Master Thesis
Entrepreneurship, Strategy and Organisation Economics

Education Self-employment and Endogeneity

Author : Sebastien Tiggelovend
Student Number: 314567

Supervisor: Dr. J. Block

University: Erasmus University
Faculty: Erasmus School of Economics
Programme: Economics & Business

Date of completion: 08-10-2009

Table of contents

1. Introduction

The impact of education and its contribution in the modern developed world are undeniable, both from an economic and social point of view according to a large volume of economic, sociologic and other types of scientific literature. The impact of entrepreneurship has a similar undeniable impact, researched in a large amount of economic and business literature and generally being accepted as a source of jobs and a positive effect on the economy of a country.[1] This concept of entrepreneurship can be seen as a choice for a specific career or "employment status choice"[2]. The term 'employment status choice' was defined by Katz (1992) as "*the vocational decision process in terms of the individual's decision to enter an occupation as a wage-or-salaried individual or a self employed one*", later similarly defined in Kolvereid (1996). In that setting it has been researched in labour economics as well. It has been tried to identify reasons why individuals choose a certain employment status, related to their characteristics (among which educational characteristics is a part).

The relationship between these two concepts of "Education" and "Entrepreneurship"[3] has been researched to a lesser extent than the individual concepts by themselves and is a lot more complex. This relationship is an important aspect in the prediction and promotion of entrepreneurship as well as being a factor in the assessment of the merits of education. One of the problems with this relationship is the danger of endogeneity that might occur in the analysis and which has not always been accounted for in prior research. A clearer understanding of this problem and how it alters the results of a relationship between education and entrepreneurship is beneficial to a better understanding. A better understanding could lead to a better utilization of educational forces to further entrepreneurship and in that sense provide a more positive effect on the economy. This assumes that entrepreneurship is indeed a strong 'motor' of the economy in providing more jobs and improving the economic welfare of a country. This latter relationship will not be researched here however as it is outside of the scope of this thesis.

In this thesis I will attempt to unravel the relationship between education and entrepreneurship in three phases. Each phase will deepen the knowledge by addressing more difficulties in the research into such relationships. The last phase will contain an attempt at tackling the concept of endogeneity from the point of view of education and entrepreneurship. The phases are structured in a similar way as the research-sub questions and this structure is as follows:

Phase I:	The relationship between education and the preference for entrepreneurship
Phase II:	The relationship between education and the propensity for entrepreneurship
Phase III:	The relationship between education and the entrepreneurial engagement levels.
Phase IV:	The effect of endogeneity on the relationship between education and the preference for entrepreneurship as well as the propensity for entrepreneurship.
Conclusion:	Conclusion based on the relationships when endogeneity is taken into account.

[1] Carree and Thurik (2003, 2006) provide overviews of literature that further supports this now generally accepted relationship.

[2] The definition is quoted from Katz (1992) and Kolvereid (1996)

[3] The terms "Entrepreneurship", "Self-employment" and "Decision to start a new business" are used interchangeable in this thesis. This underlines that the chosen definition of entrepreneurship is the choice of self-employment in the sense of deciding to start a new business. As such the three terms are equally applicable within the context of this particular research.

One of the contributions of this thesis is that the relationship of education on the choice to start a new business is researched with the use of a continuous measure of education as opposed to a categorical one. The research is done with a logistic model for the actual decision to become self employed, as well as the preference for self employment and with an ordinal logistic model for the engagement levels of entrepreneurship.

The most important contribution of this thesis however is the research into the effect of endogeneity. Endogeneity is sometimes assumed to be irrelevant or simply not accounted for in earlier research which harbours a risk of bias. To see if this bias is indeed the case, the research in this thesis contains comparisons of models with and without endogeneity being taken into account. The results from this lead to a clearer understanding of the real effect education has on the decision whether to become self-employed, separated from the potential downwards or upwards bias due to the endogeneity of education.

Research question

With the above as well as the ambiguity of the former research on this topic in mind, it will be interesting and useful to research what the exact relationship between education and the decision to start a new business really is. This should shed at least some light on whether education does have a positive effect on entrepreneurship, the way it is generally believed in society, or whether it has a negative effect the way certain studies find. As stated before another very important aspect is the endogeneity problem that education as a variable has. Since in some studies this problem has not been corrected for, it should be interesting to see if and how results change when this is taken into account.

