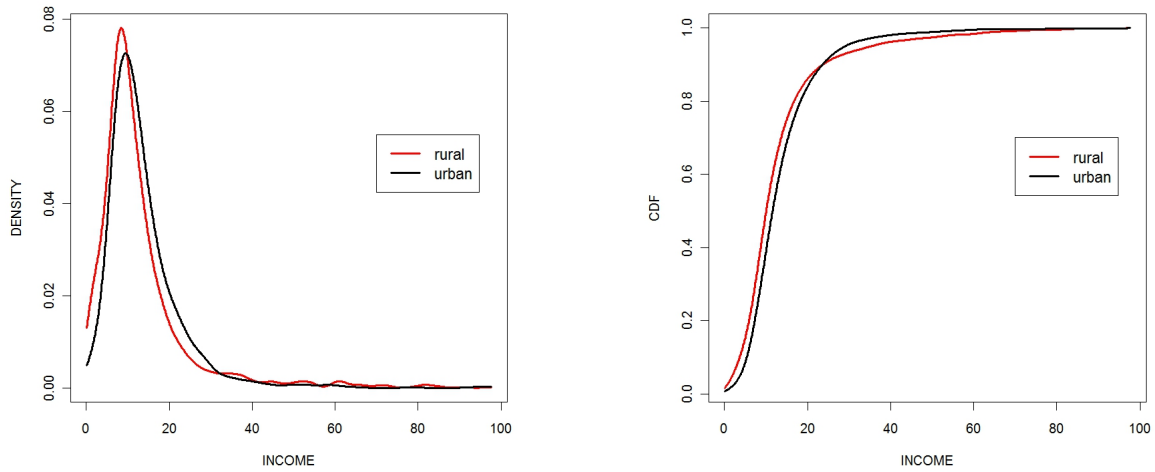
1.5.3 Urbanization

The Polish distribution of income is characterized by still considerable differences between the countryside and urban areas. We can conclude already from Table 1.1 of the higher standard of living in towns and cities. The reason for this, as already mentioned, could be the differences in educational levels: 70% of the respondents in rural areas are classified as "less than secondary level of education", in comparison to 46% in urban areas.

The graphical and formal testing results presented here may be confusing. The respective cdfs plotted in Figure 1.3 cross, which do not allow for conclusions to be made regarding first order stochastic dominance. The test results allow for the rejection of the hypothesis of rural over urban area dominance at any order, but the opposite hypothesis is not rejected, even at the first order.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 1.3. Kernel density and cdf plots for rural and urban areas



Note: the plots were obtained using the R *np* package with gaussian kernel and kernel bandwidth set to 1.56 (for rural areas) and 2.02 (urban areas)

Table 1.8 in the Appendix presents the p-values of the respective stochastic dominance tests for the rural-urban analysis. The p-values corresponding to the "urban SD1 rural"hypothesis amount to 0.330 for the KS-type tests. We could still reject this hypothesis, since Barrett and Donald (2003) considers in the general setting $\alpha < 0.5$ and the obtained level of α would indicate a 1/3rd probability that we reject a true hypothesis. Simultaneously, we could also state that the test fails to reject the null, thus assuming the first order stochastic dominance relation, and the observed fact of cdfs crossing treated as not statistically significant.

1.5.4 Regional analysis

Regional differences in economic development are very visible in Poland. There still exist disparities between the areas belonging to different occupants in 19th century and then, during the time of independence, between so called Poland A and B. There are also highly industrialized regions that constituted the engine of growth under the Communist regime, but some of them have suffered seriously from adjustment process after the restructuring began.

The most common division concerning the differences between the level of welfare based on the East-West line, with the East commonly regarded to be poorer and the West to have better infrastructure and profits from its closeness to Western Europe. In this paper, four

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

different clustering procedures are introduced, which attempts to identify the regions that are better and worse off. Table 1.4 presents some characteristics concerning different Polish regions (wojewodztwa), together with their assignment to differently defined clusters. The geographical illustration of the clustering exercise is presented in Figure 1.4.

The investigation of the East-West differences was conducted first. There are seven regions included in cluster "East", with 1332 respondents, whereas cluster "West"comprises 1729 individuals and 9 regions. East Poland is supposed to be more rural, which is supported by the data indicating that 54% respondents are classified as living in rural areas, as opposed to 41% in the "West"cluster.

The stochastic dominance tests indicate that at the 0.05 significance level, the hypothesis of *SD2* and *SD3* of West versus East can be rejected for *KS*-type tests. Simultaneously, we cannot reject East Poland regional dominance for the KS-type tests and unambiguously for dominance at the second order. We would expect that it is East Poland that should be dominated by West but we obtain the opposite result, which might be due to including rich regions Mazowieckie and Malopolskie in the East cluster.

The next attempt to construct some regions exhibiting dominance relations based on building the so called "metro(polis)-regions". They are defined as regions with big agglomerations, that included the following cities: Warsaw, Lodz, Krakow, Wroclaw, Poznan and Gdansk. This division provides a balanced distribution of respondents in the sample: 1519 of them are classified as inhabitants of the "metro-regions"versus 1542 otherwise. Also, the frequencies of urban and rural households are similar within this division. The formal results based on the *KS*-type tests indicate that "metro-regions"dominate those of "non-metro"in the 2nd and 3rd order, if the significance level is set as 0.1. However, with α set to 0.01 and 0.05 we cannot confirm any dominance relationships.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 1.4. Regional division in Poland

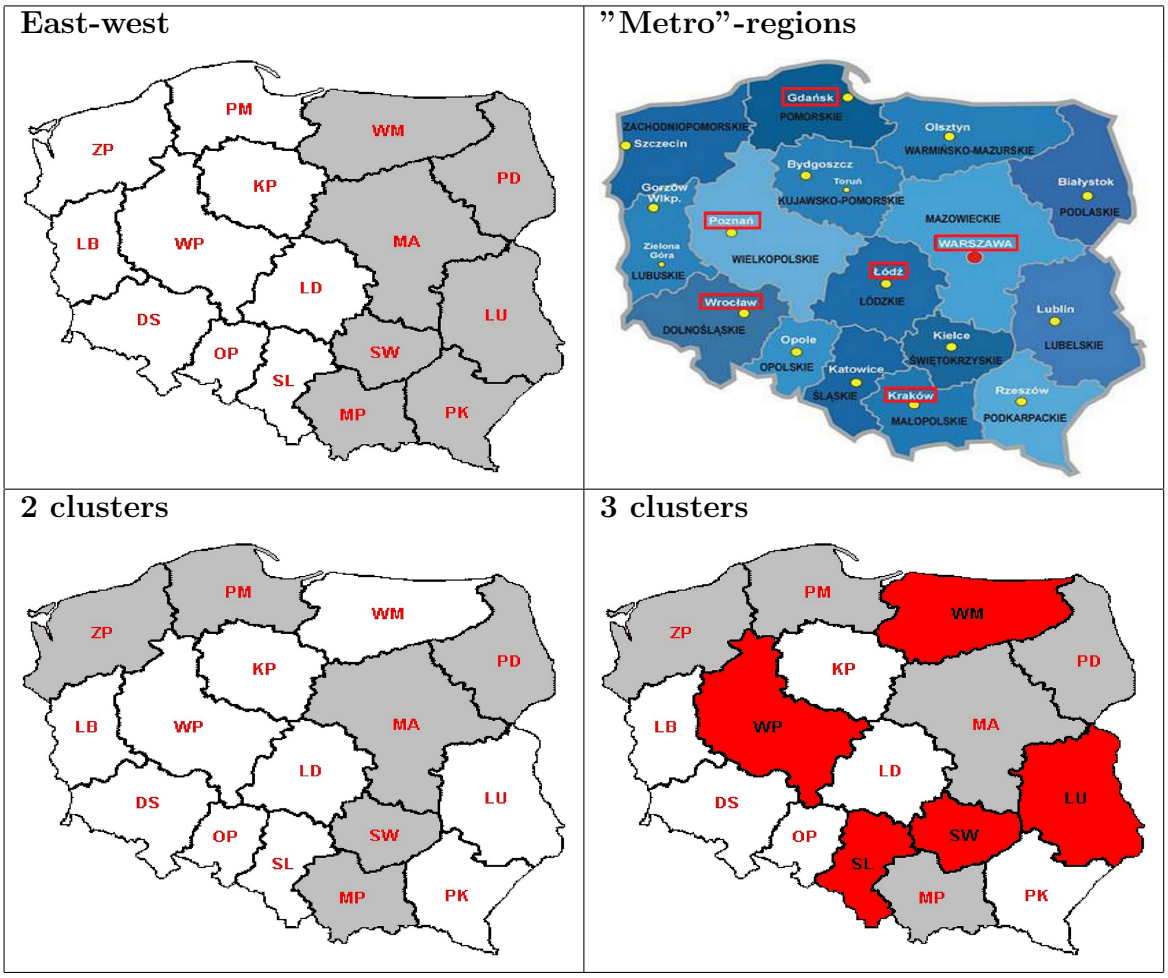
Region		Division				Sample data			Census 2002*	
Symbol	Name	East/West	Metro	2 cl.	3 cl.	Freq.	%	% rural	Freq.	%
DS	Dolnoślaskie	West	Yes	2	2	156	5.10	33.33	2 907 212	7.62
KP	Kujawsko - Pomorskie	West	No	2	2	149	4.87	36.24	2 069 321	5.43
LU	Lubelskie	East	No	2	3	189	6.17	72.49	2 199 054	5.77
LB	Lubuskie	West	No	2	2	101	3.30	39.60	1 008 954	2.65
LD	Łódzkie	West	Yes	2	2	271	8.85	50.55	2 612 890	6.85
MP	Małopolskie	East	Yes	1	1	240	7.84	45.00	3 232 408	8.47
MA	Mazowieckie	East	Yes	1	1	336	10.98	55.36	5 124 018	13.43
OP	Opolskie	West	No	2	2	70	2.29	55.71	1 065 043	2.79
PK	Podkarpackie	East	No	2	2	221	7.22	62.44	2 103 837	5.52
PD	Podlaskie	East	No	1	1	109	3.56	44.95	1 208 606	3.17
PM	Pomorskie	West	Yes	1	1	160	5.23	50.63	2 179 900	5.72
SL	Ślaskie	West	No	2	3	354	11.56	27.40	4 742 874	12.44
SW	Świetokrzyskie	East	No	1	3	99	3.23	65.66	1 208 606	3.17
WM	Warmińsko - Mazurskie	East	No	2	3	138	4.51	39.86	1 428 357	3.74
WP	Wielkopolskie	West	Yes	2	3	356	11.63	48.31	3 351 915	8.79
ZP	Zachodniopomorskie	West	No	1	1	112	3.66	28.57	1 698 214	4.45
No of observations						3061	100	-	38 141 209	100

*Population and Housing Census 2002, data refers to the population in total (without restricting to a specific age group neither to a selected employment status).

DOI: 10.2870/38699

The third clustering was based the results obtained from the *STATA kmeans* procedure, which assigns the regions to different clusters so that the Euclidean distance between them is the greatest with respect to the cluster means. Firstly two clusters were built, resulting in a rather unbalanced division of the respondents: 1056 versus 2005, with cluster 1 of smaller frequency (grey color on the map) representing the richest regions in Poland. The stochastic dominance tests clearly indicate that cluster 1 dominates cluster 2.

Figure 1.4. Different clustering of Polish territory



As the last step, 3 clusters were built again based on the *kmeans* procedure. The division provides balanced group frequencies: there are 957 respondents included in the first cluster (regions indicated in grey on the map), 968 in the second (in white) and 1136 in the third one (in red). All presented testing procedures state that cluster 1 dominates both cluster 2 and 3 with respect to all orders. Cluster 3 dominates cluster 2 according to *KS* and multiple comparison type tests at the 0.05 significance level.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Although for specific clustering choices we obtain the dominance ranking, these results do not indicate some compact regions in Poland that are characterized by higher, or respectively lower, welfare. The well-off clusters consists of "wojewodztwa"that are, in many cases, not neighbouring. Therefore, it may be difficult to explain the existing differences through geographical or historical factors.

1.5.5 *TPROB* analysis

The results presented in the sections above are usually in line with intuition and similar studies find consistent results. For instance, Szulc (2006) models the probability of falling below the poverty line based on the probit specification and shows that low education and being a farmer (as a main source of income) increases this probability. In Kot (1999), a rich set of results generally support the above findings.

The stochastic dominance analysis helps revealing some general results about the earnings discrepancies between selected subpopulations. However, the stochastic dominance test results do not inform if different types, e.g. the gender or regional, of income differences matter more. The p-values of the tests can provide only a vague idea about the distance between the two subpopulation distributions.

A distribution discrepancy measure was suggested by Gastwirth (1975), which compares the male and female earning distributions for various industries in the US. The summary measure suggested is the $TPROB$:

$$TPROB = 2\int_0^\infty [1 - F(x)]g(x)dx, \tag{1.20}$$

where $F(x)$ denotes the cumulative distribution function of the dominated group and $g(x)$ denotes the density function of the dominating group. When the two distributions are equal, $TPROB$ takes on value 1, i.e.:

$$\begin{aligned} TPROB_{eq} &= 2\int_0^\infty [1 - F(x)]f(x)dx = 2\int_0^\infty f(x)dx - 2\int_0^\infty F(x)f(x)dx \\ &= 2 - 2\int_0^\infty F(x)f(x)dx \end{aligned} \tag{1.21}$$

integrating by parts we obtain:

$$2\int_0^\infty F(x)f(x)dx = [F^2(x)]_0^\infty = \lim_{x\to\infty} F^2(x) - F^2(0) = 1;$$

from where clearly $TPROB_{eq} = 2 - 1 = 1$.

The *TPROB* measure is applied to compare all the pairs of income distributions investigated above. As far as the computational side is concerned, the R package *np* was used to estimate the respective cdfs and densities, with the kernel density bandwidths obtained in the maximum likelihood crossvalidation procedure included in this package. The integrals were approximated by the simple trapezoidal rule and a uniform grid. The calculated *TPROB* values are presented in Table 1.5, together with some additional income distribution measures discussed in the next section.

Table 1.5. Different distributions comparison measures

	Gini index	**Theil entropy**	**Coef. of variation**	**TPROB**
Gender				
Male	0.3527	0.2245	0.7672	
Female	0.3292	0.2031	0.7413	0.7155
Education				
High	0.3241	0.1901	0.7111	
Low	0.3726	0.2575	0.8419	0.7961
Urbanization				
Urban	0.3126	0.1743	0.6701	
Rural	0.3971	0.2925	0.9041	0.8567
"Metro"-regions				
Metro	0.3688	0.2480	0.8218	
Non-metro	0.3376	0.2087	0.7402	0.9846
East-west				
East	0.3717	0.2495	0.8138	
West	0.3384	0.2113	0.7559	0.9561
2 clusters				
Cluster 1	0.3713	0.2478	0.8153	
Cluster 2	0.3404	0.2134	0.7524	0.8865
3 clusters				
Cluster 1	0.3701	0.2462	0.8131	
Cluster 3	0.3426	0.2131	0.7477	0.9093 (vs c1)
Cluster 2	0.2181	0.3409	0.7676	0.8444 (vs c1)
				0.9061(vs c3)

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

The *TPROB* measure has a very intuitive interpretation giving, e.g. for the case of gender analysis, the probability that a randomly selected woman earns at least as much as a randomly selected man. The smaller this probability, the greater the discrepancies we observe between the subpopulations considered. Table 1.5 gives results that the *TPROB* measure has the smallest value when we analyse gender based subpopulations. The gap between the earnings distributions for differently educated respondents is also relatively big. *TPROB* measures for regional analysis are closer to 1, from which it may be concluded that the region of residence contributes to the existing income differences to a much smaller degree than gender and educational attainment.

1.6 Alternative welfare measures

Table 1.5 contains few welfare measures usually applied to income distribution analysis, which includes Gini, Theil entropy indices and the coefficients of variation. The ratios of medians and means, as well as some overlap measures could be applied to make conclusions concerning the inequality and efficiency characteristics of the income distributions. This section will present some basic features of alternative welfare measures and discuss their relation to the stochastic dominance approach.

One of the simplest measure of the observed wage differentials is the variance of the income (y) or its normalized version - the coefficient of variation expressed as: $C = \sqrt{Var(y)}/\mu$ (standard deviation divided by the mean). An alternative measure is the Theil entropy index that originates from the information theory and is given by:

$$T = (1/N)\sum_{i=1}^{N}(y_i/\mu)log(y_i/\mu) \tag{1.22}$$

Theil index takes on the value 0 if everyone in the society has the same income. If a situation of perfect inequality occurs, that is, one person has all the income, the Theil index equals $logN$. The normalization can be easily introduced to limit the range of the index to $(0, 1)$.

The Gini index, named after Italian sociologist Corrado Gini, is the most commonly known measure of inequality and is applied in many international rankings of country inequalities. Gini index is interpreted as the average absolute difference between the earnings of any two people in the population. Formally, it is expressed:

$$G = \frac{1}{2N^2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}|y_i - y_j| = 1 + \frac{1}{N} - \frac{2}{N^2\mu}\sum_{i=1}^{N}(N - i + 1)y_i \tag{1.23}$$

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

The higher the value of Gini, the more unequal income distribution is, with the maximum inequality value of the index equal 1. In a perfectly equal society the Gini index would be 0.

The Gini index is directly related to the concept of Lorenz curve that was introduced by Max Otto Lorenz, author of the seminal article "Methods of Measuring the Concentration of Wealth"published in 1905. Lorenz curve constitutes a visualization of wealth distribution in a given population, formally given by:

$$L_y(u) = \frac{1}{E(Y)} \int_0^u F_Y^{-1}(t)dt \text{ for } u \in [0,1] \tag{1.24}$$

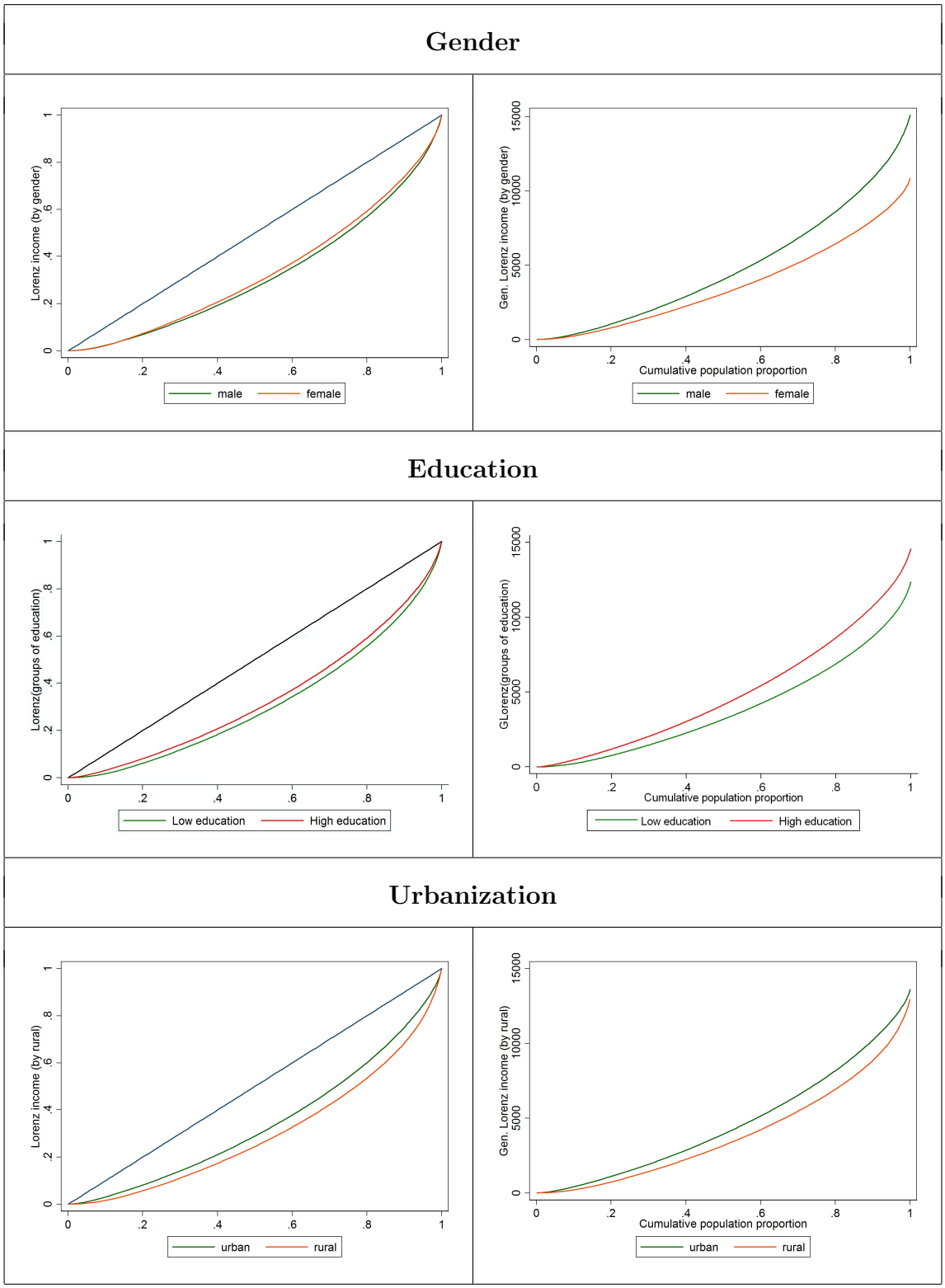
Thus, the Lorenz Curve plots the income distribution on the unit interval, showing for the bottom $a\%$ of the households the percentage of the total income they have. The illustration of perfect equality is a 45^0 line, whereas the estimated Lorenz curves lie below the diagonal. The area between the Lorenz curve and the perfect equality line (multiplied by 2 for the normalization) gives the Gini coefficient.

The Lorenz curves may also introduce an ordering between the distributions. If Y Lorenz dominates (LD) X, then the Lorenz curve corresponding to vector Y lies above that of vector X. That is, the Lorenz curve for country Y lies closer to the perfect equality line and therefore represents a more even distribution of wealth than X. Examples of Lorenz curves are given in the left panel of Figure 1.5. The Lorenz curve order is again only a partial order, whereas the Gini index introduces a complete order, even when the Lorenz curves cross.

The Coefficient of variation, Theil and Gini (together with the Lorenz order) introduce different ordering between the distributions considered in comparison to the stochastic dominance tests. For example, the stochastic dominance results show that male income distribution stochastically dominates female income distribution already at the first order. The results presented in Table 1.5 support the dominance of the female distribution. This results from the fact that female earnings are less disperse and thus are rank higher when only the preference for equality is taken into account. The indices presented do not capture the efficiency property, that is, they do not account for the social preference for higher incomes.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 1.5. Lorenz and Generalized Lorenz Curves for selected subpopulations



Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

The Lorenz curves comparison is meaningful when the assumption of mean income equality is fulfilled, which is obviously not the case here. When Lorenz curves are scaled up by the mean of the distribution, the plot obtained is called the Generalized Lorenz (GL) curve, expressed as:

$$GL_y(u) = E(Y)L_Y(u) = \int_0^u F_Y^{-1}(t)dt \; for \; u \in [0,1] \tag{1.25}$$

The GL concept is equivalent to second order stochastic dominance (directly comparable to the p-approach to SSD) incorporating both the preferences for equality and efficiency, whereas the Lorenz Order takes account of equality only. Generalized Lorenz curves are nondecreasing, continuous and convex functions with $GL_Y(0) = 0$ and $GL_Y(1) = E(Y)$. As the Gini index measures the area under the Lorenz curve, the Sen index could be computed with an analogous interpretation for the GL curves. The right panels of Figure 1.5 present the Generalized Lorenz curves for the selected subsamples and they correspond to the SSD tests results.

There exists a large variety of other possible indices measuring the general concept of welfare, or more specifically, the poverty or inequality in a society. However, many of them concentrate only on the equality issue, disregarding levels of income. Moreover, they often fail to reveal some details, especially in the tails of the distributions, not to mention the fact that the results provided may be contradictive when different indices are compared. If the income distribution plots cross, then it is always possible to find two different indices supporting two different hypotheses on the distributions ranking, enabling the manipulation of results. However, if the dominance relation is found using the stochastic dominance approach there is no need to search for any other indices.

1.7 Summary

This paper concentrates on the welfare comparison across differently defined subpopulations in Poland based on the data for the year 2000. The findings are consistent with other findings in the economic and sociology literature as well as policy documents, as concerns the conclusions on gender, educational and regional differences. This work, however, should not be considered as an exhaustive analysis of welfare distribution in Poland. A deep analysis should be conducted based on a richer dataset and using a richer set of methodological tools that are being developed. This includes new welfare measures, new statistical tests and methods of dealing with data imperfections. Nevertheless, this study constitutes a rich

framework for further research.

Many of the disputable issues were not addressed here. As far as statistical testing procedures are concerned, a deeper insight should be devoted to the discrepancies between the results for different stochastic dominance tests. From the empirical concerns, one may question the decision on the precise specification of the income variable, which is either objective or arbitrarily influenced by the data availability in many studies. The income variable may be regarded in terms of gross or net earnings, including market income, income from property or welfare benefits. Income may also be replaced by consumption, but both can still be measured at the individual level, for a household or family. This may involve the equivalence scale adjustment, while an international comparison uses the purchasing power parity correction and longitudinal analysis, an inflation adjustment.

Secondly, when the income variable is agreed on, the assumption of proxying the total welfare just by a financial dimension may be doubted. Certainly, the complex nature of a person's well-being is difficult to capture in the quantitative terms. The methodological answer includes the composite indicators research, like the Human Development Index (the UN), the Happy Planet Index (New Economics Foundation) or Gross National Happiness Index (developed in Bhutan). The stochastic dominance tools may also be extended to the multidimensional case and the joint distribution of income together with leisure, education, health or environmental related variables could be investigated. Usually, it is difficult to establish joint multivariate dominance when comparing several populations (countries). It was shown that even the marginal distributions dominance investigation often fails to point to the dominant country. McCaig and Yatchew (2007) compares Germany, the UK and the US in terms of income and leisure and they find that none of these countries dominates the other in both dimensions. The US dominates both the UK and Germany with respect to income but the Germans enjoy more leisure. Interesting results may also be obtained when poverty regions are analyzed in a multivariate way. The question whether the situation of low income households is worse than that of the richer households in other dimensions apart from income could provide interesting answers about individual choices. Nevertheless, dealing with higher dimensional distributions may be, in many cases, computationally intensive and prevent the attainment of significant statistical conclusions.

The stochastic dominance analysis could be extended to account for "Almost Stochastic Dominance", a concept described in the decision making under risk literature. It may occur that some extreme groups do not allow for the establishment of the dominance relation, e.g. the homeless or voluntary unemployed living in rich countries. It may also concern the other margin of society; the richest when the level of inequality is very high. Almost stochastic

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

dominance would correct for such occurrences.