The research question for this thesis is therefore the following:

What is the relationship between education and the decision to start a business?

To provide answers to this research question, the relationship of education is researched with respect to:

- The preference for entrepreneurship
- The propensity for entrepreneurship
- The different levels of entrepreneurial engagement

The above relationships are researched again later in this thesis but then taking the effect of endogeneity into account.

Thesis set-up

The set-up of the thesis is as follows:

- First the previous literature on the subject of education and entrepreneurship is briefly reviewed in the light of the definitions and variables as used in the empirical part of this paper to clarify the underlying concepts. The concept of endogeneity is specifically highlighted in a subsection of the literature review.
- Secondly the methodology as used in this thesis is briefly reviewed, highlighting the univariate tests and regression models in anticipation of their use.
- Thirdly the data and variables used are reviewed to illustrate and clarify the definition of the variables that are used in the univariate tests and the various regression models.
- Fourthly the aforementioned variables are empirically examined with the usage of several univariate tests, serving as a preliminary examination in anticipation of the various upcoming regression models.
- In the fifth part the results of the three logistic regression models are presented with a brief introduction. First the results and analysis of the binary logistic regression model of phase one is presented which is the model with regards to the preference for self-employment. Secondly the results and analysis of the binary logistic regression model of phase two are presented which is the model with regards to the propensity for actual self-employment. Thirdly the results and analysis of the ordered logistic model of phase three are presented which is the model with regards to the entrepreneurial engagement levels, inspired by Zwan et al. (2008) and Grilo & Thurik (2008).
- In the sixth part the instrumental variables method (IV-method) is discussed in more detail, continuing the review of the instrumental variables method from the methodology chapter.
- In the seventh part the potential problem of endogeneity is analyzed for the first two of the previous three phases using the instrumental variables method. The previous models for preference for self-employment and actual self-employment are adapted to allow correction for endogeneity and an instrumental variable approach with a probit model is used. The results of these instrumental variable models are analyzed and discussed (in comparison with the results from the previous phases).
- In the eight and final part conclusions are drawn on the basis of the previously found results with regard to the research question. The most notable findings are highlighted and recommendations for follow-up research are made which concludes this thesis.

2. Literature Review

The importance of education by itself is generally accepted and undisputed in most modern countries. The role played by education in the development of entrepreneurs however is a tricky subject and a clear conclusion in this matter has not yet been reached despite the useful research having been done in this area. This research often lead to contradicting results and remained open to interpretation in such a way that merely based on the choice of which scientific articles to cite all three options (a negative effect of education on entrepreneurship, a positive effect of education on entrepreneurship or no significant effect at all) could be supported.

Since according to European Commission Flash Barometer 192 Report, half of the Europeans have not even thought about starting up a business, let alone taken steps, the need to examine the relationships of other factors with entrepreneurship becomes very important. This to find ways in which to increase the entrepreneurial 'spirit' and reap the assumed benefits of this spirit if factors are found that can be used in such a way. Alternatively the mere understanding of these relationships and reasons behind this low rate of entrepreneurial drive is important in its own right to further the scientific research in the field of entrepreneurship.

The relationship between education and the performance of entrepreneurs, as well as the relationship between education and the longevity of newly formed small businesses of entrepreneurs has often been researched within the literature. An example of the latter would be Bates (1990) where it is found that higher educated entrepreneurs are more likely to create firms that remain in operation than their less educated counterparts. Similarly the returns to schooling has been widely researched over the years, mostly for the returns to schooling of employees. The returns to schooling for entrepreneurs was researched only to a much lesser extent since the focus lay more on the returns to schooling for employees.

The relation of education with self-employment through other methods than the research of the return to investment in education is a topic that is only scarcely found in previous literature. The topic as such is generally linked to the return on investment of schooling or is seen as a control variable of an investigation into other factors as opposed to a focus on education as a determining factor. Especially if taking the effect of endogeneity into account, this is definitely a minority in the established base of literature. Following are the views of previous research articles regarding the role of education on self-employment in either a simple linear, binomial logistic, ordinal logistic or multinomial setting. This includes mostly views related to the return on investment of education, due to the aforementioned focus of previous literature. The amount of literature related to the relationship between education and self-employment by itself is more scarce. As such the variable education tends to be merely a control variable often or is investigated in the context of returns to education. For the above reasons most of the literature mentioned is empirical in nature as the theoretical work in the area of education or the area of self-employment is most often not aimed at the relationship between the two in the same context as this thesis, especially not when taking endogeneity into account. The articles in the return to investment context, despite not being entirely comparable, will however be referred to as they are useful with regards to at least one of the aspects of the relationship between education and self-employment.

The definition of entrepreneurship as self-employment

The concept of entrepreneurship can be defined in various ways. Traditionally there are several views and focuses developed, each with their own definition of the concept. For example in following Knight (1921) the concept of risk is more emphasized when viewing entrepreneurship and it is seen as the bearing of uncertainty. Schumpeter in his work[4] sees the entrepreneur more as a mover in economic development and finds the innovate character of entrepreneurship such as 'creative destruction' to be a main facet to focus on. Historically, as noted by Acs (2006) there are at least two meanings. First the occupational notion which is the one used in this thesis, where entrepreneurship is the creation of a new business that is owned by the creator. In a simple term: self-employment. As Parker (2004) states, this is a quite thoroughly researched aspect of entrepreneurship, despite its limitations by being a rather narrowly defined scope. This definition as self-employment follows the tradition of labour economists. Secondly entrepreneurship in the sense of a behavioural notion, where an economic opportunity is seized. In this case it is not required for the entrepreneur to also be the owner of a business, merely to be the individual that observes and acts upon perceived opportunities which can possible be seen as a form of arbitrage.

The effect of self-employment on economic growth

In previous literature the majority of articles implicitly or explicitly link entrepreneurship to economic growth of a country or individuals, depending on the measure used. Acs (2006) answers the topic with a simple logic: *"Entrepreneurs create new businesses, and new businesses in turn create jobs, intensify competition, and may even increase productivity through technological change. High measured levels of entrepreneurship will thus translate directly into high levels of economic growth."*. Of course, as is also stated in Acs (2006), the reality is a bit more complex but the main logic behind this is followed nonetheless in most articles. Reality is that high level of self-employment may also indicate that there are too few conventional employment opportunities One could distinguish between necessity and opportunity entrepreneurship where the former would be more indicative of a lagging economic growth and the latter a more positive indication of economic growth. Wennekers and Thurik (1999) also find that entrepreneurship matters in the context of economic growth. They argue that a substantial reallocation of resources is required and as such a higher demand for entrepreneurship arises.

Education in the context of human capital investment

As noted by Kolstad and Wiig (2009) in the classic Mincer (1974) human capital model, education has a productive impact. Education in the sense of schooling is a classic investment decision where there is a maximization of the returns to investment comparing future net benefits and current costs of investing in education. Since education is obtained, generally this means more education, this improves the performance, otherwise the investment decision would be flawed. This is the basis of the assumption that the choice for education is dependent on the return to investment and as such that individuals base their educational decisions on this expected return.

[4] For example in Schumpeter (1934) but other work shows a similar approach.

Becker and Chiswick (1966) investigate education in the sense of years of schooling and find evidence that a significant part of the inequality in earnings is explained by schooling, indicating that education has a worthwhile impact and confirming the human capital theory.
Becker (1975) investigates the effect of formal education on earnings (as well as productivity) and find a rate of return of 10 to 12 per cent per year, which underlines that the effect is significantly present. Mincer (1970) has similar findings which indicate that human capital increases the expected earnings to some degree and notes this might be aided by underlying aspects such as ability, which would hint at the endogenous nature of education in the sense of omitted variables[5].
In Becker (1962) it is already noted that abler persons would receive more education than others and that they would invest more into education than others. In this sense ability and investment would be positively correlated. This already shows a possibility for endogeneity in the sense that this ability could be an omitted variable in equations used in further research. However although a distinction between the rate of return for employment or self-employment is not readily made, it would be sensible to assume a difference which could explain a part of the choice for self-employment.