Another drawback of the stochastic dominance analysis concerns its limit to analyze the distributional differences with respect to only one characteristic of the individuals. An analysis of welfare differences of gender, educational, regional and urban divisions could be conducted in a properly defined econometric framework. Additionally, such an approach could assess what part of the observed outcome (income) inequality may be attributed to the differences in circumstances and personal efforts.[11] This line of research is followed by Bourguignon et al. (2007), Lefranc et al. (2009), Breen and Jonsson (2005) and Checchi (2005) among others. The set of variables they define as circumstances, i.e. the factors independent of an individual's will, comprises race, place of birth and family background, whereas schooling and the job training act as effort proxies. Some of those proxies were analysed in the stochastic dominance framework but a regression-based approach allows for a simultaneous multivariate analysis. The inequality of opportunity and inequality of outcomes approach is closely related to the concepts of social mobility, inequality inheritance and the openness of society.

Certainly, the assessment of welfare should go beyond the analysis of the average income of the average person. The multidimensionality of the welfare concept and the importance of different factors shaping the whole income distribution should be addressed as well.

[11]There are also attempts to account for the double effect of circumstances on wages: the direct effect and through influencing the effort first.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

References

[1] Anderson, Gordon (1996): Nonparametric Tests of Stochastic Dominance in Income Distributions, *Econometrica*, Vol. 64 (5), pp. 1183-1193.

[2] Anderson, Gordon (2003): Poverty in America 1970-1990: Who did Gain Ground? An Application of Stochastic Dominance Criteria Employing Simultaneous Inequality Tests in a Partial Panel, *Journal of Applied Econometrics*, Vol. 18 (6), pp. 621-640.

[3] Anderson, Gordon (2004): Making Inferences about the Polarization, Welfare and Poverty of Nations: A Study of 101 Countries 1970-1995, *Journal of Applied Econometrics*, Vol. 19 (5), pp. 537-550.

[4] Atkinson, Anthony B. (1970): On the Measurement of Inequality, *Journal of Economic Theory*, Vol. 2 (3), pp. 244-263.

[5] Atkinson, Anthony B. (1987): On the Measurement of Poverty, *Econometrica*, Vol. 55 (4), pp. 749-764.

[6] Atkinson, Anthony B. (1992): Measuring Poverty and Differences in Family Composition, *Economica*, Vol. 59, No. 233, pp. 1-16.

[7] Banks, James and Johnson, Paul (1994): Equivalence Scale Relativities Revisited, *The Economic Journal*, Vol. 104, No. 425, pp. 883-980.

[8] Barrett, Garry F. and Donald, Stephen G. (2003): Consistent Tests for Stochastic Dominance, *Econometrica*, Vol. 71 (1), pp. 71-104.

[9] Bishop, John A., Formby, John P. and Smith James W. (1991): Lorenz Dominance and Welfare: Changes in the U.S. Distribution of Income, 1967-1986, *The Review of Economics and Statistics*, Vol. 73 (1), pp. 134-139.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[10] Bishop, John. A, Formby, John P., Smith, James W. (1993): International Comparisons of Welfare and Poverty: Dominance Orderings for Ten Countries, *The Canadian Journal of Economics*, Vo. 26 (3), pp. 707-726.

[11] Birch, Adrian, Haag, Antoine, Lefebure, Stijn, Villeret, Anne and Schmaus, Günther (2003): User guide, *CHER Working paper*, No. 2, CEPS/INSTEAD, Differdange, Luxembourg.

[12] Bourguignon, Francois and Ferreira, Francisco H. G. and Menedez, Marta (2007): Inequality Of Opportunity In Brazil, *Review of Income and Wealth*, Vol. 53 (4), pp. 585-618.

[13] Breen, Richard and Jonsson, Jan O. (2005): Inequality of Opportunity in Comparative Perspective: Recent Research on Educational Attainment and Social Mobility, *Annual Review of Sociology*, Vol. 31, pp. 223-43.

[14] Checchi, Daniele and Peragine Vitorocco (2005): Regional Disparities and Inequality of Opportunity: The Case of Italy, IZA Discussion Paper, No. 1874.

[15] Cowell, Frank A. (1995): Measuring Inequality, Second Edition, Prentice Hall/Harvester Wheatsheaf.

[16] Davidson, Russell and Duclos, Jean-Yves (2000): Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality, *Econometrica*, Vol. 68 (6), pp. 1435-1464.

[17] Davidson, Russell and Ducols, Jean-Yves (2006): Testing for Restricted Stochastic Dominance, *ECINEQ, Society for the Study of Economic Inequality*, Working Papers No. 36.

[18] Foster, James E. and Shorrocks, Anthony F. (1988): Poverty Orderings, *Econometrica*, Vol. 56 (1), pp. 173-177.

[19] Gastwirth, Joseph L. (1975): Measures of Earnings Differentials, *The American Statistician*, Vol 29 (1), pp. 32-35.

[20] Jenkins, Stephen (1991): The measurement of income inequality, Chapter 1 in L. Osberg (ed) Economic Inequality and Poverty: International Perspectives, pp. 3–38. M E Sharpe, Armonk NY.

[21] Kot, Stanisław M. (ed., 1999): Analiza ekonometryczna kształowania sie płac w Polsce w okresie transformacji, Warszawa, PWN, (Econometric Analysis of wages in Poland in the restructuring period, book in Polish).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[22] Krajowy Plan Działań na Rzecz Zatrudnienia na lata 2009-2011, Załacznik do Uchwały Rady Ministrów nr 111/2010, Warszawa 2010 (The Country's Plan for Employment in the years 2009-2011, An Attachment to the Council of Ministries Resolution, in Polish).

[23] Lefranc, Arnaud, Pistolesi, Nicolas and Trannoy, Alain (2008): Inequality of Opportunities vs. Inequlity of Outcomes: Are Western Societies all Alike?, *Review of Income and Wealth*, Vol. 54 (4), pages 513-546.

[24] Leshno, Moshe and Levy, Heim (2002): Preferred by "All"and Preferred by "Most"Decision Makers: Almost Stochastic Dominance, *Management Science*, Vol. 48 (8), pp. 1074-1085.

[25] Levy, Heim (1992): Stochastic Dominance and Expected Utility: Survey and Analysis, *Management Science*, Vol. 38 (4), pp. 555-593.

[26] Levy, Heim, Leshno, Moshe and Leibovitch, Boaz (2005): Expected Utility, Bounded Preferences and Paradoxes, presented at the CEMMAP symposium: Testing stochastic dominance restrictions, November 2005.

[27] Maasoumi Esfandiar, Heshmati Almas (2000): Stochastic Dominance Among Swedish Income Distributions, *Econometric Reviews*, Vol. 19 (3), pp. 287-320.

[28] McCaig Brian and Yatchew Adonis (2007): International Welfare Comparisons and Nonparametric Testing of Multivariate Stochastic Dominance, *Journal of Applied Econometrics*, Vol. 22 (5), pp. 951-969.

[29] McFadden, Daniel (1989): Testing for Stochastic Dominance, in *Studies in the Economics of Uncertainty: In Honor of Josef Hadar*, ed. by T.B. Fomby and T.K. Seo, New York, Berlin, London, and Tokyo: Springer.

[30] Mo-Yin, Tam S. and Zhang, Renze (1996): Ranking Income Distributions: The Tradeoff between Efficiency and Equality, *Economica*, Vol. 63, No. 250, pp. 239-252.

[31] Müller, Alfred and Stoyan, Dietrich (2002): Comparison Methods for Stochastic Models and Risks, John Wiley & Sons, Ltd.

[32] Nelson, Roger B. (2006): An Introduction to Copulas, Second Edition, Springer Science.

[33] Ogryczak, Włodzimierz and Ruszczyński, Andrzej (1999): From Stochastic Dominance to Mean - Risk Models: Semideviations as Risk Measures, *European Journal of Operational Research*, Vol. 116, pp.33-50.

[34] Polityka Równości Płci. Polska 2007. Raport, Warszawa, UNDP i Fundacja "Fundusz Współpracy"(Gender equality policy. Poland 2007. Report, in Polish).

[35] Post, Thierry and Versijp, Philippe (2005): Multivariate tests for stochastic dominance efficiency of a given portfolio, presented at the CEMMAP symposium: Testing stochastic dominance restrictions, November 2005.

[36] Shaked, Moshe and Shanthikumar, J. George (1994): Stochastic orders and their applications, Academic Press Inc.,U.S.

[37] Shorrocks, Anthony F. (1983): Ranking Income Distributions, *Economica*, Vol. 50, No. 197, pp. 3-17.

[38] Strategia Rozwoju Społeczno-Gospodarczego Polski Wschodniej do Roku 2020, Ministerstwo Rozwoju Regionalnego, Warszawa 2008 (The Socio-Economic Development Stategy for Eastern Poland untill 2020, Polish Ministry of Regional Development, Warsaw 2008).

[39] Szulc, Adam (2006): Poverty in Poland During the 1990s: Are the Results Robust, *Review of Income and Wealth*, vol. 52 (3), pp. 423-448.

[40] Thistle, Paul D. (1989), Ranking Distributions with Generalized Lorenz Curves, *Southern Economic Journal*, Vol. 56 (1), pp. 1-12.

[41] Warunki Życia Ludności Polski, GUS, Warszawa 2007 (Life Conditions of the Polish Population, Central Statistical Office, Warsaw 2007, in Polish).

[42] Yitzhaki, Shlomo (1982): Stochastic Dominance, Mean Variance, and Gini's Mean Difference, *The American Economic Review*, Vol. 72 (1), pp. 178-185.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix A: Tables

Table 1.6. Gender analysis

	men (Y) versus women (X)			woman (Y) versus men (X)		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.999	0.874	0.840	0.000	0.000	0.000
KS2	0.999	0.853	0.814	0.000	0.000	0.000
KSB1	0.999	0.898	0.863	0.000	0.000	0.000
KSB2	0.999	0.912	0.883	0.000	0.000	0.000
KSB3	0.999	0.907	0.877	0.000	0.000	0.000
MT(10)	1.000	1.000	1.000	0.000	0.000	0.000
W(10)	0.873	0.768	0.710	0.000	0.000	0.000
MTA(10)	1.000	1.000	1.000	0.000	0.000	0.000

Table 1.7. Education analysis

	high (Y) versus low (X)			low (Y) versus high (X)		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.9879	0.8520	0.7970	0.000	0.000	0.000
KS2	0.9879	0.8660	0.8160	0.000	0.000	0.000
KSB1	0.9879	0.8980	0.8650	0.000	0.000	0.000
KSB2	0.9879	0.9280	0.9080	0.000	0.000	0.000
KSB3	0.9879	0.9310	0.9010	0.000	0.000	0.000
MT(10)	1.0000	1.0000	1.0000	0.000	0.000	0.000
W(10)	0.8700	0.7720	0.7150	0.000	0.000	0.000
MTA(10)	1.0000	1.0000	1.0000	0.000	0.000	0.000

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 1.8. Urbanization analysis

	rural (Y) versus urban (X)			urban (Y) versus rural (X)		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.000	0.000	0.001	0.330	0.850	0.801
KS2	0.000	0.000	0.000	0.330	0.857	0.812
KSB1	0.000	0.000	0.001	0.330	0.899	0.871
KSB2	0.000	0.002	0.003	0.330	0.919	0.886
KSB3	0.000	0.000	0.000	0.330	0.943	0.915
MT(10)	0.000	0.000	0.000	0.998	1.000	1.000
W(10)	0.000	0.000	0.000	0.869	0.765	0.708
MTA(10)	0.000	0.000	0.000	0.915	0.999	1.000

Table 1.9. Regional analysis (1)

	east (Y) versus west (X)			west (Y) versus east (X)		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.252	0.522	0.559	0.093	0.028	0.037
KS2	0.252	0.543	0.579	0.093	0.028	0.036
KSB1	0.252	0.550	0.582	0.093	0.018	0.024
KSB2	0.252	0.538	0.563	0.093	0.019	0.029
KSB3	0.252	0.518	0.544	0.093	0.022	0.023
MT(10)	0.132	0.124	0.132	0.132	0.066	0.076
W(10)	0.153	0.126	0.129	0.110	0.067	0.069
MTA(10)	0.124	0.059	0.070	0.124	0.077	0.152

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 1.10. Regional analysis (2)

	metro (Y) versus non-metro (X)			**non-metro (Y) versus metro (X)**		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.458	0.547	0.563	0.304	0.069	0.099
KS2	0.458	0.562	0.572	0.304	0.050	0.074
KSB1	0.458	0.511	0.540	0.304	0.071	0.099
KSB2	0.458	0.513	0.530	0.304	0.045	0.078
KSB3	0.458	0.545	0.558	0.304	0.048	0.071
MT(10)	0.310	0.448	0.475	0.079	0.146	0.194
W(10)	0.288	0.420	0.439	0.091	0.145	0.190
MTA(10)	0.311	0.141	0.187	0.079	0.102	0.253

Table 1.11. Regional analysis (3)

	cluster 1 (Y) vs cluster 2 (X)			**cluster 2 (Y) vs cluster 1 (X)**		
	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.994	0.851	0.821	0.000	0.000	0.000
KS2	0.994	0.866	0.819	0.000	0.000	0.000
KSB1	0.994	0.921	0.889	0.000	0.000	0.000
KSB2	0.994	0.922	0.890	0.000	0.000	0.000
KSB3	0.994	0.930	0.900	0.000	0.000	0.000
MT(10)	0.946	0.965	0.955	0.000	0.000	0.000
W(10)	0.867	0.768	0.714	0.000	0.000	0.000
MTA(10)	0.949	0.835	0.805	0.000	0.000	0.000

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 1.12. Regional analysis (4)

	cl. 1 (Y) vs cl. 2 (X)			**cl. 2 (Y) vs cl. 1 (X)**			**cl. 1 (Y) vs cl. 3 (X)**		
	SD1	SD2	SD3	SD1	SD2	SD3	SD1	SD2	SD3
KS1	1.000	0.859	0.810	0.000	0.000	0.000	0.953	0.847	0.802
KS2	1.000	0.860	0.813	0.000	0.000	0.000	0.953	0.847	0.800
KSB1	1.000	0.898	0.855	0.000	0.000	0.000	0.953	0.886	0.852
KSB2	1.000	0.936	0.911	0.000	0.000	0.000	0.953	0.921	0.881
KSB3	1.000	0.935	0.900	0.000	0.000	0.000	0.953	0.930	0.905
MT(10)	0.997	0.998	0.993	0.000	0.000	0.000	0.914	0.882	0.877
W(10)	0.869	0.771	0.713	0.000	0.000	0.000	0.869	0.767	0.709
MTA(10)	0.997	0.984	0.984	0.000	0.000	0.000	0.916	0.787	0.764
	cl. 3 (Y) vs cl. 1 (X)			**cl. 2 (Y) vs cl. 3 (X)**			**cl. 3 (Y) vs cl. 2 (X)**		
	SD1	SD2	SD3	SD1	SD2	SD3	SD1	SD2	SD3
KS1	0.003	0.000	0.000	0.032	0.009	0.003	0.978	0.846	0.801
KS2	0.003	0.000	0.000	0.032	0.006	0.003	0.978	0.826	0.773
KSB1	0.003	0.000	0.000	0.032	0.010	0.006	0.978	0.876	0.831
KSB2	0.003	0.001	0.000	0.032	0.013	0.005	0.978	0.886	0.857
KSB3	0.003	0.000	0.000	0.032	0.014	0.006	0.978	0.914	0.881
MT(10)	0.003	0.004	0.002	0.042	0.006	0.014	0.998	0.990	0.964
W(10)	0.002	0.002	0.002	0.009	0.006	0.011	0.869	0.763	0.704
MTA(10)	0.003	0.001	0.007	0.044	0.005	0.008	0.998	0.987	0.986

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix B: A note on the Polish policy

Although the gender and urbanization income disparities in Poland are unambiguously stated in this and other studies, there is unfortunately no clear policy aiming at changing this situation. There are, however, some governmental and EU projects attempting to counteract the regional differences.

The gender differences found concerning income distribution are commonly observed across many countries. The gender wage gap is often discussed in terms of gender discrimination and related phenomena like "glass ceiling", "glass wall", "sticky floor"or "glass escalator", which relate to labour market segmentation in terms of employment sectors and positions occupied.[12] The gender equality rights are guaranteed by the Polish Constitution, including the equal rights to employment, promotion and the same wage for work of the same given value. Additionally, gender equality guarantees are also included in the Labour Statute Book and selected legislative acts.

However, there is actually no clear policy in Poland that aims at counteracting the existing negative discrepancies (see e.g. Polityka Równości (...)). The government concentrates rather on pro-family policy, allowing women to reconcile their family and professional life. The political proposals concern the flexible forms of employment, including the tele-work promotion and the increase of the number and accessibility of child-care institutions. Certainly, in order to change the existing situation, apart from legislative work, numerous stereotypes of gender social roles need to be overcome.

Regional and urbanization welfare differences are more often addressed by the national policy and may be easier to implement than the gender concerns. According to official estimates of Polish Central Statistical Office (GUS, 2007) the households living in the rural areas have, on average, one third lower equivalent disposable income than those inhabiting towns and cities. Moreover, those living in the agglomerations above 500 thousand people have income of around 45% higher than the country's average. These facts speak in favor of introducing a stronger state policy, creating better conditions for economic convergence for poorer subgroups.

As far as regional differences are concerned, there are plenty of programs designed to help the poorer regions, co-financed mostly by the European Union funds. For instance, in 2007

[12]The phenomena mentioned refer to the barriers in the careers of women (or any minorities); "glass ceiling"refers to the inequality in the chances of advancement to higher levels, "glass wall"describes the situation of women employed in the less prestigious and assistant-type positions, "sticky floor"refers to being trapped in low-wage and low mobility jobs, whereas "glass escalator"concerns the situation when even in the female dominated fields, the men are usually promoted.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

the European Commission launched the East Poland Development Program, which comprises five regions: Warminsko-Mazurskie, Podlaskie, Lubelskie, Małopolskie and Świetokrzyskie (that were included in the East cluster). The defined aim is to accelerate the development of this region, seen in Europe as peripheral region in a peripheral country. According to the information of Polish Ministry of Regional Development (Strategia Rozwoju (...)) these five regions are the poorest ones of the European Union with GDP not exceeding 40% of the EU average. This region accounts for 32% of Polish land area, approximately 22% of Polish inhabitants and only 16% of national GDP. The factors that strengthen the regional discrepancies are: not effective employment structure, low productive agriculture, low level of service and industry sector development, low quality of human resources (also low entrepreneurship), low indicators of urbanization and foreign capital involvement, as well as the lowest level of technical infrastructure. Moreover, according to the official documents concerning the strategy of regional development, the education level of the inhabitants of these five regions is lower than the national average. The same is true for the share of farms managed by the people with secondary or tertiary education.

The reason for the observed unequal development goes back to the times when Poland was divided between three occupants before the First World War. Also during the period of the People's Republic (1945-1989) the division into Poland A and B was still present and industry investments mainly concerned the south of Poland (Slask). Due to agricultural features, East Poland did not lure investments even after the fall of the Communist regime. At the end of the last millennium, parliamentary plans for land management pointed to the fact that the polygon between Gdansk, Bydgoszcz, Poznań, Wrocław, Kraków, Łódź and Warsaw is an EU competitive area. Other regions' capital cities have potential to become regional metropolises. However, the lack of infrastructure, in terms of airports, fast train connections, modern industry-office infrastructure or fair-congress centers, slows this process very much. The cities fulfilling these conditions were clustered in the analysis as "metro-regions" (apart from Bydgoszcz) and were proved to be characterized by higher welfare.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

CHAPTER 2

Income, Relative Social Status, and the Determinants of Happiness in Europe

Abstract

This study uses the ECHP 2001 dataset to investigate the determinants of individuals' happiness. We model happiness as a latent trait that stochastically determines responses to four satisfaction questions; the relations between social characteristics and the trait, and between the trait and the responses, are estimated in a unified semiparametric model. The results provide a succinct characterization of the role of covariates in determining the distribution of happiness within and between European states. Relative income plays an important role in the analysis and its effects are further examined by estimating sub-models for groups defined by incomes relative to cohort averages. This more detailed analysis suggests that that the genesis of happiness is affected by relative social status, and that income is more important to high status individuals.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

2.1 Introduction

> *"Men do not desire to be rich, but richer than other men"*, John Stuart Mill
>
> *"Money may be the husk of many things but not the kernel. It brings you food, but not appetite; medicine, but not health; acquaintance, but not friends; servants, but not loyalty; days of joy, but not peace or happiness"*, Henrik Ibsen

Nowadays, the amount of scientific work on happiness - satisfaction issues in different disciplines is flourishing, with a variety of questions being addressed. Among others, there is an interest in describing the relation between happiness and some demographic characteristics, together with health, income, culture, climate or political and economic freedom. Psychology and medicine studies try to discover what biochemical processes are behind happiness and how our bodies react to different psychological or physical stimuli. For sure, the nature of happiness is still not fully explored, neither is it explained and given the complicated nature of human beings, it is difficult to believe that it ever comes into being. However, the current scientific literature on this topic is prolific and allows us to understand a great deal from the observed patterns, behavior and paradoxes.

The economic literature on happiness may be divided into macro- and micro-level happiness analyses. Macro - level insights, apart from controlling for personal characteristics, try to describe the role of some general and macro-economic concepts like inequality (Gini coefficients), inflation, unemployment levels, GDP, life expectancies, divorce or crime rates in shaping individuals' levels of happiness in different countries, states or communities. In addition, the generosity of the state measured by benefit replacement rates, as well as the size of a community or some environmental variables (like SOx emission) may be included in the range of explanatory variables. The results, which are usually obtained, are in line with expectations - positive processes are positively correlated with the reported satisfaction levels and those seen as negative ones, go in the opposite direction. Moreover, the differences between different groups are outlined: Americans are less concerned about inequality than Europeans, probably because of the differences in mobility within the societies (Alesina 2004); the poor care less about the environmental degradation than the rich, and the positive effect of longer life expectancy is weaker for older people (Di Tella 2008).

By contrast, the pure micro-level analysis is currently being thought about, too. Attempts are being made to quantify the impact of gender, race, nationality, family life (marital status, children), age, education, religion, health, domicile, employment status, hours worked or amount of leisure, political attitudes and even some psychological concepts (e.g. the

"big five"personality traits). Lately, two specific and related topics, that is unemployment and relative income interdependence with satisfaction, have come to the attention of the economists. The first is dealt with by e.g. by Clark and Oswald (1994) who provide results about the involuntary feature of unemployment and its negative effects, greater than those of divorce. Korpi (1997) rejects the hypothesis about a health - based selection to unemployment and confirms the causal effect of being unemployed on diminishing satisfaction. In general, the direction of causality in the happiness field is difficult to establish; this is e.g. the case when the impact of such life events as getting married or divorced is analyzed (Stutzer and Frey 2006); the relationship between income, health and reported satisfaction is not clear either.

The issue of finding the determinants of happiness and quantifying their importance is also addressed by this study, where the analysis is conducted for six selected European countries. A special attention is given to relative income and a question whether the relatively richer and poorer subpopulations are characterized by different happiness origins. However, the main contribution of this work is methodological: the application of a semiparametric Item Response Theory Model and happiness being modeled as a latent trait believed to shape the answers to different satisfaction questions.

The paper proceeds as follows. Section 2 discusses different approaches to the inclusion of concern related to relative social status in economic analyses. The next section provides a description of the data - both the answers to 4 satisfaction questions (here called items) and personal characteristics. Section 4 presents the methodology applied in this study and is followed by the presentation of the results, while section 6 narrows the analysis to the differences between the two subpopulations - below and above the average income of the respective age cohorts. Finally, section 7 concludes.

2.2 Concern for relative social status

As already mentioned, concern about relative social status has been introduced into the happiness research recently. It is directly related to the concepts of interpersonal preferences or interdependent utility models, expressed as a view that people have utility functions depending on the perceived well-being of others, measured by their utility, income or consumption. The sociological terms also refer to the "relative deprivation"or "social frame of reference theories"and this issue was already addressed by e.g. John Stuart Mill, Karl Marx or Thorstein Veblen. There are plenty of analyses confirming the idea of people basing their happiness on the comparison to some "other individuals"or some "reference groups".

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

However, given the nature of the available datasets and the measurement limitations, it is income that is usually treated as a proxy of the absolute and relative social status.

The phenomenon of "reference drift effect" was introduced into the literature in the 1970's and it states that the individual's utility depends negatively on the income of the reference group. This idea is also applied to explain the famous Easterlin paradox, i.e. the situation when raising the income of all does not raise the general level of happiness, although people with higher incomes tend to report higher levels of satisfaction. The example of Japan, where a 5 - time increase in the GDP over 30 years in the second half of 20th century did not cause any increase in subjective well-being, may be striking, but the US and Western Europe also show similar patterns. The conclusion that income does not buy happiness, at least in the developed countries, might thus be drawn, but for sure it can play a role in the developing world where basic needs are still not met for the relevant sections of the populations. Together with the general increase in welfare, "being richer than other men" becomes more valued. Moreover, the "relevant other men" groups become bigger with the spatial borders disappearing thanks to the new and more available information channels.

The question of who constitutes the reference group of an individual arises. Do people look for comparison in the neighborhood or region, among people of similar age, with the same education, gender, of the same race, or with a similar profession? The relevant others might also be sought in favorite TV soap operas and they do not need to stay the same over time, e.g. because of job or neighborhood changes. Moreover, the habituation issue also arises, since individuals look at their past situation and experiences (Clark et al. 2003);[13] however, most of the interest covers the "external" i.e. social reference questions, and internal comparison goes beyond this analysis. In practice, "people like me" groups are constructed in different ways: some studies assume that individuals compare themselves with all the citizens of the same country. Luttmer (2005) follows the approach of "keeping up with the neighboring Joneses", using a within - region comparison, for regions taken to be the so-called PUMAs (Public Use Microdata Areas), each one with around 150,000 inhabitants. McBride (2001) includes all people in the USA who are 5 years younger and 5 years older than the individual concerned in her reference group. Ferrer-i-Carbonell (2005) introduces even some further simultaneous reference categories: education, age and region.

Relating own situation to peers certainly influences the creation of individual aspiration levels. This is exploited in the study by Stutzer (2004) where he uses Swiss data and specifically the answers to the questions about the sufficient and minimum required level of

[13]Comparison to parents' standard of living may also be introduced into the analysis; the information on this is included in the General Social Survey and was exploited by McBride (2001).