The effect of education on self-employment

The relationship between self-employment, preference for self-employment or a similar dependent variable and education has been researched to various degrees in the previous literature. The results are however not consistent over all, this is also noted in Grilo and Thurik (2006). The level of education is a variable for which contrasting results have been obtained both for the existence of a significant impact as well as the nature of this impact on preference for self-employment and actual self-employment. Grilo and Thurik (2006) state the same findings as is shown in the literature review of this thesis, which is that:

- Some studies do not find a significant impact of education.
- Among the studies that do find a significant impact, this relation is sometimes positive, sometimes negative and sometimes negative up to a certain level and positive there-after, depending on the study set-up.

This could be due to the lack of accounting for endogeneity in most studies. Grilo and Thurik (2008) note that education suffers from the risk of endogeneity, indicating that any future comparison result would need to take this into account. In that light an iv-analysis is performed in this thesis to attempt to unravel the endogeneity effect from education for a less biased result. Despite the potential endogenous bias however the results of previous literature convey that the relationship warrants further research. A further overview of the studies as meant above follows:

In Zwan et al. (2008) it is shown that education has a significant effect in an ordered regression and shows a positive sign for education and a negative sign for educations squared. The effect in the relevant range however is found to be positive (and there is a turning point of 47 years as 'age when finished full time education' after which the direction of the effect changes). Robinson and Sexton (1994) find that the number of years of formal education increases the probability of becoming self-employed (by 0.8%) and as such would suggest a positive relationship. Robinson and Sexton (1994; p. 154) state that *"higher levels of"* general

[5] The omitted variable in studies using standard data-sets would be ability which is tied into education but often not measured or only measured as the ability of the person at a later age, after the education decision is taken.

"education increase both the probability of becoming self-employed and the success of individuals in that sector in terms of the earnings.". Blanchflower (2004) finds that *"In Europe the probabilities are lower the more educated an individual is, while the opposite is true in the US."* This indicates the importance of using variables to control for geographical influences such as the country the individual is in to reduce potential bias.

In Grilo and Irigoyen (2006) the way education is measured is different from the set-up chosen in this thesis. The same question 'Age when finished full education' is used to construct three education levels as opposed to using it as one linear variable. (These education levels are separated into low, intermediate and high indicating before the age of 15, between 15 and 21 and above 21 respectively where the intermediary level is used as the base.) According to their estimates the level of education does not have a significant relation with preferences for self-employment. This result is comparable with Blanchflower et al. (2001) where years of schooling was used as measure of education. Grilo and Irigoyen (2006) also find that there is a positive effect of education on being self-employment. This effect as measured by the aforementioned levels have a u-shape where lower education and higher education have a positive effect compared to intermediate education. Blanchflower et al. (2001) uses years of education in linear form in their regression and finds a negative impact of education on the probability of being self-employed, contradicting Grilo and Irigoyen (2006).

The results from Evans and Leighton (1989, 1990), who use years of education in linear form in the regression, find a positive effect of education on the probability of being self-employed, stating: *"the probability of being self-employed is higher for more highly educated individuals even after we control for individuals in professional occupations."* This confirms Grilo and Irigoyen (2006) but contradicts Blanchflower et al. (2001). Neither of these three studies mentioned above however added a quadratic term to check the existence of a u-shaped relationship so the contradiction might be a result from a possible u-shaped relationship or from endogeneity for which no control was used.

Grilo and Thurik (2008) perform a multinomial logit regression with 7 entrepreneurial engagement levels, including the independent variable of education in the form of 3 educational levels and find that education has a significant effect. Relative to the lowest engagement level used in their research all other categories, with the exception of 'no longer being in business', have a positive relation with education. They state that this is not in contrast with earlier literature.

According to Le (1999) the level of education can influence the propensity to become self-employed in several ways. In his research he refers to Lucas' (1978) model, indicating that one way education can influence the probability of self-employment is through the enhancement of an individual's managerial ability. An opposite way in which education may influence the propensity for self-employment, according to Le (1999), is that a higher level of education may ease entry into the wage-sector and in that sense decrease the probability for self-employment. This through the creation of outside options where it would be more profitable to be under wage employment as opposed to self-employment. The net impact of both these effects would then be the resulting effect of education on the propensity for self-employment, offering another possible explanation for the conflicting evidence found in the literature regarding this subject. Furthermore Le (1999) states that *"One of the major theoretical determinants of self-employment choice is educational attainment"*. In his article he also tested risk attitude, access to capital and other potential determinants. The two

mentioned above have are generally regarded as important enough in the context of this research to be added at least as control variables.

Van der Sluis et al. (2008) show that there is a lack of uniformity in the measurement of education. They find that despite years of education often being used to build the variable, the most used proxy is a system of dummies for various educational levels. They find that this lack of agreement on the definition warrants carefulness and that it complicates comparisons. Apart from this they find that: *"Apart from the lack of agreement on the definitions of the key variables, a comparison of the compiled studies also indicates a lack of common tools and techniques. Studies differ substantially in the selection of control variables that enter the relationship between schooling and entrepreneurship outcomes."*

Thurik et al. (2007) use a set of perceptual variables alongside the more 'standard' demographic variables such as gender, age, whether parents are self-employed and of course education. The usage of these perceptual variables are a common practice in this context according to Thurik et al. (2007). In their research they find that education is not significant in the preference for self-employment in the sense of latent entrepreneurship and find a hardly significant result for a negative relation between education and actual self-employment. The reasoning behind this, as stated in the article, would be that the lower the education, the fewer the job opportunities and as such these individuals would be entrepreneurs out of necessity.

Van der Sluis et al. (2004) find no empirical evidence that a systematic relationship between the education of an individual and the probability of selection into entrepreneurship exists. They do note that this does not necessarily contradict economic theory because as stated before: there are two opposing effects which causes ambiguity as to which effect prevails in an empirical setting. They also find that it is not clear from a meta-analysis that the returns to education would be uniformly higher for employees than they are for entrepreneurs. Instead they find that in Europe, the returns to education are slightly lower for entrepreneurs than for employees and the opposite is true for the U.S.

Despite the contradictory results in the literature from which a brief overview of theoretical works has been given, the importance of education by itself is generally accepted. The role played by education in the development of entrepreneurs however is shown to be a tricky subject and a clear conclusion in this matter has not yet been reached despite the useful research done in this area. The research often lead to contradicting results and remains open to interpretation in such a way that merely based on the choice of which scientific articles to cite and which method is chosen all three options (a negative effect of education on entrepreneurship, a positive effect of education on entrepreneurship or no significant effect at all) could be supported. This emphasizes that the methodology and reasoning chosen is important, due at least in part to endogeneity. This undecided 'outcome' of what the actual relationship entails is also a reason why further research into this area (and related areas) is quite necessary.

2.1 Endogeneity

As Verbeek (2004; p. 132) states: *"it is often argued that many explanatory variables are potentially endogenous, including education level…"*. This is mentioned in the context of micro-economic wage-equations but can be noted to be equally true if moved to the context of an 'employment status choice'[6] considering the relationship between expected wages and the occupational choice mentioned earlier as an assumption based on previous research.
Grilo and Thurik (2008) as well as Parker and van Praag (2006) state that the world of entrepreneurial choice is known for its endogeneity problems. A similar thing could be said for the education research. This occurs for example when an omitted factor influences both an independent variable such as education, and the dependent variable, such as preference for self-employment or actual self-employment. The methodology behind correcting for endogeneity is commonly chosen to make use of the instrumental variable (IV) method. In this method variable(s) are to be found that are correlated with the independent variable that is endogenous, and uncorrelated with the error or disturbance term of the original equation. Generally speaking this is a difficult task due to an often low correlation with the independent variable and seems often also determined by the limitations of the data-set, although items such as family background variables have been used previously.

There are several reasons why education is seen as endogenous and would require IV analysis as opposed to an OLS approach. It is possible that the education as self-reported by the individuals may be misreported, as also stated in Murray (2006) when they refer to a study by Ashenfelter and Rouse (1998). This by itself would cause a bias in the results of a logistic regression without utilizing the instrumental variable method.