DOI: 10.2870/38699

income (which should be highly correlated with the aspiration levels). The finding is not surprising - higher income aspiration levels reduce satisfaction with life, ceteris paribus, and a higher income level in a community is reflected in higher aspirations. However, it is possible that in some specific situations higher income of peers may play a positive role, signalling the possibility of an increase in future wage (so called Hirschman effect). The positive effect of higher co-workers wages is shown in Clark et al. (2009) in the matched employer-employee dataset based on the Danish ECHP sample. Similarly, Clark et al. (2008) argue that richer neighbours influence us through two independent channels: through making others feel worse off in the relative income sense, but at the same time bringing positive spillovers in creating local communty social capital.

The next issue, characterized by a wide variability of approaches in the literature, refers to the technical part of introducing the relative income concern into the statistical analysis. This might be done by computing some cell averages or by estimating predicted incomes. Ferrer-i-Carbonell (2005) presents several possible solutions following the first approach: the proxy for the relative deprivation is assumed to be the average income of the reference group or the distance between the individual's own and the reference group income. Some comparison asymmetry may also be introduced, i.e. the happiness of individuals might be negatively affected by an income below that of the reference group, but for those above, no positive impact is expected. Since the analyzed databases often have a poor measure of income, it is possible to match a dataset from an external source (e.g. McBride 2001 uses CPS matched with GSS). Clark and Oswald (1996) introduce the comparison income as the predicted income from the conventional earning equation, which represents the income of a typical employee with given characteristics. Other ideas include the ranking in the income distribution, the quintile of the distribution the individual belongs to, or the shares of poor and rich in the communities (Tomes 1986).

2.3 Data description

This study uses the data collected in the European Community Household Panel (ECHP). The panel covers years from 1994 to 2001 (8 waves) and includes 15 countries (although Austria, Finland and Sweden joined the project later). In the first wave, more than 60 thousand households were interviewed, that is, approximately 130,000 adults aged 16 and over. The ECHP standardized methodology and procedures provide comparable information across countries, but this analysis is limited to six countries: Austria, Denmark, France, Greece, Ireland and Italy. The choice was based on the data availability (variables of interest)

and the diversity of cultural, sociological, economic or even climate factors assigned to each of these countries. This study does not exploit the panel structure of ECHP, since only the wave 2001 is used here. Moreover, the unit of the analysis is an individual, which gives in the end more than 16 thousand observations: 2,076 Danes, 2,221 Austrians, 1,465 Irish, 4,212 Italians, 2,275 Greeks and 4,337 French. The sample of respondents was narrowed to the individuals who normally work more than 15 hours a week (i.e. in paid employment or being self employed; the unemployed are excluded) and who report positive income.

2.3.1 Items

The term happiness, used interchangeably with subjective well-being (SWB) or utility, is difficult to measure directly. Usually, it is proxied by self-reported assessment referring to general satisfaction with life questions ("*How satisfied are you with your life as a whole these days?*") or directly to the level of happiness ("*Taken all together, how would you say things are these days - would you say that you are very happy, pretty happy, or not too happy?*").

An analysis may also be limited to a specific life domain - like health, financial situation, job, leisure, housing, environment, social protection - and this information is often provided in the datasets. These specific measures might be aggregated (building an index) in order to obtain a general level of satisfaction.

An alternative to subjective well-being or satisfaction questions is the application of some psychological health measures, which, among others, include information about feelings of happiness, strain, depression, stomach problems and insomnia. Such an approach is used by e.g. Clark and Oswald (2002) and Korp (1997).

However, there are clear shortcomings with the dependent variable construction described above. The answer to a single question can be easily manipulated in surveys by an appropriate ordering of issues asked before, reminding the respondents about the positive or negative aspects of their lives. On the other hand, aggregating different items together (summing up, taking the averages) may imply a loss of information and impose some level of rigidity. The aggregation strategy is often imposed by the estimation method - the commonly applied OLS or ordered probit/logit specifications require a single dependent variable. Instead, this study introduces the item response theory modeling approach, which allows several satisfaction questions to be treated separately, but at the same time to account for the common factor driving all the answers.

Specifically, there are four satisfaction items in the ECHP dataset. They concern the respondent's levels of satisfaction with work (or main activity), financial situation, housing situation and amount of leisure. There are 5 possible answers: largely unsatisfied, mildly

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

unsatisfied, mildly satisfied, largely satisfied, fully satisfied. Since the frequencies of the bottom and top answers are very low for most of the cases, the first two and last two categories are merged. In effect, there are 3 satisfaction levels in the analysis, with the last one (answer 3) expressing the highest level of satisfaction.

Table 2.1. Frequencies of the satisfaction answers

answer	main activity	finances	housing	leisure
	All countries			
1	0.064	0.125	0.052	0.135
2	0.425	0.560	0.350	0.511
3	0.511	0.315	0.597	0.355
	Denmark			
1	0.022	0.049	0.027	0.083
2	0.257	0.380	0.213	0.454
3	0.721	0.572	0.760	0.463
	Italy			
1	0.114	0.179	0.084	0.208
2	0.524	0.625	0.469	0.563
3	0.362	0.196	0.447	0.229
	Austria			
1	0.019	0.083	0.023	0.078
2	0.202	0.398	0.162	0.362
3	0.778	0.519	0.815	0.560
	Ireland			
1	0.038	0.087	0.033	0.070
2	0.470	0.599	0.383	0.565
3	0.492	0.315	0.584	0.365
	Greece			
1	0.103	0.179	0.086	0.227
2	0.588	0.693	0.557	0.646
3	0.309	0.127	0.356	0.127
	France			
1	0.048	0.117	0.036	0.091
2	0.422	0.581	0.278	0.474
3	0.530	0.302	0.686	0.435

We assume that the answers are comparable across respondents, excluding e.g. the qualia problem or anchoring effects.[14]

[14]Anchoring effects relate to a cognitive bias appearing when too much importance is assigned to one

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

There are clear differences in responses patterns for different countries, as we can already conclude from Table 2.1, which presents the answer frequencies. The Austrians and the Danes give on average higher answers than other nations. Greece and Italy are characterized by the highest percentages of individuals who evaluate their satisfaction at the lowest levels. Moreover, there is an answer pattern across the items - *finances* and *leisure* are the domains that obtain the lowest scores. In "standard"item response theory approach, usually found in ability testing, this fact would be reflected in a difficulty ranking. That is, the parametric specifications contain usually a set of the difficulty parameters, that differ across the items and allow to account for the differences in the probabilites of "correct"answers for questions included. In the semi-parametric IRT model used here, there is no clear difficulty specification, but some conclusions can be still drawn based on the results obtained.

Finally, the question of whether the items measure one concept may be raised. This unidimensionality requirement seems to be met in this setting, which is supported by the nonparametric correlation coefficients[15] presented in Table 2.2. The p-values for the pairwise associations also indicate significant associations between all items. Also, applying factor analysis to reduce the dimensionality of the problem provides the result of 1 factor explaining 53% of the observed variation in answers, which is a high value speaking in favor of the unidimensionality.

Table 2.2. Nonparamteric item correlation coefficients

	m. activity	**finances**	**housing**	**leisure**
m. activity	*****	0.471	0.389	0.291
finances	<0.001	*****	0.377	0.277
housing	<0.001	<0.001	*****	0.322
leisure	<0.001	<0.001	<0.001	*****

The upper diagonal part contains correlation coefficient estimates

The lower diagonal part contains corresponding p-values (H0: tau=0 against H1: tau≠0)

The values reported support the assumption that the items refer to the same underlying latent trait - individual happiness, although there is still a lot of unexplained item-specific

aspect of an event or a specific value. Anchoring effect in econometric model specifications might be incorporated e.g. as intercept heterogeneity (objectively the situation may be the same, but subjective evaluations differ). Qualia concerns a problem of subjective and objective character of experiences and occurances, i.e. the fact that we experience the world differently and that things seem to us in different ways (e.g. if and how we can distinguish between two "different"red colours).

[15]Kendall's tau coefficients were calculated using the R ltm package (Rizopoulos 2006).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

variability. In this study we aim at obtaining the distribution of happiness across the population of interest using the information from these four answers and personal characteristics. The technical assumption behind this is the stochastic dominance relation - i.e. the respondents with higher values of happiness tend to give higher answers to any question concerning satisfaction.

Conceptually, we need to justify the link between the latent happiness and answers to different satisfaction questions. The conceptual referents introduced by Rojas (2007)[16] can be employed to support the idea of latent happiness driving answers to different life domain satisfaction questions. Among others Rojas introduces the following definitions: "Happiness is accepting things as they are"(stoicism); "Happiness is being satisfied with what I have and what I am"(satisfaction); "Happiness is to enjoy what one has attained in life"(enjoyment); "Happiness is to seize every moment in life"(carpe diem); "Happiness is in living a tranquil life, not looking beyond what is attainable"(tranquility). All these statements support the assumption of the common factor - unobservable happiness, driving different answers. Nevertheless, providing the definition of happiness is an open and complex issue for social scientists and it can be expected that is not universal across individuals.

2.3.2 Individual characteristics

The set of the explanatory variables in the happiness/satisfaction studies is relatively standardized (see e.g. Dolan et al. 2008) and the results concerning the impact of personal, economic and social factors are usually similar. However, as already mentioned, in some cases the direction of causality is not clear.

Firstly, an attempt to quantify the impact of demographic characteristics is made. Gender is usually a significant predictor of satisfaction, with women reporting higher scores. However, some studies report diminishing gender differences or even an inversion of the pattern (Stevenson 2008), but the choice of the measure used as a proxy for well-being and the set of other control variables seems to matter. Marriage status is found to play a significant role with those being married assessing their happiness as higher. In this study, there are 4 exclusive categories: being single, married (reference category), cohabitating and divorced (merged with separated). Widowers are excluded from the sample because of their small number and some estimation problems.[17] The category cohabitation is included and

[16]His work concentrates on the relation between different attitudes to happiness and the role of income in shaping the evaluation of own life.

[17]This relates to the increasing number of "zero cells" issues, i.e. a problem of some combinations of answers and individual characteristics not supported by the data.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

comprises the individuals who are not married (but might be divorced) and who live with a partner. The scale and importance of such relationships in Europe increases as an alternative to marriages and it should also be reflected in this kind of study.

Age is the next demographic characteristic believed to shape an individual's happiness; studies suggest a U-shaped relation with the minimum occurring in middle age, between 32 and 50 years old. Here the age range was limited to 20-60, since the inclusion of pensioners and young individuals (although possibly participating in the labour market) might introduce too much heterogeneity and some spurious results. The dataset in hand does not contain any information about the number of children, nor about religion or the degree of religiosity, which are usually found to be significant predictors.

As far as some socioeconomic characteristics are concerned, we control for educational level, with secondary education (3rd and 4th levels in ISCED classification) being the reference category. However, the findings in the literature differ significantly. Clark and Oswald (1996) report that highly educated people appear less content with their jobs. However, in this case, this could be explained by their having higher expectations than employees with lower education. Nevertheless, the general relation is assumed to be positive, especially when SWB is proxied by general life satisfaction questions. Information relating to the occupational life of respondents is also included: the average number of weekly hours worked and income. The latter refers to the reported net income and is converted to common currency units correcting for purchasing power parity. Observations corresponding to equivalised net income below 300 and above 3500 are removed from the analysis. Reflecting the usual concavity assumption, we model utility as logarithmic in both income and hours worked. We regard income as a proxy for consumption and hours worked for negative leisure. An increase in the number of hours worked is mostly found to be negative, as it decreases individuals' free time. However, the opposite was observed for the German data (mentioned by Dolan at al. 2008), but was explained as the difference between part-time and full-time workers, with the former probably expressing unsatisfaction with the lack of full-time positions.

Finally, respondents were also asked to assess their health ("*How is your health in general?*"), the dummy *v. good health* reflects the answers of very good health and *bad health* the answers of bad and fair health, whereas good subjective assessment is the category omitted. Apart from enjoying good health, an active social life is also assumed to have an positive impact on SWB. The variable *social* takes on value 1 if the answer to the question "How often do you meet friends or relatives not living with you, whether here at home or elsewhere?"is "most days"or "once/twice a week"; that is, the person can be regarded as a sociable person. Country fixed effects are also introduced into the model with France being

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

the reference category.

The descriptive statistics reporting either frequencies for dummy variables or the means for the continuous (or pseudo - continuous) ones are presented in Table 2.3.

Table 2.3. Descriptive statistics of variables analyzed

	All	**Austria**	**Denmark**	**France**	**Greece**	**Italy**	**Ireland**
	Frequencies (%)						
single	24.55	27.42	11.99	18.86	29.98	27.64	37.54
married	61.36	57.50	62.62	57.87	64.40	66.76	55.49
divorced	5.55	6.26	8.77	7.17	3.82	3.58	3.41
cohabitation	9.71	9.37	17.87	18.81	1.80	2.83	3.89
male	56.20	57.23	50.92	54.11	60.70	59.85	50.78
high education*	23.72	11.03	35.79	32.98	27.56	13.11	26.00
low education*	33.55	11.75	11.46	55.4	29.27	37.63	28.05
v. good health	36.30	46.69	48.94	11.69	77.01	19.66	60.07
bad health	18.67	10.76	14.11	31.57	4.48	23.60	6.76
sociable	81.15	72.80	78.61	73.44	89.93	84.92	95.70
	Sample averages						
age	38.61	37.69	41.01	38.92	37.79	38.69	36.79
hours worked	38.26	38.57	38.22	37.34	40.89	37.84	37.37
net income**	1153.2	1079.7	1228.7	1239.5	888.4	1057.5	1588.6

*High education corresponds to ISCED levels 5-7, whereas low to 0-2

**The equivalized values (corrected for purchasing power parity)

As already mentioned, special interest is dedicated to the importance of relative social status, proxied here by relative income. The reference group for each respondent is defined as the cohort of individuals 5 years younger and 5 years older (following McBride 2001), separately for each country.[18] There are several arguments in favor of this approach: gender differences in terms of socio-economic characteristics (education, wages and expectations) decrease nowadays in developed European countries. Moreover, limiting the comparison to some selected regions might also be seen as too restricitive, especially with modern information technology. On the other hand, it seems unrealistic for 25 - year -old individuals to

[18]The age ranges of reference groups for the individuals below 25 and above 55 years old are respectively smaller because of the inclusion of only individuals aged 20-60 in the sample.

compare their standard of living with 50 - year - old citizens, so the cohort of 20 - 30 - year - old people is regarded to approximate the true group of reference in this case. Contrary to the study mentioned above, relative income concern is introduced into the model not as an average income of the reference group, but just as a dummy taking on value 1 if the respondent reports the income above the respective reference group average.[19] The alternative specification introduces the quartile of the income distribution to which the individual belongs (separate for each country), with the 2nd quartile being the reference category. Based on the previous findings (e.g. Clark and Oswald 1996, McBride 2001, Ferrer-i-Carbonell 2005), it is expected that a significant link between well-being and relativities will be found.

2.4 Methodology

2.4.1 General approach

Satisfaction analyses usually involve basic econometric tools. In the simplest case, when the OLS estimation is applied, the cardinality assumption of the responses is imposed, meaning, for instance, that an increase in evaluation from 1 to 2 is the same as from 2 to 3. This is not, however, a desired property; instead ordinality is preferred and true SWB/happiness treated as a latent variable. In this case, the ordered probit/logit regression is usually applied. Nevertheless, as mentioned in the items description section, these methods allow only for a single dependent variable (i.e. either just a single item, or an aggregated measure). If there are several responses about satisfaction with life domains, which is the case when dealing with ECHP data, these methods might be found unsatisfactory and as not exploiting all the information in hand. Factor analysis could be a solution to this problem, but in general it is applied rather as a check of dimensionality and a data reduction technique. It tries to find a set of factors/a factor able to reproduce the data accounting for the covariance between the items. Moreover, it is claimed that factor analysis does not investigate the interaction between items and respondents, which is, of course, of interest.

An alternative to the above - mentioned approaches is item response theory (IRT), whose application seems to be undervalued in the happiness field. Its origin traces back to psychometrics and the measurement of ability based on tests scores. This method allows for dealing with more than one dependent variable at the same time, without imposing any index

[19]The value of the reference group mean income was not introduced into the model, because of the likely inaccuracy. Building 40 reference groups from, in the best case, 4300 observations (France) might not represent the real situation. Matching the data for income from external datasets was not applied here.

building transformations. In general, IRT models describe the association between a respondent's underlying level on a latent trait and the probability of a particular item response. Currently, there is a wide variety of IRT models, both parametric and semi-/nonparametric ones. The parametric specifications, depending on the range of parameters introduced, allow for the item difficulty, item discrimination (how much information about the latent trait the item conveys), chances of guessing or some personal characteristics to be accounted. These models do not assume the cardinality of the answers; like logit/probit specifications the ordinarity assumption occurs corresponding here to the stochastic dominance feature that individuals with higher levels of latent trait are more likely to give higher answers.[20] Moreover, similarly to the standard models of discrete choice, the probabilities of a certain answer are estimated, but as a function of the latent trait, whereas in logit/probit approach this is a function directly of personal characteristics.

The assumptions of unidimensionality (only 1 concept is measured by the items), monotonicity (or stochastic dominance: respondents with higher values of the latent trait give higher responses) and local independence (items are uncorrelated with each other when the latent trait has been controlled for) are common for different IRT models. However, these assumptions can also be relaxed in different specifications.

2.4.2 The semi-parametric item response theory model

The methodology applied in this study is the semi-parametric IRT model developed by Spady (2006, 2007). This model assumes that the distribution of the latent trait varies across the population and can be determined as a function by both the item responses and individual characteristics. In this setting we aim at estimating each individual's distribution of happiness, given the four answers to the different life domain satisfaction questions and the individual characteristics. Technically, we obtain the set of distributions: $f(\theta \mid W, \mathbf{r})$, where θ stands for the latent trait, W for a vector of respondent's characteristics and $\mathbf{r}$ for a vector of answers to the satisfaction items. Individual happiness is the factor driving all four answers, and happiness itself is influenced by personal characteristics. However, the latter have no direct effect on the answers, only through the latent happiness,[21] which is embodied

[20]Note that the statements of kind: indivduals with X - as much of the latent trait are Y - more likely to give a respective answer, do not appear here, since the specification is nonlinear and specifically in this study - semiparametric.

[21]Relaxing this assumption for one or more items leads to the DIF (differential item functioning) specification. It is introduced when some subgroups (basing on gender, race, etc.) have different probabilites of specific answers for the same levels of the latent trait. The preliminary analysis here has not found sufficient justification for introducing DIF.

by:

$$p(r_1, r_2, r_3, r_4|\theta, W) = p(r_1, r_2, r_3, r_4|\theta) \tag{2.1}$$

and

$$p(r_1, r_2, r_3, r_4 \mid W) = \int p(r_1, r_2, r_3, r_4 \mid \theta) f(\theta \mid W) d\theta \tag{2.2a}$$

Since this model incorporates also the local independence assumption, we can write:

$$p(r_1, r_2, r_3, r_4 \mid W) = \int p(r_1 \mid \theta) p(r_2 \mid \theta) p(r_3 \mid \theta) p(r_4 \mid \theta) f(\theta \mid W) d\theta \tag{2.2b}$$

The left-hand side of this expression is a function of the observed data, but the right-hand side involves the unobservable trait θ. In order to evaluate the integral, firstly $p(r|\theta)$ and $f(\theta \mid W)$ need to be estimated and it is done in a semi-parametric framework. A parametric requirement imposed in this approach is the assumption of $f(\theta \mid W)$ being $N(\mu(W), 1)$, with $\mu(W) = W\beta$; that is, the distribution of the latent trait for each individual is assumed to be normal with the mean being a linear function of the characteristics and variance equal 1.[22] The comparison of the obtained results is drawn by referring to a baseline respondent, whose distribution of the latent trait is $N(0, 1)$. This person (who might be hypothetical, but who might also be a real respondent in the dataset) has only zeros in her W vector. In this setting this corresponds to a married French woman, of middle level of education, enjoying good health, not being sociable (meeting family and friends less than once a week). The continuous variables need some rescaling to ensure the zero-value characteristics of the baseline respondent: this is done by centering them around their means (for the values of sample averages see Table 2.3). $f(\theta \mid W)$ for various subpopulations is modeled by normal additive location shifts.

The specification of $p(\mathbf{r} \mid \theta)$ is free from parametric assumptions. However, for the model to make sense, the monotonicity requirement needs to be imposed. This is expressed here in terms of stochastic dominance relations: the responses of individuals with higher values of the latent trait first order stochastically dominate the responses of those with the lower values of the trait. In other words, happier people tend on average to give higher scores on the satisfaction question. Graphically, this is represented by the downward sloping (weakly monotonically decreasing) and non-crossing item characteristic curves, which illustrate the

[22]Since θ is not directly observed here, specifying its distribution as normal is just a matter of scaling. The choice of a uniform distribution can be implemented as well. The variance of the latent trait can be also specified as a function of the personal characteristics.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

correspondence between the responses and the latent trait. An example is Figure 1 in the next section, where the estimation results are discussed. The lowest curve in each box shows the probability of answer 1 for a given satisfaction question, i.e. $F(r_i = 1|\theta)$; the curve above of answer 2 or less: $F(r_i = 2|\theta)$, and the last curve is just the constant "1"line (omitted in the graph), which would represent the certainty of answering 3 or less: $F(r_i = 3|\theta)$. In order to obtain the specific probabilities of possible answers, we subtract the value indicated by a respective curve from this lying directly above. The non-crossing condition corresponds to the non-negativity of probability requirement. The downward sloping feature ensures that for the increasing value of the latent trait, the probability of a higher answer grows. Technically, the item characteristic curves are constructed using exponential titlting; the detailed description may be found in the Appendix.

The value of $p(r_1, r_2, r_3, r_4 \mid W)$ as found in the equation 2b is calculated for each respondent and the resulting likelihood function for the whole sample is estimated by maximum likelihood. In the last step, the posterior distribution of happiness for each individual given her answers and characteristics is obtained by applying Bayes Law:

$$f(\theta \mid W, r) = \frac{f(\theta, r \mid W)}{p(r \mid W)} = \frac{p(r \mid \theta, W) f(\theta \mid W)}{p(r \mid W)} = \frac{p(r \mid \theta) f(\theta \mid W)}{p(r \mid W)} \tag{2.3}$$

2.5 IRT estimation results

Using the data described and applying the semi-parametric IRT methodology three models were estimated. As already mentioned, 4 satisfaction questions and a range of personal characteristics are taken into account at the same time. The estimation results are threefold: the effect of the latent happiness on the probabilities of specific answers to satisfaction questions, individual distributions of the latent trait, and the role of the personal characteristics in shaping them. The estimated coefficients that are an answer to the latter issue are presented in Table 2.4 and the item characteristic curves, illustrating the items - happiness relationship, are plotted in Figure 2.1. The individual distributions of happiness are discussed at the end of this section.

2.5.1 The impact of personal characteristics on happiness

Table 2.4 provides the estimated coefficients determining the location shifts of an individual's happiness distribution corresponding to a specific characteristic, i.e. the column "coefficient"corresponds to the vector β that specifies the distribution $f(\theta \mid W) \sim N(W\beta, 1)$.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Three different specifications may be seen as a check for model stability. The first specification does not include any relative income proxy among the personal characteristics, the second one accounts already for the relative social position by introducing a variable *rincome* defined in the data description section. The last specification replaces *rincome* with 3 dummy variables, indicating the quartile of the income distribution (seperately for each country) the respondent belongs to, with the 2nd quartile being the base category. The detailed discussion of the results relates to the coefficients under specification 2, unless indicated differently.[23]

Apart from the coefficients next to the *income* variable, the impact of other characteristics remains stable across different specifications. These results are in line with intuition and are discussed here with details for the second model. Married or cohabitating respondents are in general happier than single and are much happier than divorced. Men are usually less satisfied than women, and education increases happiness. The coefficients assigned to the latter are not of a large absolute magnitude, but since income is controlled for, the education effect may capture just the satisfaction effect of higher status in the society. A very strong negative effect is assigned to bad health, and that of good health is comparable with having a relatively intensive social life.

The age effect confirms the usually assumed concave relationship. IRT results suggest that the utility drops until age 31 and then increases, ceteris paribus, with individuals of 42 years reaching the same happiness as those of 20 years. The difference in location of the individual latent trait distribution between 20 and 60 year - old people (possessing the other characteristics of the base respondent) amounts to 0.47.[24]

As far as the effect of income is concerned, it can be concluded from the results that money does make people happy. Higher income allows people to meet their desires, buy more goods or services and simultaneously reach a higher status in society. The results for specification 1, assuming only the absolute income importance, indicate that a 100 unit increase in equivalent income amounts to a 0.05 location shift if this is given to an individual earning 1200 net equivalent units (with other characteristics of the baseline respondent).[25]

[23]None of these specifications includes variables interactions (e.g. checking if higher education has stronger impact on happiness for males) in order to keep the model simple and concentrate on the methodology.

[24]The interpretation of results relating to the continuous variables is not straightforward, since the variables were centered around the means; moreover age is introduced in the quadratic form and the obtained coefficients correspond to a and b in the following expression: $a(age - mean(age)) + \frac{b(age-mean(age))^2}{100}$; the value of the mean can be found in Table 2.3.

[25]In order to find these values for the income variable, one need to bear in mind that the obtained coefficient relates to the expression: $(\ln(income) - \ln(mean(income))$; similarly for *hours worked*.