Another way is that omitted variable(s) could create a bias, in the sense that the ability of a child could impact both the education the child receives and later in life it could impact the career choice, which could then be mistakenly attributed to education. Since (an) omitted variable(s) can not always be reliably added as a variable to the regression and are not available in this data-set, the IV method can be a solution. Through this method the separate effect from the omitted variables which would be in the error of the equation are set apart from the effect from education itself. The omitted variable as mentioned could be such that: the choice of educational level is potentially the result of a person having a greater ability in which case this ability is the reason why a certain career choice is made as opposed to education. Because of this an effect previously attributed to education could actually be because of this unmeasured ability. This unmeasured ability would be present within the error term of the original logistic regression and as such cause an endogenous effect. For this reason the effect of education not related to the error term would need to be separated as mentioned.

If the education is considered as a dynamic model of schooling decisions in a sequential setting, one would have to consider the value of different opportunities and choices. As Heckman et al (2005) find there is an option value to education as well. The economic return to education includes the potential for achieving a higher amount of education. This could conclude in sizable option values. Through IV-estimates this might be lessened but the empirical effect would require a more in-depth analysis into the return to investment in education to be determined.

[6] The definition is quoted from Katz (1992) and Kolvereid (1996)

Another potential problem related to endogeneity as stated by Bascle (2008) is simultaneous causality. This occurs when the causality runs in both directions from the independent variable to the dependent variable and vice versa. This might arguably be the case with education and self-employment. Education might be chosen with the future prospect of self-employment or vice-versa this future prospect might be the result of education.

In conclusion it should be noted that the sources of endogeneity can cumulate according to Bascle (2008) and as such it can be assumed that education is endogenous, even if one of the various types above may not be a significant problem by itself, the accumulation might make it more apparent and as such the IV approach could help remedy the potential bias arising from the endogeneity issues as described above.

3. Methodology

There are several methods used in this thesis which can roughly be divided into four types. These methods were chosen due to them being the most appropriate for the type of analysis necessary to examine the relationship between education and self-employment within the boundaries of the available data-set. Considering that they are by themselves not extraordinarily adapted from the standard usage in economic analysis they will be only reviewed shortly in the context of how they were used in this thesis. It should be noted that the 'Instrumental Variable' approach is described in more detail in chapter 7 before the results from this analysis are shown.

Regarding the choice of variables in the regression models used, I have chosen for one of the most-used sets of control variables (whilst allowing for a possible u-shaped relation for age as well as education) and for a method that has been previously used incorporating in the research a binary model as well as a model based on engagement levels. The binary models (though slightly adapted) are later used for an instrumental variable analysis for the purpose of separating a possible endogeneity effect. These variables are described further in chapter 4 of this thesis. One of the reasons for these choices is to attempt a more uniform approach. This because as Van der Sluis et al. (2008) already noted, the lack of uniformity in the measurements and approaches to education in research is troublesome and warrants caution, just as the lack of uniformity in the usage and selection of methods and control variables. I also utilized a set of perceptual variables, alongside the 'standard' demographic variables, as well as a measure of risk tolerance, following Thurik et al. (2007) in that respect since as they state these are more often used than not and are found to have merit as control variables.

The four types of methods used in this thesis mentioned above are as follows:

- Univariate analysis
- Binary logistic and probit regression
- Ordinal logistic regression
- Instrumental variables method

Univariate analysis

In chapter 5 "Univariate Analysis" a number of preliminary tests are done to get an indication of the relationship between education and the preference for self-employment and actual self-employment. These tests are univariate in nature in the sense that they only take one variable into account which is education.

The Kolmogorov-Smirnov test (KS-test) is used to get an indication of whether there is a significant difference in education between people who are entrepreneurs and people who are self-employed as well as between people preferring self-employment or not preferring self-employment in the dataset. The choice for this test, which is a form of minimum distance estimation, is made due to its nature of not making an assumption about the distribution of data to avoid problems with potential non-normal distributed data.

The Wilcoxon rank-sum (Mann-Whitney) test is used to find if there is a significant difference in education between self-employed individuals and not self-employed individuals and individuals preferring self-employment and individuals not preferring self-employment.