Table 2.4. IRT estimation results

	Specification 1			Specification 2			Specification 3		
	coef	**s.e.***	**p-value**	**coef**	**s.e.***	**p-value**	**coef**	**s.e.***	**p-value**
single	-0.0262	0.0288	0.3635	-0.0239	0.0283	0.3983	-0.0222	0.0282	0.4306
divorced	-0.2779	0.0418	0.0000	-0.2749	0.0415	0.0000	-0.2746	0.0416	0.0000
cohabitation	0.0298	0.0377	0.4288	0.0353	0.0364	0.3323	0.0334	0.0360	0.3536
male	-0.1532	0.0250	0.0000	-0.1612	0.0229	0.0000	-0.1698	0.0226	0.0000
highedu	0.0414	0.0303	0.1722	0.0282	0.0272	0.2996	0.0162	0.0266	0.5417
lowedu	-0.0937	0.0312	0.0027	-0.0879	0.0276	0.0014	-0.0860	0.0266	0.0012
vhealth	0.2584	0.0240	0.0000	0.2579	0.0237	0.0000	0.2542	0.0236	0.0000
bhealth	-0.4602	0.0297	0.0000	-0.4581	0.0278	0.0000	-0.4640	0.0273	0.0000
social	0.2863	0.0323	0.0000	0.2859	0.0273	0.0000	0.2848	0.0261	0.0000
income	0.6340	0.0329	0.0000	0.4420	0.0433	0.0000	0.2894	0.0565	0.0000
rincome	-	-	-	0.2058	0.0284	0.0000	-	-	-
q1	-	-	-	-	-	-	-0.0478	0.0335	0.1537
q3	-	-	-	-	-	-	0.1402	0.0301	0.0000
q4	-	-	-	-	-	-	0.3953	0.0413	0.0000
hours	-0.6007	0.0477	0.0000	-0.5633	0.0468	0.0000	-0.5364	0.0468	0.0000
age	0.0077	0.0013	0.0000	0.0099	0.0013	0.0000	0.0068	0.0013	0.0000
age2	0.0707	0.0103	0.0000	0.0640	0.0099	0.0000	0.0708	0.0099	0.0000
denmark	0.3842	0.0405	0.0000	0.3874	0.0367	0.0000	0.4077	0.0359	0.0000
ireland	-0.5902	0.0528	0.0000	-0.5361	0.0492	0.0000	-0.4992	0.0486	0.0000
italy	-0.6747	0.0413	0.0000	-0.6964	0.0347	0.0000	-0.7128	0.0334	0.0000
greece	-0.9851	0.0493	0.0000	-1.0405	0.0438	0.0000	-1.0881	0.0443	0.0000
austria	0.6410	0.0433	0.0000	0.6136	0.0382	0.0000	0.5983	0.0370	0.0000
Log-Likelihood	51994.8277			51968.7044			51938.9598		

*the reported standard errors are the robust estimators

DOI: 10.2870/38699

If the income is increased by the same amount but for someone reporting a 900 net income, the increase in the mean of the latent trait distribution is already close to 0.07. This effect is much stronger for the poorer (e.g. 0.14 if the income is raised from 400 to 500), which is due to the logarithmic specification of income introducing the diminishing marginal utility. However, if the relative position is controlled for, these effects are smaller and they drop respectively to 0.04, 0.05 and 0.10. In other words, adding the variable relative income decreases the impact of absolute level of income. Moreover, the relative income effect seems to be much stronger than that of the absolute income: for a respondent with a reported net income of 900, it should be increased to around 1400 equivalent units to bring about the same change in average happiness as just moving from the group of people with lower than average incomes. However, the change in income of 500 equivalent units corresponds to 45% of the mean for the whole sample, so it does not seem to be a realistic scenario. Being above or below the average income of the age reference group makes a significant difference in the level of happiness, confirming the theory that relative income position matters, with its absolute value comparable to the effect of divorce and enjoying a very good health.

Hours worked have an expected negative effect on happiness and e.g. increasing the weekly working time from 35 to 40 hours for someone with the baseline characteristics amounts on average to a 0.075 drop in the mean of $f(\theta \mid W)$. Finally, bearing in mind that France is the reference category, we see that on average the Austrians, followed by Danes are the most happy. The inhabitants of the Mediterranean area seem to be the least happy from the European countries chosen. These country shifts can reflect the differences in the mentality, but also the general macroeconomic situation, the role of social security systems, religion-related issues might be captured here.

2.5.2 Latent happiness and the answers to satisfaction questions

The estimation of the distributions $F(r_i|\theta)$ for each item is represented by the item characteristic curves which are illustrated in Figure 2.1. The dashed lines indicate the estimated curves applying the parametric Graded Response Model,[26] assuming the logit link function, the item difficulty, and item discrimination parameters. The parametric specification imposes the lower level of the curve shapes' flexibility, which is especially visible for the *leisure* item. However, for the remaining three items the shapes are relatively similar. The

[26]Formally, the Graded Response Model is represented by the formula: $P(r_{ij}) = \frac{\exp[a_i(\theta - b_{ij})]}{1+\exp[a_i(\theta - b_{ij})]} - \frac{\exp[a_i(\theta - b_{ij-1})]}{1+\exp[a_i(\theta - b_{ij-1})]}$, with a_i being the discrimination parameter of item i and b_{ij-1} - the difficulty parameter of answer j for item i.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

item category response functions illustrating directly the probabilities $p(r_1, r_2, r_3, r_4 \mid \theta)$ are presented in the Appendix (Figure 2.3)

The probabilites of the specific answers for chosen values of the latent trait $p(r \mid \theta)$ are presented in Table 2.5.

Table 2.5. Estimated probabilites for different positions on the happiness scale

	main activity			**finance**			**housing**			**leisure**		
theta	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
-2.5	0.728	0.271	0.001	0.827	0.173	0.000	0.604	0.389	0.007	0.731	0.265	0.004
-1.5	0.109	0.851	0.039	0.273	0.724	0.003	0.045	0.825	0.129	0.168	0.748	0.084
-0.5	0.023	0.692	0.286	0.072	0.878	0.050	0.025	0.487	0.488	0.105	0.645	0.250
0.0	0.023	0.457	0.521	0.069	0.760	0.171	0.025	0.318	0.656	0.105	0.583	0.312
0.5	0.023	0.234	0.743	0.068	0.523	0.409	0.025	0.188	0.787	0.105	0.509	0.386
1.5	0.019	0.033	0.948	0.047	0.102	0.851	0.024	0.038	0.938	0.095	0.213	0.692
2.5	0.005	0.004	0.991	0.010	0.013	0.977	0.009	0.005	0.986	0.029	0.037	0.934

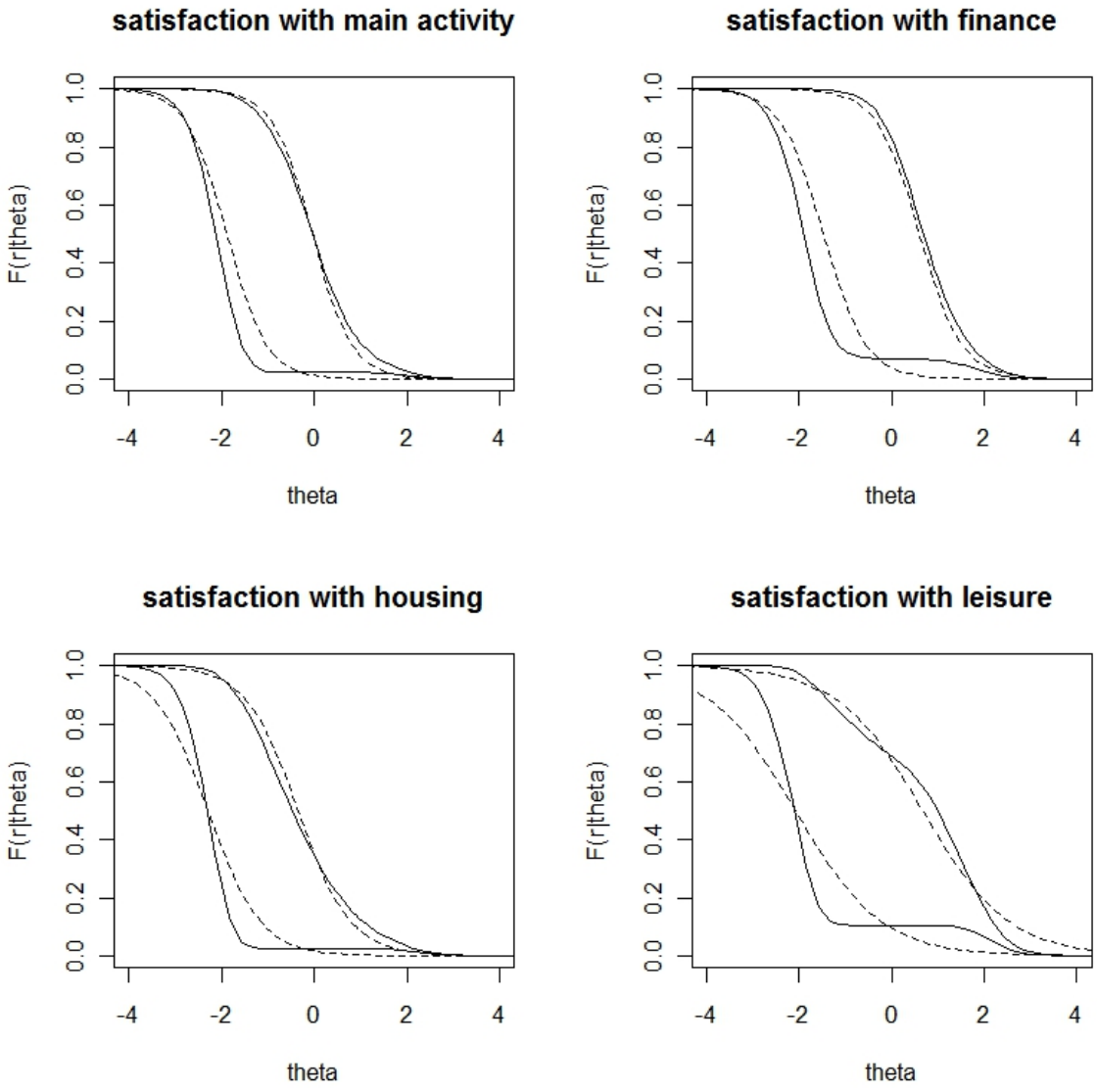
The fact that for greater values of theta the probabilities of lower answers drop and for higher answers grow, illustrates the required monotonicity or stochastic dominance condition. Moreover, as already indicated, for the items *leisure* and *finances* the estimated probabilities of answer 3 are lower than for the other 2 items for different values of *theta*. This confirms the idea that these two life domains are more unlikely to achieve full satisfaction, which could be also due to more limited possibilities of influencing own financial situation and not being able to realize leisure plans.

2.5.3 Individual posteriors of happiness

Finally, with the results on $f(\theta \mid W)$ and $p(\mathbf{r} \mid \theta)$, as well as $p(\mathbf{r} \mid W)$ (as given in equation 2b), the distribution of the latent trait for each respondent may be calculated following the expression in formula 3. Figure 2.2 presents the posterior happiness distributions for the chosen respondents. This figure shows the complexity of the estimation methodology, i.e. accounting for individual characteristics and answers to the satisfaction questions. Both of these play a role in obtaining the individual's happiness distributions.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 2.1. Item characteristic curves



The first plot illustrates the significance of the country - fixed effects. The chosen respondents are fairly similar to each other, apart from nationalities: they are married women, of middle education, good health, being sociable, with incomes above the respective cohort's mean and with absolute income, hours worked and age close to sample averages. None of them was characterized by extreme values of item answers. However, the plot shows clear shifts in the posteriors, with Austrian, Danish and French representatives moved to the right part of the plot.

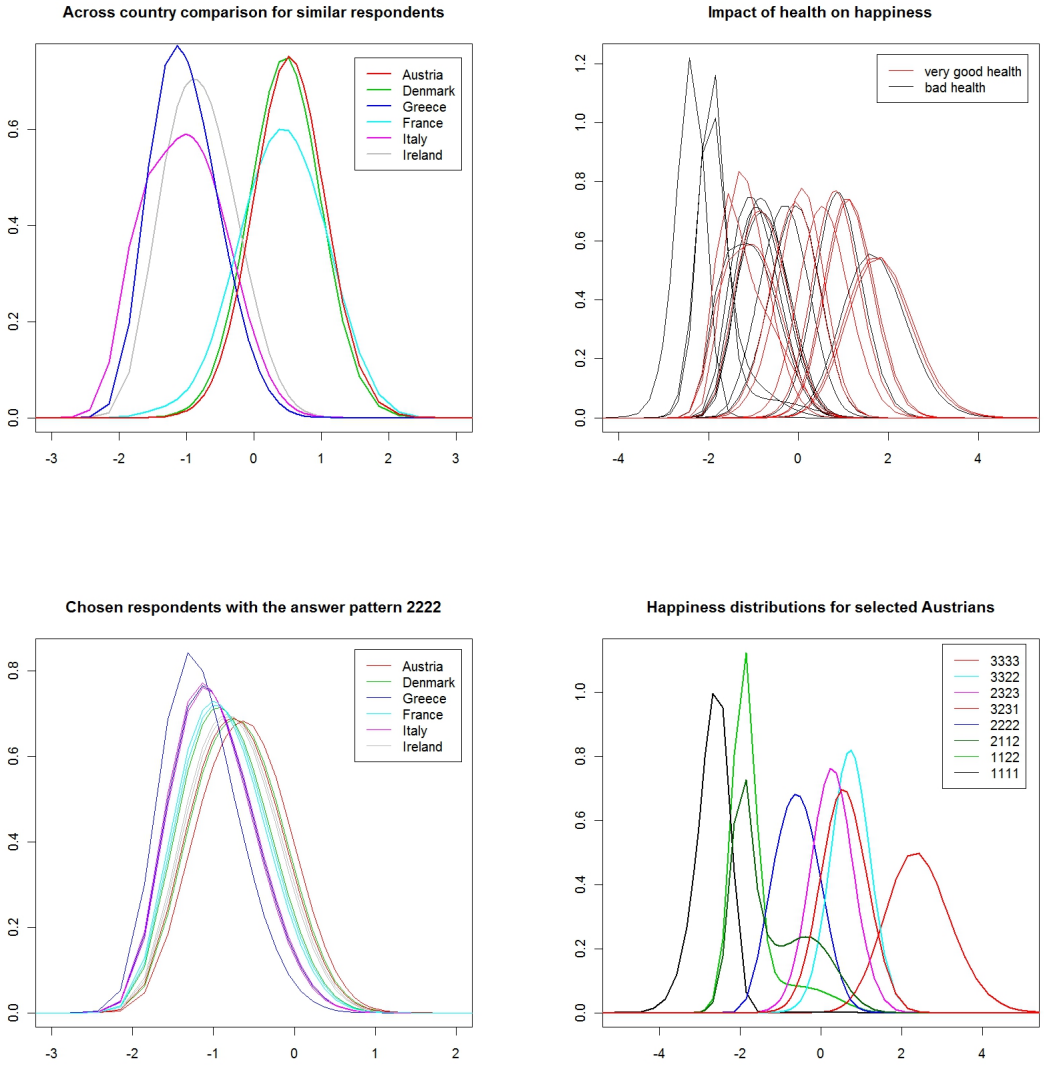
The second plot presents posteriors of 24 respondents - four for each country: two who

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

reported very good health and two bad. The distributions for those with very good health (red color) are shifted more to the right side of the plot, which illustrates the average positive effect of high health self-evaluation.

The next subplot shows the posterior happiness distributions for 12 respondents (2 from each country) who gave the "2"answers to all items. If, as in standard IRT models, only responses to satisfaction questions were used, then there would be just one curve common for everyone.

Figure 2.2. Estimated posteriors of happiness for selected individuals



Finally, the last plot takes a more thorough look at the happiest nation found. The red

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

curve is the posterior estimate for the happiest Austrian and the happiest respondent in the whole sample at the same time (measured by the expectation of theta: $\int \theta \ f(\theta \mid r, W)d\theta$). This is a married woman, of middle education, good health, having social life and an income higher than her cohort average, 58 years old, working slightly less than the sample average. The black curve indicates the posterior estimation for the least happy Austrian - a married male, aged 34, of middle education, being social, reporting bad health and an income below his comparison value. The curves in-between are the posteriors for selected Austrians with different response patterns, as assigned in the legend. Their location shifts are in line with expectations, i.e. higher answers are linked to the posteriors located more to the right.

2.6 Relative social status and materialistic values

The above - stated results show that relative income and, thus, relative social status have a significant impact on an individual's well-being. However, it is possible that the feeling of relative deprivation changes the role and importance of other factors. Does earning a higher income reduce our happiness from a serious relationship? Can it diminish the negative well-being consequences from getting divorced? And can we expect richer people to be less influenced by their health problems, since they may feel more financial security if they need to face sudden expenses? In order to answer these and other questions, the IRT analysis was run for two groups obtained from dividing the sample according to the value of the relative income dummy. "The relatively poorer"are those who report lower incomes and "the relatively richer"higher incomes than the average of each individual's cohort. The baseline respondent is the same for the two subsamples: as before, it is a married French woman, with middle level of education and enjoying good health, not having an intensive social life, aged 39, working 38 hours a week and with net earnings of 1153 equivalent units. The descriptive statistics for both groups can be found in the Appendix (Table 2.9) and show that the lower social status subsamples are composed of higher fractions of women, low educated people and those who evaluate their health lower. They usually work less and there are more single or divorced individuals than in "the relatively richer"group. The estimated coefficients presenting the role of individual characteristics in shaping the happiness distribution are reported in Table 2.6 below.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

Table 2.6. IRT results for the subpopulations based on the relative income

	The "poorer" subgroup			The "richer" subgroup		
	coef	**s.e.***	**p-value**	**coef**	**s.e.***	**p-value**
single	-0.0250	0.0378	0.5086	-0.0007	0.0427	0.9868
divorced	-0.2922	0.0520	0.0000	-0.2200	0.0701	0.0017
cohabitation	0.0324	0.0491	0.5100	0.0365	0.0547	0.5046
male	-0.2037	0.0299	0.0000	-0.0965	0.0349	0.0057
highedu	0.0001	0.0430	0.9976	0.0516	0.0349	0.1394
lowedu	-0.1094	0.0363	0.0025	-0.0573	0.0423	0.1754
vhealth	0.2223	0.0326	0.0000	0.3036	0.0349	0.0000
bhealth	-0.4371	0.0366	0.0000	-0.5162	0.0422	0.0000
social	0.3421	0.0376	0.0000	0.2295	0.0377	0.0000
income	0.3367	0.0555	0.0000	0.8119	0.0706	0.0000
hours	-0.4958	0.0610	0.0000	-0.6025	0.0763	0.0000
age	0.0101	0.0017	0.0000	0.0064	0.0021	0.0024
age2	0.0661	0.0130	0.0000	0.0064	0.0156	0.0000
denmark	0.3594	0.0539	0.0000	0.4455	0.0504	0.0000
ireland	-0.4937	0.0687	0.0000	-0.6885	0.0667	0.0000
italy	-0.9058	0.0489	0.0000	-0.3956	0.0493	0.0000
greece	-1.2873	0.0583	0.0000	-0.6912	0.0651	0.0000
austria	0.5165	0.0551	0.0000	0.7816	0.0524	0.0000
No. of obs.	9365			7221		
LLF	29956.90			21597.94		

*The reported standard errors are the robust estimators

It seems that possessing a higher social status among peers has a "buffering"effect: most of the coefficients have smaller absolute values. The well-being of "the richer"is less influenced by family life. For "the relatively poorer", cohabitating relationships contribute more to the increase in happiness and at the same time they suffer more where there is divorce or they stay single. This shows that reaching a higher status is connected with a higher level of materialism and a decrease in the importance of the interpersonal relationships. This is also confirmed by the lower level of "social"variable coefficient - seeing family and friends more

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

often seems to play a bigger role for the poorer. Although some studies find (e.g. Alesina et al. 2004) that an absolute income is relatively more important for the poorer, here we obtain an opposite result: giving the same amount of money to two people earning the same, but being in the higher and lower social status groups respectively, brings a greater increase in satisfaction to the former one. This finding again confirms the materialistic views of the richer parts of European societies.

When we look at the results concerning the effect of gender, there is almost a twofold difference between both groups, with the decreasing gender discrepancy for the richer group. This could be explained by the social role of men, who are supposed to be the breadwinners in the households and to provide the families with financial stability. Therefore, the feeling of relative deprivation has much stronger negative effect for men than for women. The positive effect of high education on well-being appears in both groups, but for the poorer one is of a much smaller magnitude and is not significant. It is possible that higher expectations and then likely disappointment with earnings being lower than those of peers, despite being a graduate, drives this result. On the other hand, the negative effect of low education decreases for the richer group, which could mean that lower feelings of self-esteem (possible lack of abilities, willingness to study, lower social background) often matter less if earnings are relatively high.

As far as the age and hours worked effects are concerned, there are no significant differences between the two subsamples. However, there is a big discrepancy between the estimated coefficients related to self-evaluated health, with their greater absolute values obtained for the richer subgroup. Health seems to be a bigger concern for individuals when they reach higher social status, which could be because of the relatively smaller importance and severity of problems related to housing or financial issues than for the poorer. The elimination of such problems might cause an increase in concern for those that are being beyond the own control.

Finally, the comparison of the country - fixed effects provides the conclusion that relative social status position is very important in Southern Europe. The negative shift in location on the happiness scale in relation to the baseline French is much smaller for the richer group. The negative effect for Italy becomes even smaller than for Ireland, which is here the only example of a country where being in a higher social status group is actually connected to a drop in average happiness. The explanation for this is not clear, but might be connected to the strong Catholic tradition in this county.

Another way of looking at these results is to compare compensating differentials, i.e. the amounts of money that would bring the same mean location shift in the latent happiness

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

distribution as the respective variable of interest. In other words, these are the changes in income that are required to exactly off-set a particular life occurrence.[27] For instance, from Table 2.7 we see that for the baseline respondent earning 1153 equivalent units, having a partner brings the same increase in happiness as a rise in income by 116 and 53 units for a respectively "relatively poorer"and "richer"individual. Again, the materialistic views of the richer subpopulation are clearly visible here, since the obtained values of compensating differentials are significantly lower for this group. This is due to two parallel effects: the lower importance of the respective issues represented by chosen variables and the higher income valuation.

Table 2.7. Compensating differentials

	poorer subsample		**richer subsample**	
		baseline respondent		
	860 net	**1153 net**	**1153 net**	**1534 net**
single	-61.45	-82.40	-1.01	-1.34
divorced	-498.91	-669.02	-273.72	-363.98
cohabitation	86.75	116.32	53.05	70.54
male	-390.37	-523.47	-129.27	-171.90
high education	0.34	0.45	75.65	100.60
low education	-238.59	-319.94	-78.64	-104.58
v. good health	804.22	1078.42	522.90	695.32
bad health	-625.18	-838.34	-542.59	-721.52
sociable	1515.34	2031.99	376.79	501.03
hours*	-142.29	-190.80	-100.47	-133.60
age**	188.99	253.43	48.35	64.29

*Calculations for an 5 hours increase (from the sample mean) in the weekly working time

**Calculations for an increase in age by 5 years

As we can see in Table 2.7, all the numbers obtained for the baseline respondent are much higher for the relatively poorer subgroup, meaning that it is more difficult "to buy"their happiness. The higher layer of the society can compensate much more easily for some negative life occurrences or for a lack of the positive ones. Even the health issues seem to

[27]Following Clark and Oswald (2002).

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

off-set more easily, although the coefficients corresponding to the location shifts were greater for the richer subgroup, but the difference in income valuation appears to be high enough to drive such results. Only high education seems not to matter for the lower-income group and has a positive effect for the relatively richer group. However, the differences in compensating differentials diminish if the computed values for the two hypothetical respondents with the baseline characteristics and incomes fixed at the mean values of the respective groups are compared. Apart from the "bad health" occurrence, the differences remain of the same sign.

Table 2.8. Estimated probabilites of answers for different latent trait values

		main activity			**finances**			**housing**			**leisure**		
theta	**group**	1	2	3	1	2	3	1	2	3	1	2	3
-1.85	poor	0.21	0.76	0.02	0.44	0.56	0.00	0.07	0.85	0.08	0.20	0.75	0.05
	rich	0.21	0.77	0.02	0.35	0.65	0.00	0.25	0.71	0.05	0.47	0.50	0.03
-1	poor	0.03	0.82	0.15	0.11	0.87	0.02	0.03	0.63	0.34	0.10	0.72	0.19
	rich	0.02	0.86	0.12	0.05	0.94	0.01	0.02	0.73	0.25	0.14	0.72	0.14
0	poor	0.03	0.43	0.54	0.10	0.73	0.17	0.03	0.26	0.71	0.10	0.52	0.39
	rich	0.02	0.48	0.51	0.04	0.79	0.17	0.02	0.38	0.60	0.12	0.65	0.24
1	poor	0.03	0.10	0.87	0.09	0.29	0.62	0.03	0.07	0.90	0.10	0.52	0.63
	rich	0.02	0.11	0.87	0.04	0.26	0.69	0.02	0.12	0.86	0.11	0.49	0.39
1.85	poor	0.02	0.02	0.96	0.05	0.07	0.88	0.03	0.01	0.96	0.08	0.08	0.84
	rich	0.01	0.02	0.97	0.03	0.05	0.92	0.01	0.03	0.96	0.06	0.19	0.74

The last set of results presented here (Table 2.8) are the estimated probabilities of satisfaction answers for a given position on happiness scale. Figure 2.4 in the Appendix presents the item characteristic curves for both "relatively richer" and "relatively poorer" subgroups. Satisfaction with the finances item is more likely to be evaluated as higher for the "relatively richer" individuals. However, the respective probabilities for satisfaction with the main activity are similar for both subsamples. For the housing and leisure items, the situation reverses: the "relatively" poorer are more likely to give higher scores. Bearing in mind that this group has more limited financial resources, their ability to be happy with these life domains, where more money improves the standard of living and allows for plans to be realized, speaks in favor of their positive attitudes towards what they have and what they can attain.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

2.7 Discussion

This work presents the importance of different factors in shaping happiness. Concerns for family, health and economic issues are generally regarded as the top three worries in our lives, which is also confirmed by the results presented in this paper. Moreover, also gender, social life and high education are of big importance. However, the findings indicate that there are some trade-off effects basing on the social status: higher (relative) incomes decrease the role of family life and other different life occurances in shaping the individual's happiness. According to the provided results it is only the health health status that has a higher impact on the richer group welfare, but generally they are characterized by more materialistic values.

Modern European societies, represented here by the Austrians, Danes, French, Irish, Italians and Greeks, may be ranked (with the given order) according to their general levels of happiness. This ordering might be surprising, especially with the low satisfaction scores for the citizens of the sunny Mediterranean areas. However, Austria and Denmark represent small countries, characterized by a higher degree of equality and a strong welfare state tradition, which probably significantly influences the well-being of the citizens.

The semi-parametric IRT model applied here, allows us to account for both answers to different life domain satisfaction questions and individual characteristics in shaping the individual's happiness. The obtained results should not be interpreted as causality statements, but rather as conditional judgements. Nevertheless, the estimation strategy presented here allows for dealing with several observed dependent variables at the same time assumed to be a manifestation of the unobserved one. Moreover, the semi-parametric specification allows for a higher level of flexibility without imposing some rigid assumptions. However, since it is impossible to control for all the factors shaping happiness, panel data methods should be introduced. Joining the panel technics with the semi-parametric IRT model presented here is the next challenge in happiness techniques field.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

References

[1] Alesina, Alberto, Di Tella, Rafael and MacCulloch, Robert, Inequality and happiness (2004): are Europeans and Americans different?, *Journal of Public Economics* Vol. 88 (9-10), pp. 2009 - 2042.

[2] Clark, Andrew E., Diener, Ed, Georgollis, Yannis and Lucas, Richard E. (2003): Lags and leads in life satisfaction: a test of the baseline hypothesis, DELTA Working Paper No. 2003-14.

[3] Clark, Andrew E., Kristensen, Nicolai and Westergard-Nielsen, Niels (2009): Job Satisfaction and Co-worker Wages: Status or Signal?, *The Economic Journal*, Vol. 119, No. 536, pp. 430-447.

[4] Clark, Andrew E., Kristensen, Nicolai and Westergard-Nielsen, Niels (2008): Economic Satisfaction and Income Rank in Small Neighbourhoods, IZA Discussion Paper, No. 3813.

[5] Clark, Andrew E. and Oswald, Andrew J. (1996): Satisfaction and comparison income, *Journal of Public Economics*, Vol. 61 (3), pp. 359-381.

[6] Clark, Andrew E. and Oswald, Andrew J. (1994): Unhappiness and unemployment, *The Economic Journal*, Vol. 104, No. 424, pp. 648-659.

[7] Clark, Andrew E. and Oswald, Andrew J. (2002): Well-being in panels, Working paper 2002, downloadable from *http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/oswald*.

[8] Deaton, Angus (2007): Income, aging and wellbeing around the world: evidence from the Gallup World Poll, NBER Working Paper 13317.

[9] Diener, Ed and Suh, Eunkook M. (editors, 2000): Culture and subjective well-being, The MIT Press, Cambridge Massachusetts.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[10] Di Tella, Rafael and MacCulloch, Robert J. (2008): Gross national happiness as an answer to the Easterlin Paradox?, *Journal of Development Economics*, Vol. 86 (1), pp. 22-42.

[11] Di Tella, Rafael, MacCulloch, Robert J. and Oswald, Andrew J. (2003): The macroeconomics of happiness, *The review of Economics and Statistics*, Vol. 85 (4), pp. 809 - 827.

[12] Dolan, Paul, Peasgood, Tessa and White, Mathew (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*, Vol. 29 (1), pp. 94-122.

[13] Easterlin, Richard A. (1995): Will raising the incomes of all increase the happiness of all?, *Journal of Economic Behavior & Organization*, Vol. 27 (1), pp. 35-47.

[14] Ferrer-i-Carbonell, Ada (2005): Income and well-being: an empirical analysis of the comparison income effect, *Journal of Public Economics*, Vol. 89 (5-6), pp. 997-1019.

[15] Ferrer-i-Carbonell, Ada and Van Praag, Bernard M.S. (2004): Happiness Quantified: a Satisfaction Calculus Approach, Oxford University Press.

[16] Frey, Bruno S. and Stutzer, Alois (2002): What can economists learn from happiness research, *Journal of Economic Literature*, Vol. 40 (2), pp. 402-435.

[17] Frey, Bruno S. and Stutzer, Alois (2006): Does marriage make people happy, or do happy people get married, *The Journal of Socio-Economics*, Vol. 35 (2), pp. 326-347.

[18] Korpi, Tomas (1997): Is utility related to employment status? Employment, unemployment, labor market policies and subjective well-being among Swedish youth, *Labour Economics*, Vol. 4 (2), pp. 125-147.

[19] Luttmer, Erzo F. P. (2005): Neighbors as negatives: relative earnings and well-being, *The Quarterly Journal of Economics*, Vol. 120 (3), pp. 963 - 1002.

[20] McBride, Michael (2001): Relative-income effects on subjective well-being in the cross-section, *Journal of Economic Behavior & Organization*, Vol. 45 (3), pp. 251 - 278.

[21] Rabe-Hesketh, Sophia and Skrondal, Anders (2004): Generalized latent variable modelling: multilevel, longitudinal, and structural equation models, Princeton University Press.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[22] Rizopoulos, Dimitris (2005): ltm: An R package for latent variable modeling and item response theory analyses, *Journal of Statistical Software*, Vol. 17 (5), pp. 1-25.

[23] Rojas, Mariano (2007): Heterogeneity in the relationship between income and happiness: A conceptual-referent-theory explanation, *Journal of Economic Psychology*, Vol. 28 (1), pp. 1-14.

[24] Spady, Richard H. (2006): Identification and Estimation of Latent Attitudes and their Behavioral Implications, CEMMAP working paper CWP12/06.

[25] Spady, Richard H. (2007): Semiparametric Methods for the Measurement of Latent Attitudes and the Estimation of their Behavioral Consequences, CEMMAP working paper CWP26/07.

[26] Stevenson, Betsey (2008): Happiness inequality in the United States, *NBER Working Paper Series*, No. 14220.

[27] Stutzer, Alois (2004): The role of income aspirations in individual happiness, *Journal of Economic Behavior & Organization*, Vol. 54 (1), pp. 89-109.

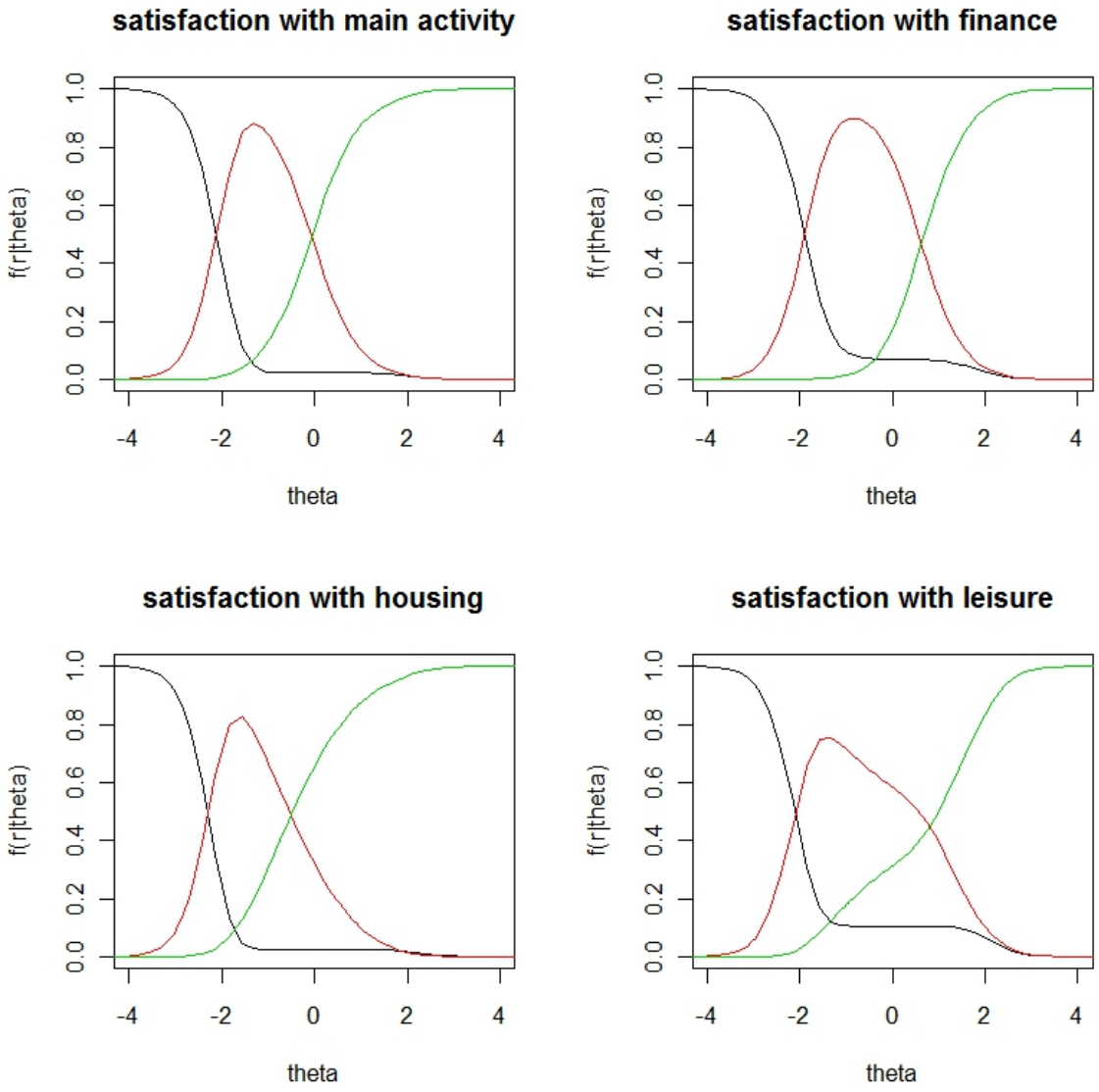
[28] Terrell, George R. (1999): Mathematical statistics: a unified introduction, New York, Springer-Verlag.

[29] Tomes, Nigel (1986): Income distribution, happiness and satisfaction: a direct test of the interdependent preferences model, *Journal of Economic Psychology*, Vol. 7 (4), pp. 425-446.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix A: Tables and Figures

Figure 2.3. Item category response function



Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 2.9. Descriptive statistics for two subsamples (1)

	All		Austria		Denmark		France	
	rich	**poor**	**rich**	**poor**	**rich**	**poor**	**rich**	**poor**
	Frequencies (%)							
married	63.84	59.44	58.47	56.70	63.70	61.78	62.44	54.52
divorced	4.46	6.39	5.68	6.74	8.80	8.74	5.44	8.44
cohabitation	9.43	9.94	10.66	8.30	18.15	17.65	17.31	19.92
male	69.46	45.96	76.20	41.58	67.66	37.87	64.13	46.76
high education	37.28	13.27	17.53	5.70	52.26	22.96	52.20	17.12
low education	20.61	43.52	3.98	18.16	5.28	16.28	35.76	69.84
v. good health	38.50	34.60	47.31	46.18	53.69	45.24	13.06	10.68
bad health	15.93	20.78	8.37	12.74	9.90	17.40	27.87	34.28
sociable	81.76	80.67	74.40	71.49	75.69	80.90	76.21	71.40
	Sample averages							
age	38.71	38.54	37.94	37.47	41.29	40.80	39.22	38.70
hours worked	40.26	36.72	42.09	35.66	41.32	35.80	39.03	36.12
net income	1533.50	860.00	1422.22	797.13	1530.65	993.59	1729.32	879.57
no of obs.	7221	9365	1004	1217	909	1167	1837	2500

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 2.10. Descriptive statistics for two subsamples (2)

	Greece		Italy		Ireland	
	rich	**poor**	**rich**	**poor**	**rich**	**poor**
	Frequencies (%)					
married	68.43	61.42	68.18	65.67	57.58	53.69
divorced	2.90	4.51	2.30	4.57	2.21	4.45
cohabitation	1.66	1.90	2.52	3.06	4.27	3.56
male	70.19	53.70	72.62	50.08	66.86	36.90
high education	45.76	14.13	20.81	7.21	38.29	15.39
low education	14.70	40.04	26.18	46.40	18.11	36.64
v. good health	79.40	75.25	20.76	18.82	63.48	57.12
bad health	3.52	5.19	21.36	25.31	5.89	7.51
sociable	89.03	90.60	85.54	84.45	95.29	96.06
	Sample averages					
age	38.23	37.47	38.32	38.97	36.78	36.80
hours worked	41.14	40.70	39.13	36.85	40.86	34.36
net income	1188.64	666.91	1344.02	838.25	2172.65	1084.03
no of obs.	966	1309	1826	2386	679	786

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

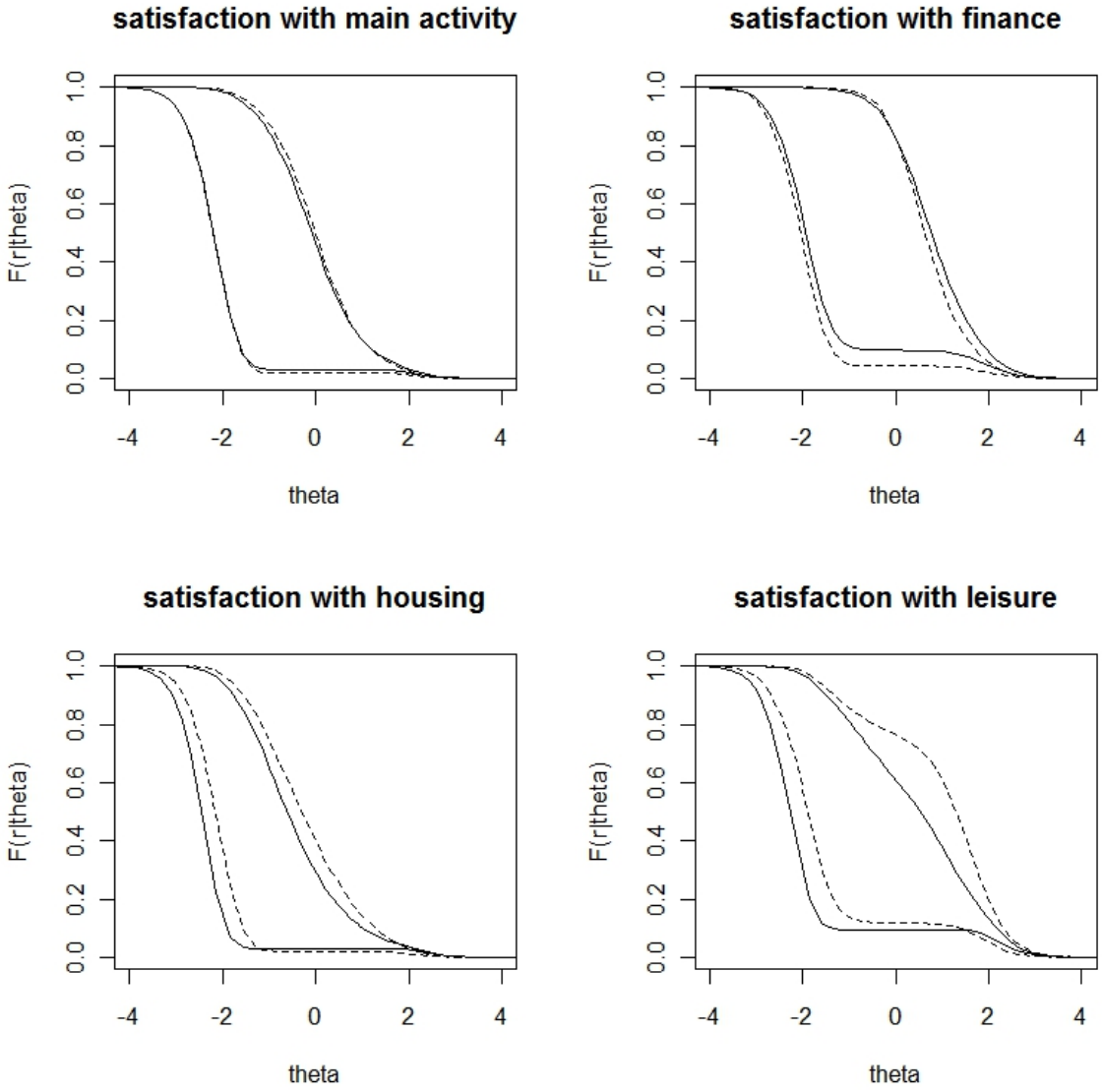
Table 2.11. Estimated tilting parameters

			All (Specification 2)		"the relatively poorer"		"the relatively richer"	
			estimate	s.e.*	estimate	s.e.*	estimate	s.e.*
activity	$F(r_i = 1 \mid \theta)$	t1	-1.9887	0.0488	-1.8245	0.0550	-2.0735	0.0665
		t2	6.9982	1.1913	8.6429	2.1272	8.3467	1.8590
	$F(r_i = 2 \mid \theta)$	t1	-0.1493	0.1345	-0.2430	0.1741	-0.0657	0.1473
		t2	-0.4843	0.0868	-0.3138	0.1150	-0.4433	0.1264
finances	$F(r_i = 1 \mid \theta)$	t1	-1.4379	0.0328	-1.2157	0.0360	-1.6741	0.0514
		t2	4.4920	0.6347	5.2291	1.0807	5.9699	1.2607
	$F(r_i = 2 \mid \theta)$	t1	1.8203	0.1949	1.6362	0.2254	1.9435	0.2717
		t2	-0.8075	0.1122	-0.5017	0.1427	-1.0084	0.1778
housing	$F(r_i = 1 \mid \theta)$	t1	-1.8886	0.0372	-1.7928	0.0437	-2.0317	0.0642
		t2	11.4236	2.1977	17.1373	4.8522	7.5641	1.5391
	$F(r_i = 2 \mid \theta)$	t1	-0.6627	0.0870	-0.9197	0.1237	-0.4299	0.0996
		t2	0.3718	0.0765	0.3302	0.1258	0.3022	0.1054
leisure	$F(r_i = 1 \mid \theta)$	t1	-1.1246	0.0241	-1.1623	0.0335	-1.0995	0.0297
		t2	7.9979	1.8039	11.5463	3.8001	4.9332	0.8675
	$F(r_i = 2 \mid \theta)$	t1	0.4487	0.0494	0.2240	0.0893	0.6816	0.0505
		t2	1.5200	0.0809	1.0004	0.0927	1.9182	0.1584

*Reported standard errors are the robust estimates

DOI: 10.2870/38699

Figure 2.4. Estimated item characteristic curves for poorer (solid lines) and richer (dashed lines) subsamples



Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix B: Estimation of item characteristic curves

The item characteristic curves are estimated by exponential titlting, which is commonly used as a technique of numerical (saddlepoint) approximation of distributions from the general exponential family. The main concept of tilting is based on relating the density of interest to another density. Following Terrell (1999), the tilted version of the distribution $f(x)$ is:

$$f_t(x) = \frac{e^{-xt} f(x)}{\int e^{-Xt} f(x) dX} \tag{2.4}$$

The choice of denominator is such to make the tilted density integrate to 1. The families of log-densities approximating the unknown density comprise polynomials, splines or trigonometric series.

Exponential tilting (ET) provides a good fit and a big flexibility of shapes of the obtained curve introducing a small number of parameters. This is illustrated in Figure 2.5, where the fitted curves for the artificially generated data are presented. The *grm* black line corresponds to the 2-parameter logistic specification as in the Graded Response Models. The other 3 ET lines are the respective integrals of exponential tilting estimates for different choices of the number of parameters. We can see that the two parameter ET provides already a better fit than the logistic specification. Moreover, ET is regarded to be more robust to misspecification problems. It is, however, not recommended to approximate heavy tailed distributions.

As explained in section 4, item characteristic curves (*icc*) are a collection of downward sloping and not crossing curves. The first feature suggests choosing a hazard function as a general specification, that is $1 - G_i(\theta)$, where $G_i(\theta)$ is a cumulative distribution function. Since the IRT framework does not aim at imposing any specific assumptions on the parametric family (usually taken as logistic or normal) the iccs should belong to, the exponential tilting technique is applied here. Specifically, $G_i(\theta)$ is estimated as the *cdf* corresponding to an exponential tilt of the uniform density with two tilting parameters and shifted Legendre polynomials as the basis function. The value of function G_i for any θ is given by:

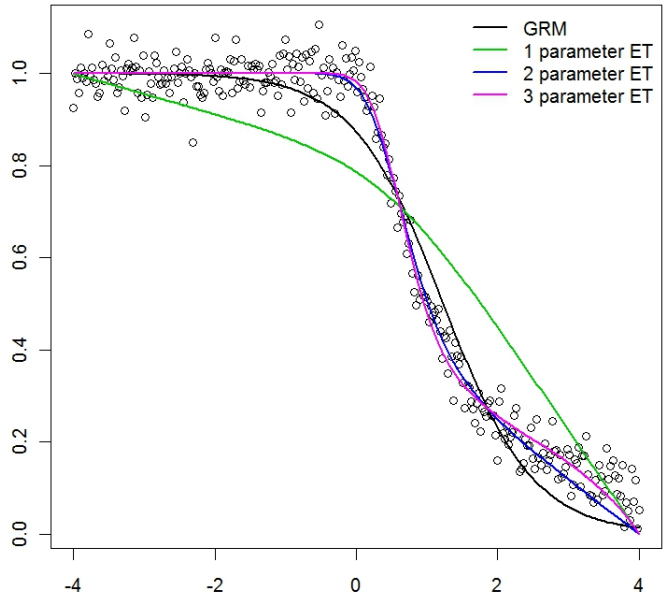
$$G_i(\theta) = \int_0^\theta \frac{e^{t_1(6u^2-6u+1)+t_2(2u-1)} du}{\int_0^1 e^{t_1(6u^2-6u+1)+t_2(2u-1)} du} \tag{2.5}$$

This function is evaluated for θ from $[0, 1]$, which is imposed by introducing the shifted Legendre polynomials, which are orthogonal on the unit interval. The results are than scaled to cover the whole interval corresponding to the support of theta. The estimates of the tilting

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

parameters for the each items' *iccs* are presented in Table 10 (Appendix 1).

Figure 2.5. The comparison of logistic and ET fits



However, G functions obtained here need some manipulation to act as the *iccs* and to fulfill the stochastic dominance condition. The highest item characteristic curve, i.e. a probability of answer 3 or less to item i: $F(r_i = 3|\theta)$, is by definition a straight "1"line. The next curve is taken as $F(r_i = 2|\theta) = 1 - G_2(\theta)$, and thus the monotonicity assumption is enforced. However, constructing the next curve as $F(r_i = 1|\theta) = 1 - G_1(\theta)$ does not assure that this curve lies below the previous one. Therefore, it is obtained as: $F(r_i = 1|\theta) = (1 - G_1(\theta)) * F(r_i = 2|\theta)$. Since the first factor takes on positive values no greater than 1, the curves do not cross and preserve the stochastic dominance assumption.

Technically, the numerical integration is carried out using Gauss-Legendre quadrature with 72 grid points (for evaluating the integrals in (5) almost 600 points), spaced unevenly in 7 segments, with the increasing number of grid points toward the middle of the interval $[-8, 8]$.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

CHAPTER 3

Framing effects and the latent trait measurement: An analysis of the self-reported happiness changes

Abstract

This work summarizes and extends the quantitative analysis of framing effects, which probably concern a vast majority of statistical surveys. Framing effects may bias the inference on the phenomena measured and cause the results incomparability, especially if the population under study is exposed to framing to varying degree. Framing effects are analyzed here parallelly with the differential item functioning concept that is widely studied in the psychometric literature. The contribution of this paper is mainly methodological. This work presents an extended semiparametric IRT framework allowing to model flexibly the potential differences in the response formation process. The empirical analysis is conducted for the 1986 wave of the General Social Survey, which implements an experiment verifying the significance of question order in shaping the general happiness answers. Finally, the discussion about the results reliability in the light of framing effects is presented.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

3.1 Introduction

The latent trait measurement is a broad and still expanding research field in different areas of science, including social and political sciences, cognitive psychology, statistics and economics. At the same time, the latent trait measurement addresses many concerns about the empirical and methodological appropriateness of the survey design and results validity. It is certainly difficult to capture the respondents' non-observed attitudes or beliefs and it may be even more difficult to assure their comparability.

Each survey should consist of understandable and non-confusing questions. Moreover, there should be no factors giving potential incentives to the interviewees to report values different from the real ones or to refuse to answer a question at all. These rules are normally applied to any of the survey questions in order to make sure that even the reported socio-demographic characteristics are correct (see e.g. *Handbook of Recommended (...)*). The latent trait measurement attempts require even more precaution and detailed survey preparation work, so that any undesired and measurement error enhancing influences on the respondents are eliminated.

However, different studies and survey analyses report many cases of response instability, which are often related to seemingly trivial changes in the questionnaire forms. Such phenomenon is referred to as framing and generally describes a situation in which changes in the presentation of an issue produce changes in answers, without changing beliefs or attitudes. This effect is usually achieved by altering the weight of particular considerations, which affects the way how the respondents recall or express their attitudes. Framing can be considered in both positive and negative terms. For example, marketing or political framing can sometimes be viewed as a strategy to manipulate the individuals. On the other hand, framing may be perceived in a positive light when it refers to the respondent's learning process, facilitating the recall of different events necessary to provide a meaningful answer.

This paper presents an econometric strategy to deal with framing effects in case of latent characteristics measurement. Specifically, the empirical analysis is conducted for the General Social Survey 1986 dataset, whose design implements a question ordering experiment. The ordering effect concerns answers to the general happiness question and their relation to the existence of prior questions about satisfaction with different life domains. The simple analysis of response frequencies inclines towards accepting the framing effects significance.

This work provides a thorough analysis of framing effects, addressing the detection, statistical significance verification, the econometric modeling and overall importance issues. Specifically, Sections 2 and 3 of this paper explain the concept of framing and the related

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

concept of differential item functioning, discussing also the examples where systematic groups differences arise. Section 4 provides the description of the data and the IRT methodology used in the analysis. Section 5 introduces methods of framing effects detection and presents the tests results for the GSS experiment. The next Section addresses the overall framing effects importance, attempting to assess the extent to which the results may be unreliable if the framing effects occur. Finally, Section 7 concludes.

3.2 The understanding of framing effects

A comprehensive and illustrative definition of framing effects may be found in Entman (1993): "To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described". Nevertheless, framing should not be confused with changes in beliefs, neither used interchangeably with such expressions as persuasion, nor manipulation. Nelson et al. (1997) explains that frames appear to activate existing beliefs and cognitions, rather than adding something new to an individual's beliefs about the issue.

The literature points out different types of framing effects, however the classification and nomenclature are not standardized. Zaller and Heldman (1992) mentions, for example, the framing effects of question wording and question order, which often coexist with equivalency effects. Ordering effects describe the situation in which we observe the influence of the prior question on the response. Zaller and Heldman (1992) gives the Cold War experiment on attitudes towards the Soviet journalists as an example: the number of students agreeing to allow Soviet journalists into the US doubled when the item was preceded by the question if whether American journalists should be allowed into Russia. Similarly, it has been shown that attitudes towards abortion can be easily influenced by a proper choice of preceding questions, e.g. asking about children, welfare, etc. If the abortion question follows a religion question, we can additionally consider the reference group effect, when the respondents tend to identify themselves with the values assigned to a group they belong to. Equivalency effects exist when logically equivalent but differently phrased questions are not answered alike, which typically involves casting issues in either a positive or a negative light. Chong and Druckman (2007) provides an example where 85% of respondents would allow a hate group to hold a political rally if the question was prefaced with the suggestion "Given the importance of free speech", whereas only 45% were in favor when the introductory phrase was changed to "Given the risk of violence".

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

Changes in the broadly understood survey design may unintentionally influence the responses. Obviously, differences in answers can also arise when the question is posed in a different way from a technical point of view, by applying, for instance, an open-ended formulation or a multiple choice. Conti and Pudney (2010) analyses the effect of survey design on reported job satisfaction. They draw on two quasi-experiments in the BHPS: the use of textual labels for the answer scales and the impact of the interview modes (face-to-face versus self-completion). The results confirm that these factors significantly influence the interviewees' responses. The studies also show the existence of effects of the race/gender of interviewer, which may result in more feminist answers if the interviewer is a woman (Huddey et al. 1997). Smith (1986) summarizes that the changes in the sample universe, sampling method, interviewer training, question wording, item order or coding procedure can distort the survey comparability across time and space. Further, in his 1987 paper (*The Art of Asking...*) he writes: "the choice of what questions to ask and how to ask them determines what answers one gets and what analyses might emerge. Deciding what to cover and how to measure it are basic, routine steps in fielding surveys."