This approach tests the difference in medians between the two groups and as such can form an addition to the results of the previous test.

The variance ratio test is used to test whether the variances of the two groups (preference for self-employment and no preference for self-employment and actual self-employment and non-self-employment respectively) differ significantly with respect to education. This test is used to determine whether to use the two sample t-test with equal or unequal variances.

The two-sample t test with (un)equal variances is used as a parametric counterpart to the previous Wilcoxon rank-sum (Mann-Whitney) since the latter is similar in a way with the difference that ranking has taken place over the combined samples. The t-test in this case is used while drawing on the Central Limit Theorem in the sense that the amount of observations should be large enough to avoid breaking the assumption of normality which could interfere with the results.

Binary logistic and probit regression

For the regression models in phase two and three, the binary regression methods of logistic regression and probit regression are used. Logistic regression is primarily used for the analysis before endogeneity is taken into account. In that part it is used to model the non-linear relation between education and self-employment which can not be modelled with a standard linear OLS regression. The probit method is used only in chapter 7 to provide an easier comparison with the instrumental variable model which is an IV-probit model and for this reason the coefficients are better comparable with a probit model than the logistic model. Both these models are also convenient as they are binary in nature due to the nature of the dependent variables.

The form of these models is as follows: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ etc. where β_0 is the intercept where all X-factors are 0 and β_1 and so on are the coefficients measuring the size of the impact of each factor. The dependent variable is calculated as the logit variable[7] and then features as Y, mentioned in the equation above.

A similar reasoning is used for the probit regression with the difference that it uses a probit link function. This function is the inverse of the standard normal cumulative distribution. It also utilizes standard maximum likelihood procedure, similar to the logistic regression. The two models are both alternatively used in previous literature and economics research and as such both are appropriate and their results are usually not very different.

Ordinal logistic regression

Since the third phase has a dependent variable that is not binary but has multiple classes, the choice for the non-linear regression model in this case is limited to either multinomial or ordinal logistic regression. These are in essence a sort of extension of the logistic regression model except that it allows for more categories, be it ordered or not ordered, utilizing an assumption of proportional odds. Since the classes as defined of the entrepreneurial engagement levels can be ranked in a way where they follow a logic order, ordinal logistic

[7] The logit variable here is the natural log of the odds of the dependent occurring or not occurring.

regression is preferred over multinomial logistic regression due to the potential loss of information in the latter since any information pertaining the order would be lost. For this reason the ordinal logistic regression is used for the third phase.

Instrumental variables method

The instrumental variable approach (IV) is the method used to attempt to compensate and correct for the bias that is expected by the endogeneity of the education variable. Since chapter 7 "Instrumental Variables Method" is dedicated to this method, the review here will be more brief in nature.

The IV technique itself is a method commonly used to correct against endogeneity where other regression techniques such as logistic regression or OLS regression fall short and produce biased results. It operates on the basis of excluding instrument variables which are used in a separate regression to approximate the endogenous variable. More accurately: instrument variables which are uncorrelated with the error term of the original equation are used to approximate the endogenous variable as this variable is in some way correlated to the error term in the original equation.

Instrumental variable approach can deal with the potential problems of an endogenous variable such as:

- Misreporting in a self-reported variable such as education.
- Omitted variables either due to the data-set or due to a variable not being measurable.
- Self-selection which is a potential problem with education.
- A potential option value for the endogenous variable, such as in the case of education where more education could have a value due to it 'creating' an option.
- Simultaneous causality when the causality runs in both directions between the endogenous variable and the dependent variable.

IV can correct for omitted variables which is a problem with endogenous variables such as education. Similarly it can help in the other situations as well and as such the method is valuable since the instruments ensure that the problems do not produce inaccurate results.
As Leamer (1983) has shown the mere choice of which variable to include in an analysis can skew the results to a large amount and in instrumental variables this is also a large potential problem. The usage of the IV-method is therefore not without flaws by itself as its scope is more narrow and the limitations of data-sets more stringent due to the use of the technique. Similarly there are a number of problems that need be watched out for:

- The instrument used needs to be relevant, tested by the correlation between the endogenous variable and the instrument.
- The instrument needs to be valid and thus exogenous itself, tested by an over identification test if enough instrument variables are available.