Framing effects are also analysed in the context of psychological phenomena, such as anchoring or a tendency to satisfy. This is the subject of van Soest and Hurd (2008), which argues that the responses are infuenced by cues contained in the question and creates the anchoring effect: *"a psychological explanation is that if people are unsure about the exact amount, the entry point serves as an anchor that provides some information about this amount"*. They also investigate the "yea-saying"- the phenomenon that people have tendency to answer "yes"rather than "no". The second is found to be correlated with respondent's characteristics - more uncertainty leads to more "yea-saying". Toepoel et al. (2009) similarly claims that people with less cognitive sophistication are probably more affected by contextual issues.

The tendency to satisfy, apart from "yea-saying"may also relate to the tendency of choosing the first satisfactory or acceptable response rather than selecting the true answer. In terms of framing effects, the preliminary analysis suggest that mode effects may arise here: internet surveys may enhance more superficial cognitive processes and introduce more satisficing.

Framing effects may also be analysed within the framework of the cognitive model describing the process of answers formation. *Handbook of Recommended Practices (...)* distinguishes five stages of the survey response formation, that is:
0) Encoding: the process of forming memories from experiences;
1) Comprehension: the process of interpreting the question, trying to identify its meaning;

2) Retrieval: recalling information relevant for answering the question from memory;
3) Judgment: the process of combining or supplementing what has been retrieved;
4) Reporting: the process of selecting and communicating an answer.

The "zero"encoding stage can take place a long time before the actual survey, therefore this analysis is limited to the core stages, 1 to 4. Framing may then relate to any of the four stages, sometimes influencing just a single stage, but usually having an impact on several cognitive processes at the same time. Moreover, respondents may shortcut the cognitive process when forming an answer, which can be directed by some cues in the questionnaire. The above mentioned abortion example, framed in the religion context, would probably leave the comprehension stage unaltered, but would certainly have on impact on the information retrieval and judgment. Another example, the gender of interviewer effect, concerns probably only the reporting stage at most.

Moreover, Krosnick and Alwin (1987) points out that the cognitive processes can follow differently depending on the position of the items in the list. Items presented early in a list are likely to be subjected to deeper cognitive processing; by the time a respondent considers the later alternatives, his or her mind is likely to be cluttered with thoughts about previous alternatives that inhibit extensive consideration of later ones. This may be related to the order effects.

A meaningful analysis of the response effects requires an underlying assumption of the respondents possessing well formed attitudes or beliefs. Nevertheless, there exist claims that most citizens carry around in their heads a mix of only partially consistent ideas and considerations (e.g. Zaller and Heldman 1992). This is supported by selected examples of significant response instabilities, when in some repeated interviews only about half of the respondents give the same answer. Certainly, the process of mapping the attitudes onto the set of survey questions is a potential source of many measurement problems and imperfections, also including the whole range of different response effects. However, only the assumption of the stability of underlying true attitudes, general answering consistency or the existence of some answering regularities can allow for any reasonable analysis.

3.3 Framing and differential item functioning

From a modeling point of view, the framing effects mechanism may be classified in the same category as differential item functioning (hereafter DIF). Basically, DIF is defined as a situa-

tion where individuals with the same value of the latent trait[28] have different probabilities of particular answers. In other words, DIF occurs if the item under consideration is measuring a quantity in addition to the one the test was designed to measure, a quantity that both groups do not possess equally (Shealy and Stout in DIF 1993, Ch. 10). This answering effect is extensively studied in the psychometric and educational measurement literature. A typical DIF illustration (e.g. Angoff in DIF 1993, Ch. 1) describes a verbal ability test that includes questions about expressions of Latin origin, which may favor students of Spanish or Italian origin. Often, the gender effect is found significant. The literature provides another verbal ability testing example, when questions about fishing related vocabulary are more likely to be answered correctly by boys. DIF, in this case, may arise due to the gender social role division and again, not because of the differences in real ability. Psychometric studies attempt to assure test fairness or comparability for different subpopulations, as relates to gender, race, ethnicity and minority status. In the ability testing framework, DIF may also be identified for groups of students nested within classrooms, which is usually dealt with by applying the hierarchical modeling strategy.

The term DIF is sometimes used interchangeably with item bias, stressing the fact that some items function differently in different groups of examinees / respondents and may therefore provide systematically biased estimates. Technically, in the psychometric literature the examined groups are called the focal (or minority, disadvantaged) and references groups (majority, advantaged), analogically to the treated and control groups in the causality literature. DIF can be classified as exhibiting uniform and non-uniform effects, with the former implying the item being easier for the whole reference group and the latter indicating that the advantage may exist for only a part of the subpopulation.

The DIF studies deal with the question of how item scores are affected by external variables that do not belong to the construct being measured, with these variables usually being arbitrarily known or assumed. This measurement concern, as already mentioned, has been thoroughly examined in the psychometric literature. Recently, DIF has also drawn the attention of economists, normally skeptical about the comparability of various surveys. The assumption of the response scales being the same across countries, across time and across groups of respondents within a country is often doubtful. Therefore, the measurement in the latter situation has to face also the potential danger of individuals' scale incomparability. In the case of satisfaction questions, higher values of satisfaction scores can be confounding and relate not to better life conditions, but to an optimistic nature and seeing the world through

[28]Generally, it is difficult to speak about the value of the latent trait, since there is no unique and commonly accepted scale on which a latent variable is measured, but through normalization a scale is usually introduced.

rose-colored glasses, which may sometimes be related to the nationality. For instance, someone in country A with life satisfaction at level X reports that they are not satisfied, but a person in country B with the same actual satisfaction would report to be satisfied. This is a motivation for many recent scientific insights.

Kristensen and Johansson (2008) attempts to account for varying perceptions of subjective questions in different countries, concentrating on the job satisfaction measurement. Many studies identify Danes as the most satisfied workers, which raises an important policy question whether working life in other countries should be organized as in Denmark. In order to verify the correctness of the international ranking, Kristensen and Johansson (2008) applies a vignette methodology, asking the respondents how good or bad a set of hypothetical jobs or life situations are. This approach corrects for systematic differences in the way the subjective questions are answered, assuming that each respondent uses the same scale for the vignette and the self-report (response consistency). The estimation of a chopit model, which rescales the ordered probit threshold parameters through anchoring vignettes, provides an answer that Danes just tend to give higher satisfaction scores and it is the Dutch model that is objectively the best in terms of job satisfaction. A detailed description of the vignette methodology and the chopit model may be found in the Appendix of this paper.

Similarly, concern about the cultural differences in survey responses constitutes the motivation for Kapteyn et al.(2007) to investigate the discrepancy in the reported work disability in the Netherlands and the US. The approach is also based on the vignette methodology and draws the conclusion that the observed differences stem to a large extent from incomparable response scales for the nationalities under study. Another analysis by Kapteyn et al. (2009) compares the global life satisfaction in the US and the Netherlands using the self-reports and a battery of vignette questions. They find a difference in the way that satisfaction is reported in the two countries and it is stated that it is not just a uniform scale shift.

Vignettes are often treated as a powerful tool allowing for correction in the way different respondents provide answers. However, the stong assumption underlying the vignette methodology is the response consistency requirement and in many cases its validity is questioned. Gupta et al. (2010) analyzes the dataset, including the self-assessment of health and also an objective measure of health (hand-grip strength). Their model provides better predictions when they do not impose the response consistency assumption and show its violation in many cases. Also Kapteyn et al. (2011) questions the validity of response consistency assumption for different domains of health. This work is based on answers obtained in 2 interviews with the respondents: in the first one the respondents were asked to describe and self-assess their health, whilst in the second interview they were presented with

the replica vignette, giving them the description of their own health as described several months before (i.e. the respondents were shown "their own vignettes").[29] The expectation of the same score for health self-assessment in wave 1 and the replica vignette in wave 2 was, however, not fulfilled. The authors draw the conclusion that the vignette description is not complete, so that there is still room for individual interpretations. This proves that the vignette methodology is still in its infancy.

Independently of the potential drawback of the vignette methodology, there are no vignettes available in many surveys nor does knowledge exist as to which subpopulations answer questions in a systematically different way. In this case, instead of dividing the sample arbitrarily according to some specific characteristics (e.g. by gender or nation), the analysis may first identify groups displaying DIF and only after, the dimension causing the response differences is examined. This strategy is followed by Cohen and Bolt (2005) in the case of ability testing and Clark et al. (2005), which provides evidence that individuals do not transform income into well-being in the same way. The model endogenously divides the observations, in a probabilistic sense, into four separate classes (with distinct demographic and country patterns) differing by the comprehension of satisfaction concept.

To sum up, there exists an analogy between the framing effects and differential item functioning, which justifies the application of the DIF detection and modeling techniques for framing effects analysis. The DIF analysis, in its narrow sense, concerns the unequal chances of some specific subpopulations to give a particular answer. Framing effects usually do not relate to a specific group, but rather to the whole population or sample being surveyed, unless only a part of this population is exposed to framing or an experiment is conducted. In this situation, the analysis of the framing-treated and untreated groups is analogical to the focal and reference groups DIF examination. In both cases, individuals from two groups answer questions differently, not because they differ in the level of the latent trait that shapes the answers, but because there are some additional factors having an impact on their responses. In the case of framing, those factors are external to the respondent and their isolation may be theoretically possible. DIF, however, usually relates to some individual characteristics that cannot be eliminated, like gender or nationality. The modeling framework stays, however, the same: individuals do not provide answers as would result from their underlying attitudes, beliefs or opinions, with this discrepancy being systematic and related to specific personal, survey or environment characteristics. Therefore, the methodology described below may be applied to both cases and the terms DIF and framing effects are

[29]The interviees were also presented with some other vignettes so that they were less likely to recognize that they obtain the description of their own health.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

hereafter used interchangeably in the econometric modeling context.

3.4 Data and methodology

Modeling DIF or framing effects requires a proper and usually non-standard estimation methodology. However, in many cases, the econometric remedy to these problems may still have significant limitations, unless the dataset analysed displays certain properties. An example of such a property is a randomized experiment structure, which allows for the isolation of the considered framing effects from other effects. Various waves of the General Social Survey provide different randomized experiments, with the aim being to identify the impact of isolated factors on the distribution of responses.

The General Social Survey (GSS) monitors social changes in the United States since 1972, constituting the base for a wide range of scientific investigations of the structure and development of American society. A rich set of survey information and documentation, together with the datasets itself, may be found on the project website.[30] The GSS contains a standard "core"of demographic, behavioral and attitudinal questions, which in many cases have remained unchanged since 1972 in order to facilitate time-trend studies. GSS has consistently repeated question wording and tried to standardize other measurement procedures, e.g. by designing the changes in religion and income response categories so that they were collapsable into original categories.

Since the first GSS wave there has been a constant growth in the number of items included in the survey, which could dangerously increase the respondents burden. Therefore, item rotation design was introduced since the beginning of the survey, resulting in selected groups of items appearing in two out of every three surveys. However, such rotation design causes difficulties in keeping the constant order of the items and may result in unintended ordering effects. If there is a suspicion of a response effect, the GSS usually implements methodological experiments based on the split-ballot surveys comparisons. An ordering effect (however, not related to the rotation scheme) concerned the variable "general happiness", whose score suddenly rose between 1972 and 1973. A spilt ballot experiment was conducted in 1978 and the positive influence of the added "marital happiness"variable was confirmed (Smith 1986).

An unexpected drop in "general happiness"scores occurred in 1985 and amounted to 6 percentage points in comparison to 1984, as far as the number of people reporting being "very happy"is concerned. Preliminary analysis indicated the alteration in question order as a potential explanation. From 1973 to 1984 the immediate question order was fixed:

[30] $http://www.norc.org/GSS+Website/$

five satisfaction items (city, hobby, family, friends, health) were followed by the marital happiness and general happiness items. In 1985 the domain satisfaction questions were dropped to make room for some other items, within the questions rotation design. In order to verify the hypothesis an experiment was designed in 1986; half of the sample received the two happiness items (marital and general) immediately after the five satisfaction items, while the other half had the satisfaction items follow the happiness questions, resembling the situation in 1985. The experiment results (Table 3.1) were in line with the expectations: there was an observed difference in the "very happy"scores and Smith (1986) finds it to be statistically significant.

Table 3.1. The general happiness score in the two subsamples

		HAPPINESS SCORE			
FORM		**1**	**2**	**3**	**Total**
		not too happy	pretty happy	very happy	
1	Freq.	66	347	227	640
	%	10.31	54.22	35.47	100.00
2	Freq.	70	393	196	659
	%	10.62	59.64	29.74	100.00
Total	Freq.	136	740	423	1299
	%	10.47	56.97	32.56	100.00

At the same time, while the general happiness item shows differences across groups, the domain items concerning satisfaction with city and hobby are characterized by nearly identical response distributions (see Table 3.4 in the Appendix). Items related to satisfaction with family, friends and health exhibit slightly bigger frequencies of highest answers for the $FORM = 1$ respondents. Therefore, higher scores for general happiness item for them may result from randomization problems, i.e. assigning people who were happier on average to the first group. This hypothesis could be accepted if framing effects were insignificant. If, on the contrary, the testing procedures prove that the question ordering influences the answers, then framing effects may be interpreted as a learning process and would influence the retrieval stage of the cognitive model. It is likely that the domain satisfaction questions, apart from the existing differences in happiness levels, additionally lead respondents to recall a mixture of positive memories and therefore, to score higher on the general satisfaction item.

The experimental design of the 1986 GSS wave constitutes the empirical basis of this paper's analysis. For notational purposes, the $FORM = 1$ subsample refers to the standard

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

GSS question ordering, where the domain satisfaction questions preceded the general happiness question. The value $FORM = 2$ is assigned to the respondents whose questionnaires were replicating the 1985 survey and whose scores are lower on average.[31] Not all of the 1986 respondents were included in the analysis; those who have missing values on the basic variables were dropped. Therefore, the final sample amounts to 1299 observations, instead of an original size of 1470 .[32] The list of variables used in the analysis and their descriptive statistics are presented in the Appendix of this paper (Tables 3.3, 3.4 and 3.5).

The econometric strategy applied here is based on the Item Response Theory (IRT), a methodology of psychometric origin designed to model latent traits. The latent trait modeled here is the happiness variable, measured by six conditionally independent items: five domain satisfaction items and a general happiness item. The first five items measure the satisfaction with city, hobby, family, friends and health. Originally, these items were measured on a 7-point scale, but some categories were merged due to low frequencies and the final analysis is based on a 5-point scale.[33] As presented in Table 3.1 above, general happiness is expressed on a 3-point scale and is left in this form for the IRT estimation. The IRT framework includes items with different ranges of measurement scale, which does not deplete the validity of the results.

The aim of the analysis is to verify whether the ordering effect is significant and in the case it is, to show how to incorporate the differences in answering formation into the IRT model. A set of additional results shows the relation between the latent trait and the examined personal characteristics, as well as between the latent trait and the probabilities of particular answers to any of the six items. Specifically, the econometric model allowing for such an analysis is the semi-parametric IRT methodology developed by Spady (2006, 2007) and extended here to account for DIF. This methodology is based on the quasi - maximum likelihood framework, maximizing the probability of observing a vector of answers $(r_1, r_2, .., r_6)$ conditional on the personal characteristics W, which in turn influence the latent characteristic θ. Mathematically, it is expressed as:

$$p(r_1, r_2, .., r_6 \mid W) = \int p(r_1 \mid \theta)p(r_2 \mid \theta)..p(r_6 \mid \theta)f(\theta \mid W)d\theta \tag{3.1}$$

Let ζ denote the variable that divides the sample into subgroups characterized by an

[31]The item "marital happiness"is left outside the analysis, mainly due to the relatively frequent "not applicable"responses.

[32]The aim of this paper is the framing effects modeling and not handling of the missing data problem nor generalizing the results to the entire population. Therefore, it should not lessen the results validity.

[33]The original 7-point scale comprised the answers: none, a little, some, a fair amount, quite a bit, a great deal and a very great deal.

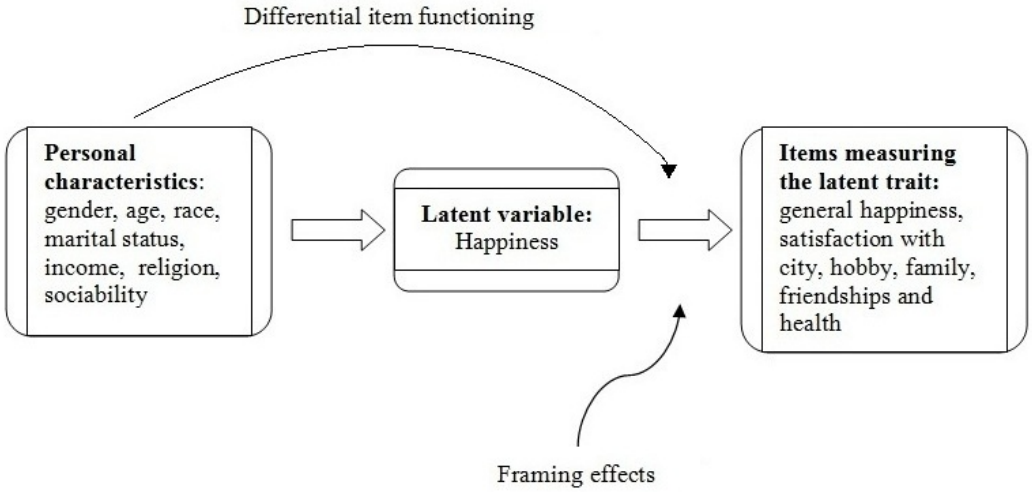
Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

item's differential functioning. In the case of a standard DIF situation, ζ is a personal characteristic, such as gender, race, etc. When the framing effects are considered, ζ captures the exposure of the respondents to some external factors. In the GSS example considered here, ζ corresponds to the $FORM$ variable, capturing the kind of questionnaire the respondent was asked to fill in, and thus the framing treatment status (e.g. treated if $FORM = 2$). Keeping in mind that framing concerns solely the last item, "general happiness", the relationship being modeled between the observables and unobservables may be written as:

$$p(r_1, r_2, .., r_6 \mid W, \zeta) = \int p(r_1 \mid \theta)p(r_2 \mid \theta)..p(r_6 \mid \theta, \zeta)f(\theta \mid W)d\theta \tag{3.2}$$

The intuition behind this mathematical formulation is illustrated in Figure 3.1. The diagram stresses that both DIF and framing effects have an impact only on how the latent variable is mapped onto the satisfaction items measurement space, and not on the level of the latent trait itself nor on the way the latent trait relates to personal characteristics.[34]

Figure 3.1. The semiparametric IRT mechanism



A few assumptions are imposed to ensure the model identification. The first concerns the local independence of items, that is:

[34]An additional arrow could be drawn on this figure pointing from personal characteristics to framing effects, which would capture different degrees of framing effects impact on people of different characteristics (e.g. men being influenced more by the question wording than women). This is, however, not modeled here. The differences in the intensity of framing effects may arise in the considered framework only due to differences in the level of the latent trait.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

$$p(r_1, r_2, .., r_i|\theta, \zeta) = p(r_1|\theta)p(r_2|\theta)..p(r_6|\theta, \zeta) \tag{3.3}$$

The local independence assumption claims that all covariation among the item responses is attributable to the item's individual relationship with the latent trait. The other assumption allows only for the indirect effect of personal characteristics on the responses through the latent trait, unless DIF effects occur and some of the personal characteristics appear in ζ:

$$p(r_1, r_2, .., r_i|\theta, \zeta, W) = p(r_1, r_2, .., r_i|\theta, \zeta) \tag{3.4}$$

Finally, the latent trait itself is modeled here as $N(\mu(W), 1)$ with $\mu(W) = W\beta$, i.e. the distribution of the latent trait for each individual is assumed to be normal with the mean being a linear function of the characteristics and variance equal 1.

The specifications of both $p(r_i \mid \theta)$ and $p(r_i \mid \theta, \zeta)$ are free from parametric assumptions and are estimated using the exponential tilting technique, under the monotonicity assumption. The latter is expressed here in terms of stochastic dominance relations: the responses of individuals with higher values of the latent trait first order stochastically dominate the responses of those with the lower values of the trait. In other words, happier people tend, on average, to give higher scores on the satisfaction question.[35]

The semiparametric IRT framework allows us to model the relations between personal characteristics, the DIF-factor and the latent trait, as well as between the trait and the responses. Similar to the model developed by Adams et al. (1997), the estimation strategy applied here allows for a simultaneous estimation of these relations, rather than introducing a two step procedure. The IRT approach may account for different discrimination power of items (without arbitrarily imposing the weights) and provides as one of the outputs the posterior latent trait distribution for each individual, with its general formulation given by:

$$f(\theta \mid \mathbf{W}, \mathbf{r}, \zeta) = \frac{p(\mathbf{r} \mid \theta, \zeta) f(\theta \mid \mathbf{W})}{p(\mathbf{r} \mid \mathbf{W}, \zeta)} \tag{3.5}$$

3.5 Framing effects detection

There are many cases in which the presence of DIF or framing effects is relatively clear. However, the formal detection and quantification of the effects is often a difficult job. Firstly,

[35]For a detailed description of the methodology, see Spady (2006, 2007) and the second chapter of this thesis.

it should be assured that the assumption of comparing the comparable is fulfilled. This constitutes the basic problem of DIF detection, since the latent trait is not observed and the probabilities of answers are conditioned upon it. In the case of the GSS 1986 survey and the investigation of framing effects, the randomized experiment solves the comparability dilemma to a large extent. Since framing effects do not alter the latent trait, the probabilities of answers given the position on the latent scale should differ for the experimental treatment and control samples due to framing, if framing effects are significant.

There exists a variety of methods aiming at formal DIF detection, which may generally be divided into two categories: IRT and non-IRT based. Howard (in DIF, 1993, Ch. 6) classifies the Mantel-Haenszel statistic, logistic regression methods and standardization procedures into the latter group (called also empirical methods). The model-based procedures comprise, among others, the general IRT likelihood ratio, limited information IRT-LR, or full-information IRT-D^2.

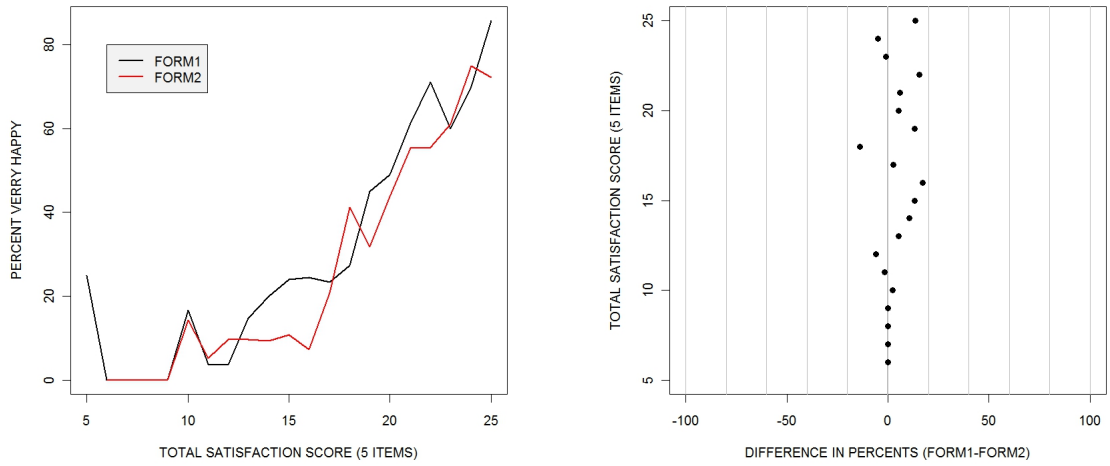
The empirically based procedure refers, first of all, to the analysis of the answers distribution, as presented in Table 3.1. The two-proportions Z-test is designed to determine whether the difference between two proportions is significant, with the null hypothesis stating the equality of two proportions. Since the biggest difference in answering pattern occurs for the highest happiness score, the test is conducted for the share of respondents giving "3"as a response versus the share of those giving "1"or "2"as a response. The calculated Z-statistic amounts to 2.2 and allows for the rejection of H0 at the significance level α=0.05. However, the test fails to reject the null hypothesis when α is decreased to 0.01. The difference in the response pattern seems to be not large enough to unambiguously reject the null hypothesis.

Similar results may be obtained by applying the Mantel-Haenszel chi-square test (e.g. Angoff in DIF, 1993, Ch. 1, or Ayala, 2009, Ch. 12). This procedure, known also as an analysis of the three-way contingency table, allows for the determination of whether two variables are independent of one another while conditioning on a third variable. The null hypothesis implies no DIF, or expressed differently, it implies that the odds of reference group members responding with "1"are the same as those of focal group members. The MH procedure tests the 0-1 answering patterns, therefore similarly as above, the answers 1 and 2 are merged. The choice of the conditioning third variable may be crucial for the analysis and is often selected as the total test score. Here, it is assumed to be the sum of scores on the five domain satisfaction questions. A standardization plot for the dependence between the total score and the general happiness answers is presented below (Figure 3.4). The right subplot shows that the $FORM = 1$ subsample has higher happiness answers on average, but it seems only for the individuals with higher total satisfaction scores. This may be the sign

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

that the response effect has a non-uniform character.

Figure 3.2. DIF detection: matching the happiness score on total score



Several versions of the MH test were computed, matching the individuals from the FORM 1 and FORM 2 subsamples on the sum of five satisfaction answers. The individuals were matched not on each possible result on the single score (ranging from 5 to 25), but were classified into 2, 3 or 4 subsamples, in order to assure the reasonable cell frequencies. The highest test statistic (4.75) was obtained for the test on three subsamples, with the total score classes established to be 5-11, 12-18 and 19-25 respectively. Comparing the MH test statistic to the χ^2 critical values with one degree of freedom allows for the rejection of the "no DIF"hypothesis at the 0.05 significance level, but supports the contrary for $\alpha = 0.01$.

The remaining non-IRT based strategies of DIF detection comprise the logistic regression, where the significance and equality of the variables in separate regressions is verified, potentially giving some indication of DIF. Dorans and Holland (DIF, 1993, Ch. 3) also describes a DIF detection procedure based on the calculation of items' discrepancies indices. The indices they propose are based on a weighting function to average differences across levels of the matching variables. This procedures were not applied to the analysis in this paper.

The detection of DIF or the presence of framing effects may also be conducted by estimating the IRT models and comparing the estimated parameters, the areas between the item characteristic curves, or applying the likelihood based test procedures. Thissen, Steinberg and Wainer (DIF, 1993, Ch. 4) suggests estimating two parametric models for the

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

focal and reference groups and testing whether the parameters are the same or alternatively, whether the differences between trace lines are significant. Since the no-DIF model can be considered as a compact model being nested in the augmented DIF-allowing specification, the likelihood-based tests may be applied.[36] For instance, Glas (1998) suggests a testing method applied to the Rasch IRT formulation based on the Lagrange multiplier test, with the test statistic constructed as a difference between the expected and the observed number of persons in the focal group scoring in a given category. A more commonly applied test is the likelihood ratio test, because of easier implementation by comparing the fit of the augmented and restricted IRT models. The null hypothesis tested is of no group differences in the item parameter estimates, i.e. the DIF does not occur. The test statistic is asymptotically distributed as a χ^2 with degrees of freedom equal to the difference in the number of parameters estimated in the augmented and compact models.

Several model specifications underwent LR testing. The results are presented in Table 3.2.

Table 3.2. Results of IRT model estimation and DIF determination

	Model 1		**Model 2**		**Model 3**	
	coef	p-value	coef	p-value	coef	p-value
married	0.3817	0.0000	0.3851	0.0000	0.3855	0.0000
divorced	-0.1195	0.2146	-0.1194	0.2149	-0.1194	0.2142
widowed	0.1061	0.4521	0.1084	0.4444	0.1076	0.4463
male	-0.1252	0.0394	-0.1244	0.0406	-0.1241	0.0409
black	-0.3833	0.0001	-0.3860	0.0000	-0.3854	0.0000
age	-0.0019	0.5131	-0.0019	0.5068	-0.0019	0.5045
age2	0.0307	0.0070	0.0306	0.0072	0.0307	0.0070
income	0.0465	0.0000	0.0467	0.0000	0.0466	0.0000
religiosity	0.3525	0.0000	0.3525	0.0000	0.3528	0.0000
sociability	0.3128	0.0000	0.3130	0.0000	0.3132	0.0000

Notes:

Model 1: DIF not allowed (restricted model); llf=10207.42595

Model 2: DIF allowed for each category in HAPPY item; llf=10204.2124

Model 3: DIF allowed for answers 3 vs 2 in HAPPY item; llf=10204.4080

The estimation of one additional curve implies an increase of two in the number of

[36] Alternatively, the no-DIF model may be regarded as a restricted model, where the constraint enforces the parameters equality in two groups.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

model parameters. Therefore, there are two models estimated that allow for DIF. The first specification allows DIF for any answer (for both curves modeled), but the LR test does not allow for the rejection of H0. Since the preliminary analysis of the answer frequencies suggests that DIF mostly concerns the difference in replying at the level "3"or "2", the next estimated model implements DIF only for one item characteristic curve that models the probability of answering in category 3. This allows for a gain of two degrees of freedom and therefore, although the likelihood value remains very similar to the "full"DIF specification, the χ^2 critical values are lower. In this case, we can reject the null hypothesis assuming no DIF at the 0.05 significance level (2 degrees of freedom).

Further, Table 3.2 provides estimates on personal characteristics and shows their robustness across different specifications. The results on the dependence between personal characteristics and happiness are in line with expectations; being married increases the probability of higher answers, while divorce leads to a decrease (widowed is not significant). Males and black respondents have lower levels of happiness, whereas religious or sociable people report higher levels. Income is positively related to happiness and age exhibits a U-shaped relation, with the minimum found for age 43.[37] The preliminary model specification included additional controls for respondent education, children (of different ages) and labor status. However, these variables were found to be strongly insignificant. The results on model robustness to DIF specification are also supported by the comparison of the three sets of item characteristics curves for each specification (Appendix, Figures 3.5, 3.8 and 3.7).

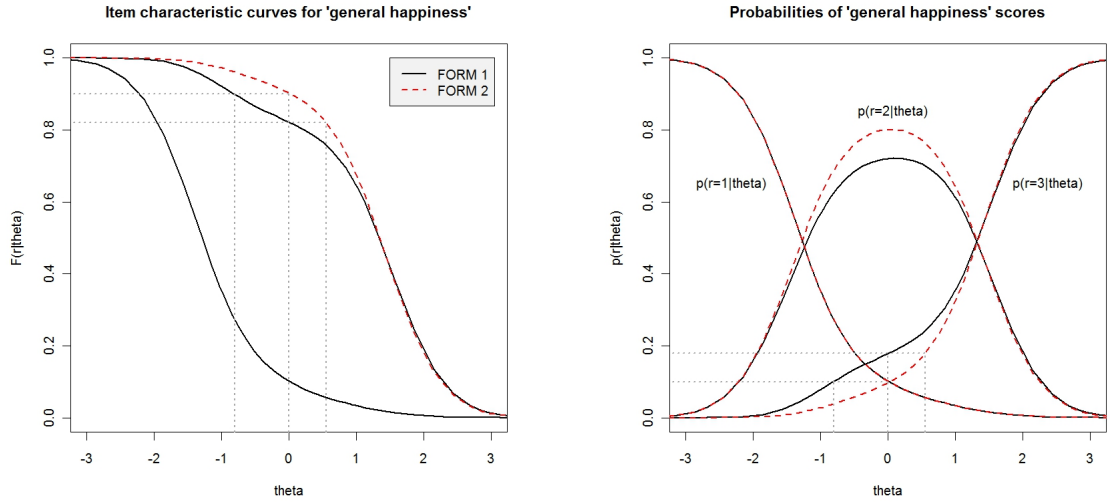
A visualization of the probabilities differences due to framing is presented in Figure 3.3. The left panel of this figure is identical to the 6th subplot of Figure 3.7 in the Appendix, where the item characteristic curves for each item are presented. The item characteristic curves measure "cumulative"probabilities, i.e. the highest curve gives the probability of answer "2 or less", whereas the lowest corresponds directly to the probability of answer "1", conditional on the latent trait θ. Therefore, the probability of answering "very happy"is given by the distance between the horizontal line $y = 1$ (not marked on the plot) and the highest curve, black or red, respectively for the $FORM = 1$ or $FORM = 2$ subsamples. This distance is smaller for the $FORM = 2$ subsamples, which is consistent with observed differences in answers frequencies. The right panel of Figure 3.3 simplifies the analysis by directly presenting the response probabilities for different θ levels. Let us analyse the response probabilities of a respondent with $\theta_0 = 0$. The probability of answering "3"("very happy") for this representative respondent almost doubles (from 0.1 to 0.18) when the five domain

[37]The coefficient next to the linear component of age effect is not significant. However, estimating the model just with the squared age brings a very similar pattern for the happiness-age relation.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

satisfaction items are presented before the "general happiness" item. In order to ensure the probability $p(r_3 = 3 \mid \theta) = 0.18$ for the $FORM = 2$ respondents, their happiness level (measured on our normalized scale) should increase from $\theta_0 = 0$ to $\theta_1 = 0.55$. Similarly, the probability $p(r_3 = 3 \mid \theta)$ amounts to 0.1 for either a $FORM = 2$ individual with the $\theta_0 = 0$ position on the latent scale, or for a $FORM = 1$ respondent with $\theta_2 = -0.8$. Similar analysis may be conducted for various latent scale locations to illustrate the effects of framing.

Figure 3.3. General happiness item and answers probabilities



The important implication of the probabilities illustration is that framing mostly influences people around the average level of happiness, for whom we observe the discrepancy between the red and black lines. Those who are either very happy or very unhappy (people at extremes) seem to be resistant to framing effects. This may speak in favor of relatively small framing power.

In summary, various testing procedures confirm the presence of framing effects at the 0.05 significance level. The differences between the frequencies or log-likelihood values are not big enough to unambiguously agree on the existence of framing effects, as would be the case if the significance level was at 0.01 or lower. Nevertheless, the existence of framing effects in the analysed GSS example is very likely, especially when the DIF analysis is restricted to the differences in answering patterns for the "very satisfied" and "pretty happy" scores. The IRT semiparametric specification shows considerable differences in certain answers probabilities when conditioning on the latent trait.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

3.6 Framing effects and their impact on the results validity

The presence of framing effects in a survey surely influences results, in particular the distribution of responses is usually examined in the first turn. Framing effects introduced into the IRT modeling framework allow the probability $p(r_i \mid \theta)$ to differ across the framing of treated and untreated groups. As in the GSS example, two sub-models are estimated, but are kept in one framework:

$$p(r_1, r_2, .., r_6 \mid W, FORM) = \begin{cases} \int p(r_1 \mid \theta)p(r_2 \mid \theta)..p(r_6 \mid \theta, FORM = 1)f(\theta \mid W)d\theta \\ \int p(r_1 \mid \theta)p(r_2 \mid \theta)..p(r_6 \mid \theta, FORM = 2)f(\theta \mid W)d\theta \end{cases} \tag{3.6}$$

The only difference across groups resulting from different values of the $FORM$ variable concerns the functional form of $p(r_6 \mid \theta, FORM)$. The elements $f(\theta \mid W)$ and $p(r_i \mid \theta)$ for $r = 1, 2..5$ stay the same in the two "submodels". There are no restrictions on the relation between the functions $p(r_6 \mid \theta, FORM = 1)$ and $p(r_6 \mid \theta, FORM = 2)$, since they are estimated separately. The stochastic dominance relation between the respective item characteristic curves for $FORM = 1$ and $FORM = 2$ may hold, implying a uniform DIF. The red and black curves are also allowed to cross, which would imply a non-uniform DIF. The alternative probit and probit-based models that could incorporate framing effects or DIF are discussed in the Appendix.

Independently of the variable's $FORM$ values, the mechanism for the IRT estimation is the same and the item characteristic curves are downward sloping. In any of the two "submodels"the ordering of the respondents, in a probabilistic sense, is preserved. That is, the individuals who had higher probabilities of a particular answer under a "no framing"scenario, are also characterized by higher probabilities when exposed to framing, but the values of those probabilities change. In this sense, framing is harmless if it concerns the whole population or is introduced as a controlled experiment. Exposing all the respondents to framing is actually analogical to making the item more difficult or easier, which in the parametric IRT is controlled for by a special parameter shaping the steepness of the item characteristic curves. Framing changes the look of the answers frequencies tables, but a more thorough analysis (as suggested here by the IRT framework) would allow for the capture of changes in the answering probabilities without drawing false conclusions about the changes in the latent traits levels and the ranking of the individuals. However, the uncontrolled

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

framing exposure may bias the results, as in the ability tests when Hispanic origin may help guess the right answer and therefore, overestimate ability.

An additional concern in this analysis relates to the sample randomization correctness. Despite methodological work, the randomization problems also occur in the GSS, as reported by Smith and Peterson (1986). A deeper insight into the structure of the two subsamples suggests that it is actually the subgroup $FORM = 2$ that should score higher on the happiness item. The $FORM = 2$ group is characterized by a slightly greater percentages of "married" and "sociable"respondents and at the same time, lower percentages of "divorced"and "black"(here the difference amounts to 12 percentage points). The $FORM = 1$ group consists of slightly more females. The remaining components of the W vector are very similar for both groups. When we compute the contribution of personal characterisitcs to each individual's happiness, i.e. we weight the vector W for each individual by the IRT coefficients, the $FORM = 2$ group is characterized by a higher value of both median and mean $W\beta$ (independently of the IRT model specification that the coefficients come from). This difference is not very large, but indicates that if the analysis was restricted only to personal characteristics, the results would be biased.

However, the question arises of how much framing changes in the analysed setting. Framing is shown to matter at the 5% significance level, but the statistical significance does not need to have an implication on the real framing importance. In order to answer this question, an exercise was conducted. Firstly, the ranking of individuals based on the mean of their happiness posterior distribution was calculated for Model 1 (no DIF) and Model 3 (DIF for one curve). The comparison of these two rankings does not show striking differences in the ordering of individuals: 1/3 of the respondents did not change their rank or changed it only by 1 position, 54% of the individuals change their ranking by no more than 3 and 90% by no more than 20. In terms of absolute shifts on the latent scale, we obtain the result that 99% of individuals achieve the estimated absolute difference below 0.085, with 90% below a difference of 0.04. The changes are contained within the range (-0.1,0.115).

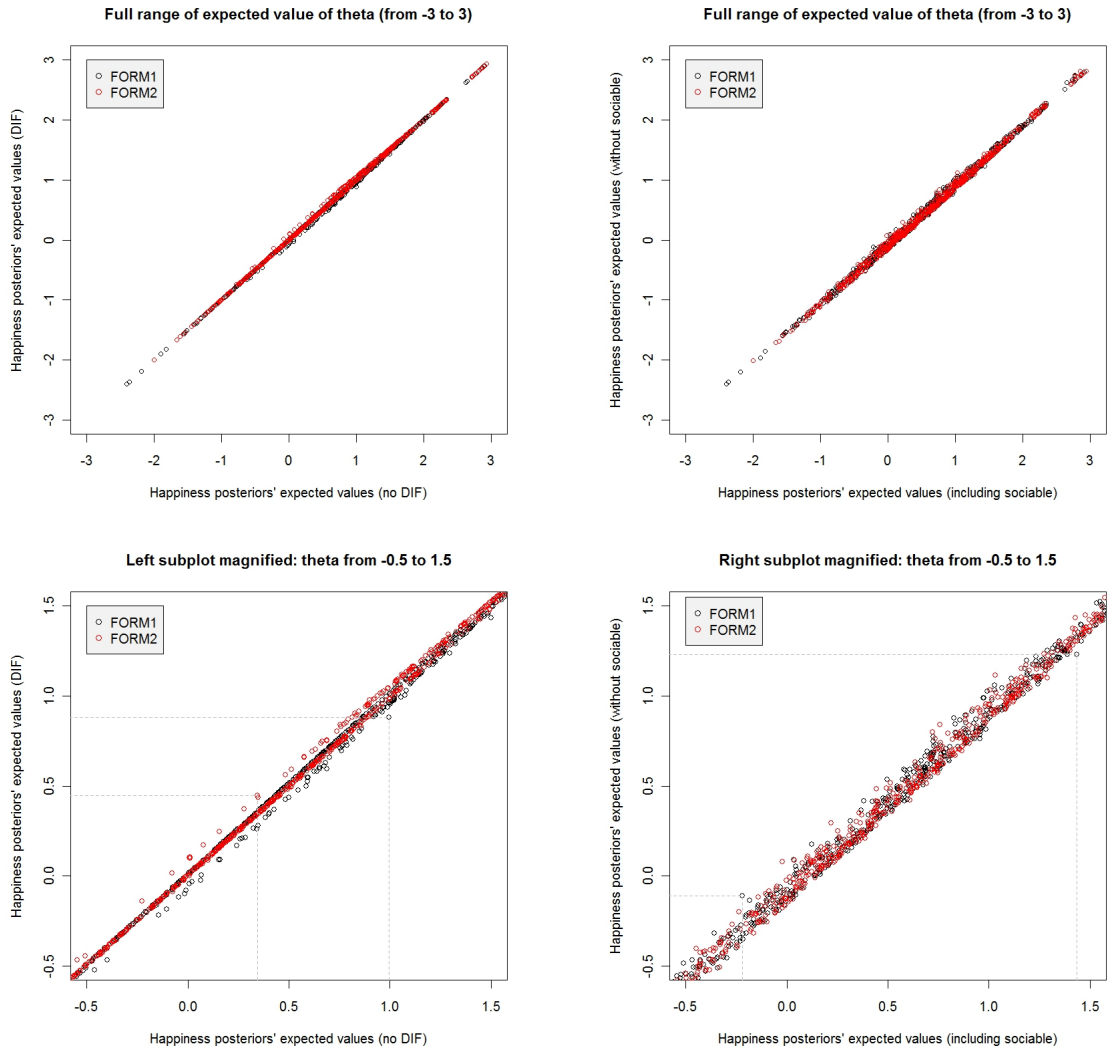
A similar analysis was conducted for the GSS data by excluding one of the significant explanatory variable in the IRT model, chosen here to be the "Sociable"variable. In this case, the differences in ranking, as well as in the values of the happiness posterior means, were bigger: 99% of the individuals achieved the estimated absolute difference below 0.172 and 90% below 0.145. In terms of ranks changes, only 8% of the respondents did not change their rank at all or change it by 1, 16,5% did not change rank by more than 3 and only 69,1% by not more than 20.

The differences between the values of the posterior means dependent on accounting for

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

framing effects and the case of differences dependent on the variable "Sociable" are illustrated in the Figure below. This plot is called a "mean-mean"plot, analogically to the "qq"plot, summarizing the changes in the estimated posteriors.

Figure 3.4. Mean of happiness posterior in Model 1 plotted against Model 3



For the left subplot we observe that there are more red points above the $x = y$ line, which corresponds to the fact that accounting for DIF allowed the $FORM = 2$ subgroup to correct happiness levels upwards. However, as noted above, it is the left panel of the Figure that shows larger discrepancies between the two estimated models (even more visible in the second row of the Figure). This finding allows us to conclude that the issue of including

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

important characteristics among the set of explanatory variables matters more than the issue of accounting for the framing effects.

In the 1986 GSS Methodological Report Smith concludes: "Alteration of the content of the GSS either by the addition or deletion of items, by the switching of items from permanent to rotating status, or by switching items from one rotation to another hampers our ability to keep measurement conditions constant and therefore increases the danger that true change will be confounded with measurement effects. This appears to have occurred on the 1985 GSS with regards to happiness. Users of the happiness items should adjust for this artifact."

Certainly, this adjustment is important for the frequency tables but, as shown above, for econometric modeling it is more important to identify all of the important factors that contribute to happiness.

3.7 Conclusions

Framing effects can probably be identified in most surveys. The broad definition comprising the TV, gender of interviewer or ordering effects makes it doubtful if we may find any measurements of respondents' opinions, believes or attitudes that are free of framing effects. The General Social Study certainly constitutes an example of a very good methodological work, controlling the distributional changes of variables, introducing experiments if needed and generally making a large big effort to ensure the comparability of waves. The rich GSS documentation, easily available on the survey website, makes researchers aware of existing problems, thanks to which the quality of the undertaken work may be improved.

The GSS example analysed here shows that framing effects appear due to a change in the ordering of questions. General happiness scores dropped when they were not preceded by the five domain satisfaction items. A possible explanation relates to possible differences in question interpretation and memory recall depending on the presence of the domain satisfaction questions.

The survey methodological teams attempt to design each survey in a way that is free of framing effects. However, many solutions on how to counteract framing effects are costly. The general line of work aims at increasing the respondent's motivation to participate in, and concentration during, the survey and decreasing, at the same time, the satisficing approach.

However, despite all of the methodological work framing effects will still exist, partly because of human nature and partly because there are always factors beyond our control. Certainly, the analysis of descriptive statistics should be conducted with caution. However, framing does not need to be so dangerous when proper econometric strategies are applied.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

The IRT framework presented here provides robust estimates of individual latent characteristics and the relation between the observables and unobservables. Finally, framing does not necessarily cause the survey data to become unreliable. When we apply econometric modeling, ignoring framing may be less harmful than misspecifying the model by omitting important variables. The dataset, even when incorporating framing effects, may still convey a lot of information about the surrounding world if we carefully choose and specify the modeling framework.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

References

[1] Adams, Raymond J., Wilson, Mark and Wu, Margaret (1997): Multilevel Item Response Models: An Approach to Errors in Variables Regression, *Journal of Educational and Behavioral Statistics*, Vol. 22 (1), pp. 47-76.

[2] Ayala, Rafael J., The Theory and Pratice of Item Response Theory, 2009 The Guilford Press, New York, US.

[3] Chong, Dennis and Druckman, James N. (2007): Framing Theory, *Annual Review of Political Science*, Vol. 10, pp. 103-126.

[4] Clark, Andrew, Etile, Fabrice, Postel-Vinay, Fabien, Senik, Claudia and Van der Straeten, Karine (2005): Heterogeneity in Reported Well-Being: Evidence from Twelve European Countries, *The Economic Journal*, Vol. 115, No. 502, pp. C118-C132.

[5] Cohen, Allan S. and Bolt, Daniel M. (2005): A Mixture Model Analysis of Differential Item Functioning, *Journal of Educational Measurement*, Vol. 42 (2), pp. 133-148.

[6] Conti, Gabriella and Pudney, Stephen (2010): Survey Design and the Analysis of Satisfaction (Note), *The Review of Economics and Statistics*, accepted for publication.

[7] Differential Item Functioning, ed. Holland, Paul W. and Wainer, Howard, Erlbaum Lawrence Associates, Publishers, 1993 Hillsdale, New Jersey, US.

[8] Entman, Robert M. (1993): Framing: Toward clarification of a fractured paradigm, *Journal of Communication*, Vol. 43 (4), pp. 51-58.

[9] Glas, Cees A. (1998): Detection of Differential Item Functioning Using Lagrange Multiplier Tests, *Statistica Sinica*, Vol. 8, pp. 647-667.

[10] Gupta, Nabanita D., Kristensen, Nicolai and Pozzoli, Dario (2010): External Validation of the Use of Vignettes in Cross-Country Health Studies, *Economic Modelling*, Vol 27 (4), pp. 854-865.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[11] Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System, Version 1; freely downloadable from Eurostat website: *http : //epp.eurostat.ec.europa.eu/portal/page/portal/quality/ess_practices/methodology*

[12] Huddy, Leonie, Billig, Joshua, Bracciodieta, John, Hoeffler, Lois, Moynihan, Patrick J. and Pugliani Patricia (1997): The Effect of the Interviewer Gender on the Survey Response, *Political Behavior*, Vol. 19 (3), pp. 197-220.

[13] Jacoby, William G. (2000): Issue Framing and Public Opinion on Government Spending, *American Journal of Political Science*, Vol. 44 (4), pp. 750-767.

[14] Kepteyn, Arie, Smith, James P. and van Soest, Arthur (2011): Are Americans Really Less Happy with Their Incomes?, RAND Working Paper, No. WR-858.

[15] Kepteyn, Arie, Smith, James P., van Soest, Arthur and Vonkova, Hana (2011): Anchoring Vignettes and Response Consistency, RAND Working Paper, No. WR-840.

[16] Kepteyn, Arie, Smith, James P. and van Soest, Arthur (2009): Comparing Life Satisfaction, RAND Working Paper, No. WR-623-1.

[17] Kepteyn, Arie, Smith, James P. and van Soest, Arthur (2007): Vignettes and Self-Reports pf Work Disability in the United States and the Netherlands, *The American Economic Review*, Vol. 97 (1), pp.461-473.

[18] King, Gary, Murray, Christopher J.L., Salomon, Joshua A. and Tandon, Ajay (2004): Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research, *American Political Science Review*, Vol. 98 (1), pp. 191-207.

[19] Kristensen, Nicolai and Johansson, Edvard (2008): New Evidence on Cross-country Differences in Job Satisfaction Using Anchoring Vignettes, *Labour Economics*, Vol. 15 (1), pp. 96-117.

[20] Krosnick, Jon A. and Alwin, Duane F. (1987): An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement, *The Public Opinion Quarterly*, Vol. 51, No. 2, pp. 201-219.

[21] Nelson, Thomas E., Oxley, Zoe M. and Clawson Rosalee A. (1997): Towards the Psychology of Framing Effects, *Political Behavior*, Vol. 19 (3), pp. 221-246.

[22] Smith, Tom W. (1986): Unhappiness in the 1985 GSS: Confounding Change and Context, *GSS Methodological Report*, No. 34.

[23] Smith, Tom W. Peterson,Bruce L. (1986): Problems in Form Randomization on the General Social Surveys, *GSS Methodological Report*, No. 36.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute DOI: 10.2870/38699

[24] Smith, Tom W. (1987): That Which We Call Welfare by Any Other Name Would Smell Sweeter: An Analysis of the Impact of Question Wording on Response Patterns", *The Public Opinion Quarterly*, Vol. 51 (1), pp. 75-83.

[25] Smith, Tom W. (1987): The Art of Asking Questions, 1936-1985, *The Public Opinion Quarterly*, Vol. 51, Part 2 (Supplement: 50th Anniversary Issue), pp. 95-108.

[26] Smith,Tom W. (1989): Thoughts on the Nature of Context Effects, *GSS Methodological Report*, No. 9.

[27] Spady, Richard H.(2006): Identification and Estimation of Latent Attitudes and Their Behavioral Implications, *Cemmap Working Paper*, CWP12/06.

[28] Spady, Richard H.(2007): Semiparametric Methods for the Measurement of Latent Attitudes and the Estimation of their Behavioural Consequences, *Cemmap Working Paper*, CWP26/07.

[29] Toepel, Vera, Vis, Corrie, Das, Marcel and van Soest, Arthur (2009): Design of Web Questionnaires: And Information-Processing Perspective for the Effect of Response Categories, *Sociological Methods & Research*, Vo. 37 (3), pp. 371-392.

[30] van Soest, Arthur and Hurd, Michael (2008): A Test for Anchoring and Yea-Saying in Experimental Consumption Data, *Journal of the American Statistical Association*, Vol. 103 (1), pp.126-136.

[31] van Soest, Arthur, Delaney, Liam, Harmon, Colm, Kepteyn, Arie and Smith, James P. (2007): Validating the Use of Vignettes for Subjective Threshold Scales, RAND Working Paper, No. WR-501.

[32] Wolfe, Rory and Firth, David (2002): Modeling Subjective Use of an Ordinal Response Scale in a Many Period Crossover Experiment, *Applied Statistics*, Vol. 51 (2), pp 245–255.

[33] Zaller, John and Feldman, Stanley (1992): A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences, *American Journal of Political Science*, Vol. 36 (3), pp. 579-616.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix A: Tables and Figures

Table 3.3. The GSS variables used in the empirical analysis

GSS Variable	Question wording
HAPPY	Taken all together, how would you say things are these days? would you say that you are very happy, pretty happy, or not too happy?
	For each area of life I am going to name, tell me the number that shows how much satisfaction you get from that area.
SATCITY	The city or place you live in.
SATHOBBY	Your non working activities, hobbies and so on.
SATFAM	Your family life
SATFRND	Your friendships
SATHEALT	Your health and physical condition
MARITAL	Are you currently married, widowed, divorced, separated, or have you never been married?
SEX	CODE RESPONDENT'S SEX
RACE	What race do you consider yourself?
COHORT	Birth cohort of respondent
INCOME86	Family income
ATTEND	How often do you attend religious services?
SOCFREND	Spend a social evening with friends who live outside the neighborhood?

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 3.4. The response frequencies of satisfaction questions

	HAPPY				
SCALE	1		2	3	
	not too happy		pretty happy	very happy	
Freq.	136		740	423	
%	10.47		56.97	32.56	
FORM 1 (%)	10.31		54.22	35.47	
FORM 2 (%)	10.62		59.64	29.74	
SCALE	1	2	3	4	5
	SATCITY				
Frequency	201	294	248	347	209
%	15.47	22.63	19.09	26.71	16.09
FORM 1 (%)	15.94	22.66	19.22	26.88	15.31
FORM 2 (%)	15.02	22.61	18.97	26.56	16.84
	SATHOBBY				
Frequency	189	173	249	447	241
%	14.55	13.32	19.17	34.41	18.55
FORM 1 (%)	15.78	12.66	18.59	34.06	18.91
FORM 2 (%)	13.35	13.96	19.73	34.75	18.21
	SATFAM				
Frequency	96	112	167	440	484
%	7.39	8.62	12.86	33.87	37.26
FORM 1 (%)	7.66	8.44	12.66	30.47	40.78
FORM 2 (%)	7.13	8.80	13.05	37.18	33.84
	SATFRND				
Frequency	74	149	225	553	298
%	5.70	11.47	17.32	42.57	22.94
FORM 1 (%)	6.88	12.03	14.84	42.19	24.06
FORM 2 (%)	4.55	10.93	19.73	42.94	21.85
	SATHEALT				
Frequency	152	199	215	431	302
%	11.70	15.32	16.55	33.18	23.25
FORM 1 (%)	11.25	15.00	16.41	32.19	25.16
FORM 2 (%)	12.14	15.63	16.69	34.14	21.40

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Table 3.5. The frequencies and descriptive statistics of socio-demographic characteristics

SEX		female	male		**RACE**		white	black	other
	Freq.	734	565			Freq.	1,112	155	32
	%	56.51	43.49			%	85.60	11.93	2.46
	For the analysis dummies MALE and BLACK were used.								
MARITAL		married	widowed	divorced		separated	never married		
	Freq.	746	138	133		55	227		
	%	57.43	10.62	10.24		4.23	17.47		
	A set of 3 dummies was created for the analysis: MARRIED, DIVORCED (divorced or separated), WIDOWED, with SINGLE as the reference category.								
SOCFREND		almost daily	several times a week	sev times a month	once a month	several times a year	once a year	never	
	Freq.	21	266	241	299	271	69	129	
	%	1.62	20.52	18.60	23.07	20.91	5.32	9.95	
	A dummy variable SOCIAL was generated taking on value 1 if respondent meets friends at least once a month.								
ATTEND		never	< once a month	once a year	sev times a year	once a month	2-3 times a month	nearly every week	every week	> once a week
	Freq.	180	92	167	153	109	135	60	287	115
	%	13.87	7.09	12.87	11.79	8.40	10.40	4.62	22.11	8.86
	A dummy RELIGION was created taking on value 1 if respondent attends religious services at least once a month.									
		min	1st Q	median	3rd Q	max				
COHORT		1897	1927	1946	1956	1967				
	Year of birth was transformed into variable AGE; AGE centered around the median (40) for the IRT analysis.									
		min	1st Q	median	3rd Q	max				
INCOME86		less then $1000	11249,5	21249,5	37499,5	more then $60000				
	Variable INCOME86 is coded as a discrete variable with 20 intervals; the values given for the 1st, 2nd and 3rd quartiles are the midpoints of the proper intervals.									

DOI: 10.2870/38699

Table 3.6. Different IRT and ordered probit specifications

	Model 1		Model 3		Model 4		Model 5		Model 6	
	coef	p-value	coef	p-value	coef	p-value	coef	p-value	coef	p-value
married	0.3817	0.0000	0.3855	0.0000	0.3953	0.0000	0.1914	0.0310	0.2004	0.0490
divorced	-0.1195	0.2146	-0.1194	0.2142	-0.1064	0.2863	-0.2132	0.0480	-0.3073	0.0130
widowed	0.1061	0.4521	0.1076	0.4463	0.1189	0.4105	0.0144	0.9150	-0.2145	0.1660
male	-0.1252	0.0394	-0.1241	0.0409	-0.1208	0.0483	-0.0858	0.1440	-0.1564	0.0210
black	-0.3833	0.0001	-0.3854	0.0000	-0.3819	0.0001	-0.3642	0.0000	-0.2234	0.0280
age	-0.0019	0.5131	-0.0019	0.5045	-0.0021	0.4701	-0.0147	0.1870	-0.0285	0.0270
age2	0.0307	0.0070	0.0307	0.0070	0.0316	0.0052	0.0184	0.1030	0.0349	0.0080
income	0.0465	0.0000	0.0466	0.0000	0.0463	0.0000	0.0473	0.0000	0.0406	0.0000
religiosity	0.3525	0.0000	0.3528	0.0000	0.3550	0.0000	0.3462	0.0000	0.2665	0.0000
sociability	0.3128	0.0000	0.3132	0.0000	0.3182	0.0000	0.2541	0.0000	0.1210	0.0840
FORM					-0.0209	0.6870	-0.0536	0.3370	-0.1363	0.0350

Notes:

Model 1: DIF not allowed (restricted model); llf=10207.42595

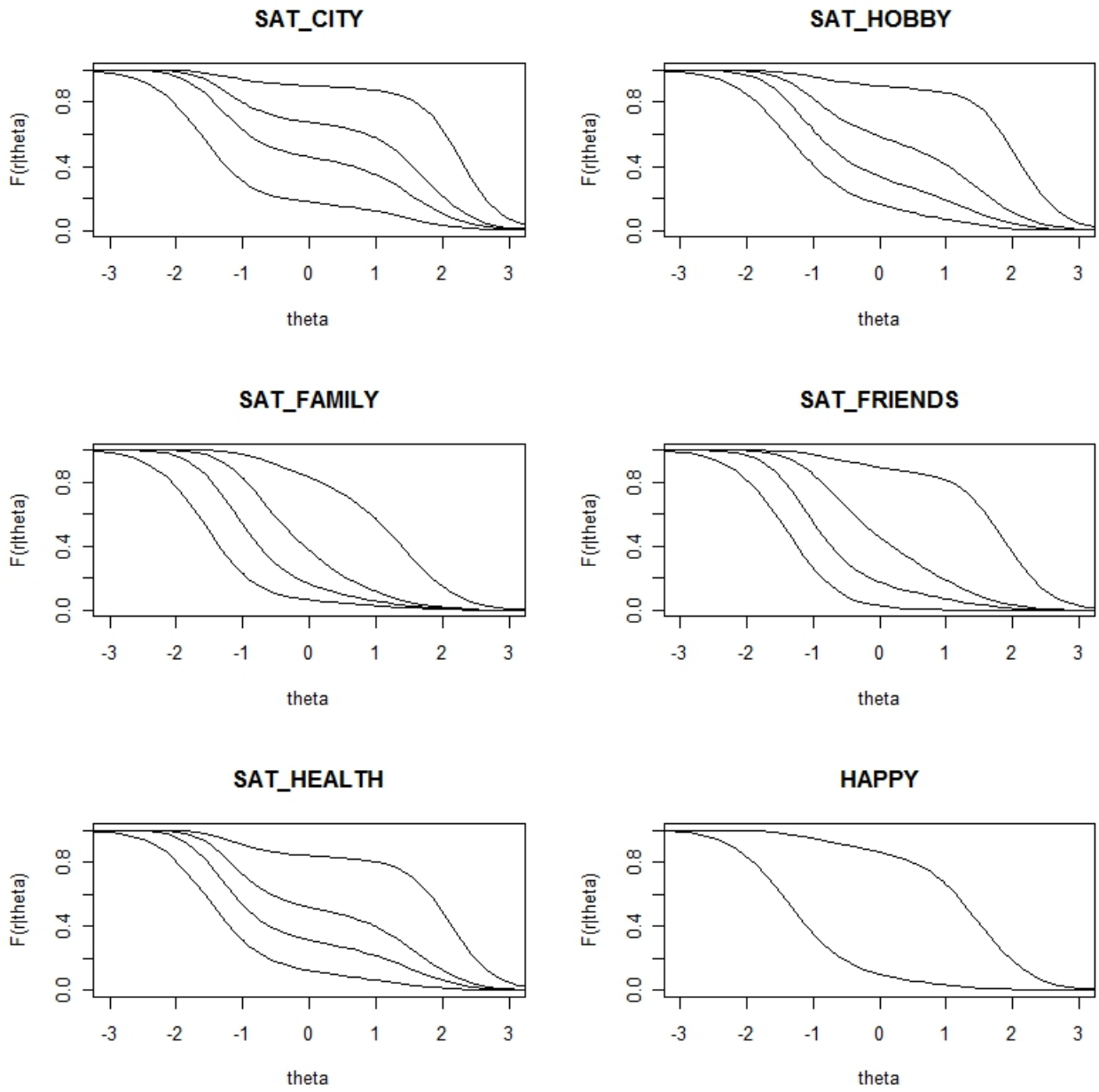
Model 3: DIF allowed for answers 3 versus 1 or 2 for HAPPY item; llf=10204.4080

Model 4: variable FORM included in vector W; llf=10207.3506

Model 5: ordered probit with sum of 6 satisfaction items as dependent variable; llf=3620.7992

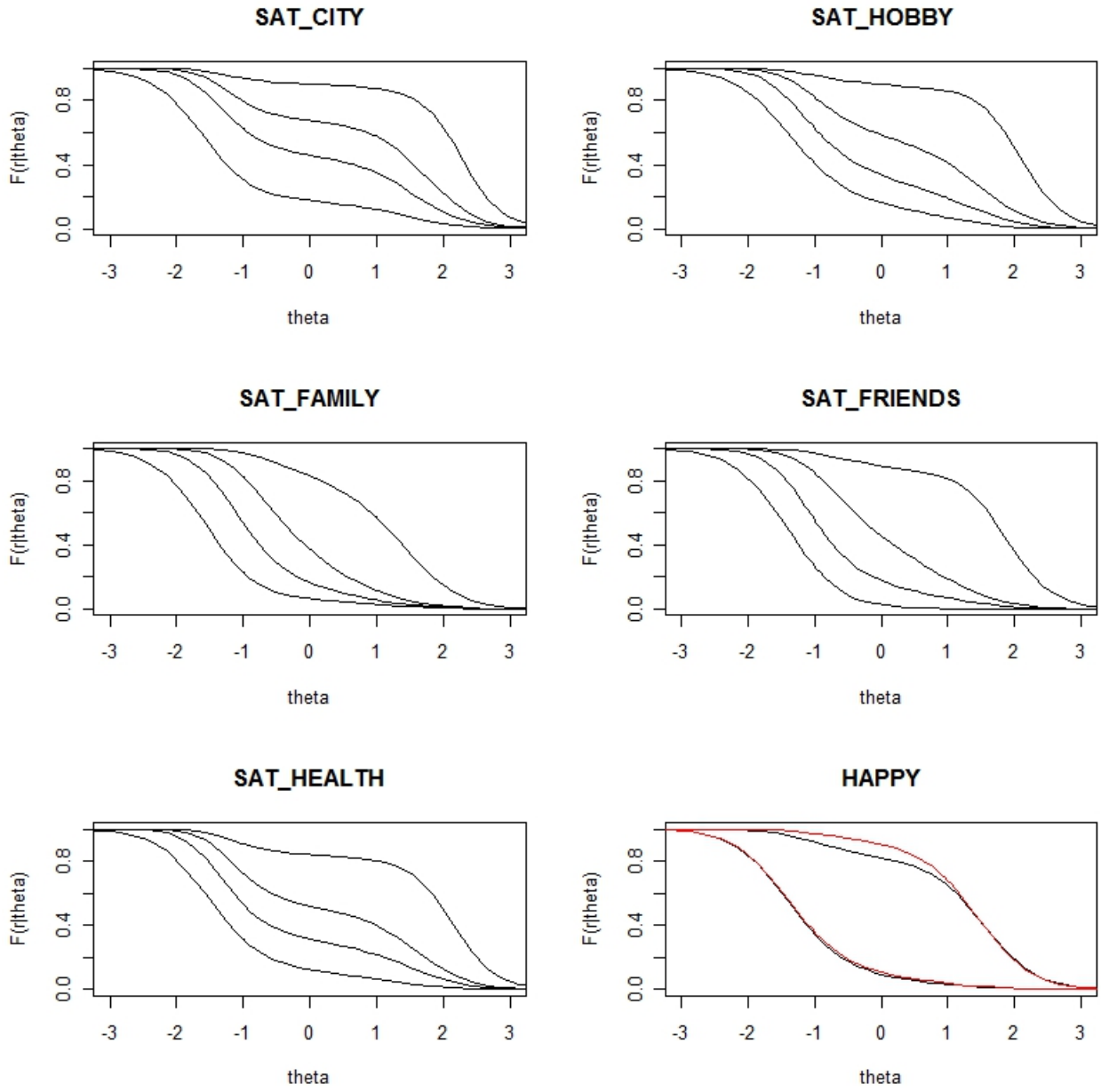
Model 6: ordered probit with happy item as dependent variable; llf=1127.8817

Figure 3.5. IRT estimation - no DIF (Model 1)



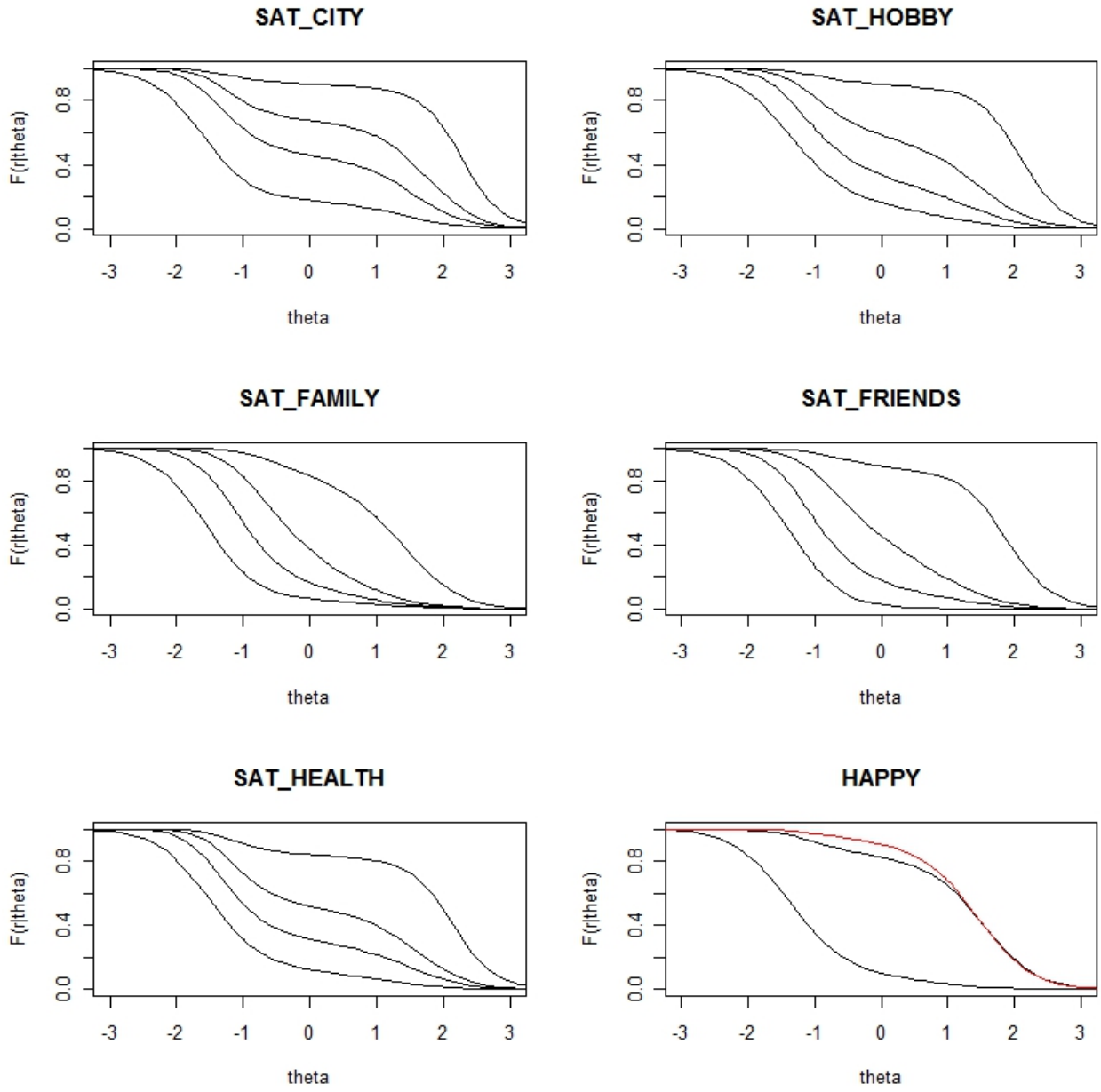
Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 3.6. IRT estimation - DIF allowed for all categories in item Happy (Model 2)



Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 3.7. IRT estimation - DIF allowed for answers in category 3 for item Happy (Model 3)



Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Appendix B: Framing effects in the probit framework

One of the most commonly applied econometric strategies for ordered responses modeling is the ordered probit model. In the GSS context considered here, the main probit drawback is its restriction of modeling basically only one item, not allowing for a parallel analysis of several dependent variables. However, let us limit the analysis to one item: general happiness in the controlled framing experiment. One method of accounting for the framing effects is to run separate ordered probit regressions. This, similar to a separate estimation of two IRT models, may not directly compare the results due to latent scale indeterminacy and may also alter other results, like the estimates on personal characteristics, which should not change due to framing. An alternative way, how framing could be incorporated into the ordered probit model, is through the inclusion of an additional framing indicator, resulting in the specification:

$$p(r_6 = j) = \Phi(\alpha_j - W\beta - FORM\gamma) - \Phi(\alpha_{j-1} - W\beta - FORM\gamma) \tag{3.7}$$

The $FORM$ variable is transformed to a dummy with value 1 for the framing case and α's are the probit threshold parameters. The sign of the obtained estimate for γ is generally interpreted as determining whether or not the latent variable (modeled only indirectly in probit) increases with the regressor. However, this interpretation can be reformulated as changes in the threshold parameters determining the level of the response, thus leaving the latent variable unaltered. The probit specification still has some drawbacks, which can be illustrated with the "general happiness"item. Assume we want to investigate the change in the probability of the highest answer "very happy", dependent on the exposure to framing. Taking into account that $\alpha_3 = \infty$, the marginal effect is thus given by:

$$\frac{\partial\, p(r_6 = 3)}{\partial\, FORM} = \gamma\, \phi(\alpha_2 - W\beta - FORM\gamma), \tag{3.8}$$

The formula implies that the effect of framing on the probability of giving the highest (and similarly the lowest) answer is always of the same sign along the whole latent trait scale. This corresponds to the concept of a uniform DIF. In the IRT framework, DIF effects could be both uniform and non-uniform, which shows the shortcoming of the probit specification. In the GSS example, asking five domain satisfaction questions before the general happiness item increases the probability of giving higher answers to the latter question. However, it could happen that such question ordering has a negative effect on individuals with lower happiness levels (reminding them how unsuccessful they are in different life domains). This

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

effect could be easily captured by the IRT framework, since the item characteristic curves are shaped only by the likelihood function maximization criterion. Probit specification would fail to account for such a relation, similarly as IRT would if $FORM$ was included among the personal characteristics W.

An extension of the probit model that incorporates DIF and the vignettes results is the so-called chopit (compound hierarchical ordinal probit) specification. The basic specification draws on King et al. (2004). We present this framework briefly below.

Denote the actual level of respondent's i latent trait by μ_i. Respondent i perceives μ_i only with random error in the self assessment question s, $(s = 1, ...S)$ and his/her unobserved perceived level is given by:

$$Y_{is}^* \sim N(\mu_i, 1). \tag{3.9}$$

The actual level varies over i as a linear function of observed covariates X_i and an independent normal random effect η_i:

$$\mu_i = X_i\beta + \eta_i,\ \eta_i \sim N(0, \omega^2) \tag{3.10}$$

The reported answer of respondent i to self-assessmnet question s with K_s ordinal response categories is modeled by the mechanism:

$$y_{is} = k,\ \text{if}\ \ \tau_{is}^{k-1} < Y_{is}^* < \tau_{is}^k \tag{3.11}$$

with the vector of thresholds such that:

$$\tau_{is}^0 = -\infty,\ \tau_{is}^{K_s} = \infty,\ \tau_{is}^{k-1} < \tau_{is}^k,\ \text{for}\ k = 1, ...K_s. \tag{3.12}$$

The thresholds are allowed to vary over the observations as a function of a vector of covariates V_i (which may overlap with X_i):

$$\tau_{is}^1 = \gamma_s^1 V_i,\ \ \tau_{is}^k = \tau_{is}^{k-1} + exp(\gamma_s^k V_i),\ \ \text{for}\ k = 2, ...K_s - 1. \tag{3.13}$$

The exponential transformation of $\gamma_s^k V_i$ is introduced in order to allow for modeling the spread of the thresholds on the latent scale. If $\gamma_s^k V_i < 0$, then we introduce the cutoff points shrinking together, meaning that the responses are spread to extreme categories. In the case of $\gamma_s^k V_i > 0$ the thresholds are spread apart, i.e. the responses are clustered in central categories. This property is directly used in Kapteyn et al. (2009), which attempts to explain the differences between the Americans and the Dutch evaluating their life satisfaction and

claim that the Dutch have a tendency to avoid extreme scores.

The second part of the model relates to the vignette evaluation. Denote the actual latent trait level of the person described in the vingette j as θ_j, $(j = 1, .., J)$. θ_j is not subscripted by respondent and is not influenced by respondent characteristics that correspond to the assumption of vignette equivalence. The actual reported evaluations, z_{lj}, do depend on respondent characteristics, but only through the thresholds. The respondents in the sample are allowed to be clustered according to the vignette asked and therefore, are indexed here differently by l. Respondent l perceives θ_j with normal random error:

$$Z^*_{lj} \sim N(\theta_j, \sigma^2) \tag{3.14}$$

Again, Z^*_{lj} is the unobserved respondent's l real-valued perception of the level of the variable being measured described in vignette j. This perception is elicited by the investigator via a survey question with the same K_1 ordinal categories as the first self-assessment question. Thus, the transformation of the latent perception to the categorical responses is analogical to the self-assessment question:

$$z_{lj} = k, \text{ if } \tau^{k-1}_{lj} < Z^*_{lj} < \tau^k_{lj} \tag{3.15}$$

with the threshold being determined by the same γ coefficients, but with the values of V measured for units l:

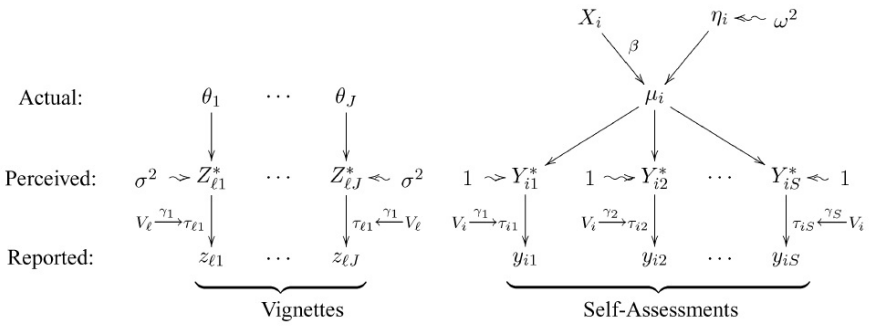
$$\tau^1_{l1} = \gamma^1_1 V_l, \;\; \tau^k_{l1} = \tau^{k-1}_{l1} + exp(\gamma^k_1 V_l), \;\; \text{for } k = 2, ...K_1 - 1. \tag{3.16}$$

Imposing that γ_1 is the same in self-assessment and vignette question corresponds to requiring the response consistency assumption to hold. The likelihood of the joint model is obtained by multiplying the self-assessment likelihood and the vignette answers likelihood, which automatically imposes the equality of the parameters. The summary of the joint self-assessment, vignettes model mechanism is presented in the Figure below.

The empirical vignette literature refers, in the vast majority of the cases, to the specification found in King et al. (2004). For instance, Kristensen and Johansson (2008) estimates the probit and the chopit models, both $V = 0$ and allows for the covariates V to influence the threshold levels. Depending on the methodology applied, the authors present different country rankings related to job satisfaction levels, claiming that there are significant differences in the respondents' latent scales.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

Figure 3.8. The summary of the vignette model (King et al. 2004)



Note: Vignette questions are on the left, with perceived and reported but not actual levels varying over observations ℓ. Self-assessment questions are on the right, with all levels varying over observations i. The first self-assessment question (see Y^*_{i1}) is tied to the vignettes by the same coefficient on the variables predicting the thresholds, γ_1, and to the remaining self-assessment questions by person i's actual value, μ_i. Each solid arrow denotes a deterministic effect; a squiggly arrow denotes the addition of normal random error, with variance indicated at the arrow's source.

Similarly, Kapteyn et al. (2009) uses the same model to consider the satisfaction differences between the Americans and the Dutch, with the extension of equation 13:

$$Z^*_{lj} = \theta_j + kI_{lj} + \epsilon_{lj}, \tag{3.17}$$

where I_{lj} is the log of income assigned to vignette j randomized across the respondents, with the values being equal to either the median income in the Netherlands or the US, or a value that is half, twice, or four times the median income in each country.

Gupta et al. (2010) applies an Ochopit model (objective-extended chopit), since apart from the self-assessment and the vignette questions, the paper uses the objective measure for health (hand-grip strength). Further, the thresholds obtained in the vignette and self-assessment submodels are allowed to differ and they categorize grip strength as an order variable, modeling it as an ordered probit:

$$Y^*_{i0} = X_i\beta_0 + \zeta_i, \text{ and } y_{i0} = m \text{ if } \tau_0^{m-1} < Y^*_{i0} < \tau_0^m \tag{3.18}$$

Here, the threshold parameters are treated as objective and are constant across the individuals. However, the assumption of one factor driving the subjective and objective measures is imposed. This corresponds to the requirement of: $\beta_0 = \beta$ (with β relating to the equation 10). The error terms η_i and ζ_i are allowed to be correlated and are modeled as being bivariate normally distributed. The likelihood function estimated in this framework incorporates the subjective, objective and vignette evaluations.

The ochopit specification is also used by Van Soest et al. (2007) to test the response

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699

consistency in the case of drinking behavior. Certainly, the vignette questions contained in a survey joined with an appropriate methodology may correct for DIF. However, the range of problems that can be addressed by the vignette methodology is limited. The vignette methodology could obviously not be applied in the happiness experiment presented in this paper. Moreover, there are many surveys that still do not incorporate vignette methodology. There are also too many questions related to subjective evaluations. Therefore, keeping the respondents' burden at a minimal level may not allow for the inclusion of many of the desired vignette questions.

Rynko, Maja (2012), On the Measurement of Welfare, Happiness and Inequality
European University Institute
DOI: 10.2870/38699