For the above problems the Wald-test for exogeneity was conducted after each IV-model to test if endogeneity of education is truly a factor in the model. And similarly the Amemiya-Lee-Newey minimum chi-square test is used as an over identification test to test the validity of the instruments. This apart from the correlations as found in the coefficients of the first equation in the IV-models to indicate the relevance of the instruments.

4. Data

The data used for the analyses in this thesis is from the Flash Eurobarometer survey on Entrepreneurship (no 192)[8]. This survey was conducted by telephone in January 2007 on a random representative sample. The characteristics of this survey are shown in short in the table below. It should be noted that the countries includes the United States, next to 27 European Member States.

Sample	People age 15+
Countries	28
Respondents	20,674

The survey provides information about socio-demographics variables, perception and preference variables and several other indicative measures. The socio-demographic variables are variables such as age, gender, education level, employment-type of parents. Perception and preference variables contain perceptions of obstacles to entrepreneurship such as availability of financial support, accessibility of information and a crude measure of risk tolerance. For the measure of entrepreneurship, three different dependent variables were possible from the data and all three were used: The preference for entrepreneurship in binary form, actual entrepreneurship in binary form and an ordinal indicator separated by engagement level of self-employment. Each of these dependent variables is analysed in different phase:

- Phase 1 for the preference for self-employment
- Phase 2 for the actual self-employment
- Phase 3 for the engagement level of self-employment.

The independent variables used are the same for all three phases. They are divided in three categories: socio-demographic variables, perception or preference variables and country dummies. Below they are shown with the questions from the questionnaire that were used to create the variables as well as the method of construction of the variable, where clarification was necessary. The chosen variables are typically used in previous research such as Zwan et al. (2008), Blanchflower (2004), Davidson (2006), Parker (2004) among others.

Independent variables

Socio-demographic variables

- Age: The age of the individual in years.
- Age/100 squared: This is a quadratic variable created by dividing the age by 100 and then squaring the result. Although not used in all previous literature as mentioned above, a fair share does incorporate this variable and has found a quadratic effect of age[9], making the inclusion of this variable worthwhile.
- Education: Education was measured as "age when finished full-time education" to have a continuous variable as opposed to having three dummy variables. This was a choice in order to facilitate the possibility that this type of variable would contain more information

[8] This survey was conducted on behalf of the European Commission's Enterprise Directorate-General.
[9] For example Grilo and Thurik (2006) or Zwan et al. (2008)

than a separation of three categories. The "Age when finished full-time education" was discounted with 6 years[10] to make sure the variable more closely captures actual education.

- Education/100 squared: This is a quadratic variable created by dividing the aforementioned education variable by 100 and then squaring the result. Although not used in all previous literature as mentioned above, it is an analogue to the age/100 squared variable aimed at determining a potential quadratic effect of education, inspired by its use in Zwan et al. (2008).
- Gender: A simple binary variable, either male or female with male = 1 and female = 0.
- Location: This variable is based on the question whether the individual lives in a metropolitan zone, a town/urban centre or a rural zone, where the coding is such that 1 = metropolitan area, 2 = town/urban centre and 3 = rural area.
- Self-employment father: The occupation of the father coded as 1 = self-employed and 0 = employee
- Self-employment mother: The occupation of the mother coded as 1 = self-employed and 0 = employee

Perception and preference variables

- Financial support: This variable measures the perception of a lack of financial support with the question "It is difficult to start one's own business due to a lack of available financial support" where agree is coded as 1 and disagree as 0.
- Administrative complexity: This variable measures the perception of a lack of financial support with the question "It is difficult to start one's own business due to the complex administrative procedures" where agree is coded as 1 and disagree as 0.
- Information lack: This variable measures the perception of a lack of available information with the question "It is difficult to obtain sufficient information on how to start a business" where agree is coded as 1 and disagree as 0.
- Risk tolerance: This variable measures the perception of a lack of financial support with the question "One should not start a business if there is a risk it might fail" where agree is coded as 0 and disagree as 1 so that the 1 indicates risk tolerance. This is a very rough measure of risk tolerance of course but it is the best possible variable in the data-set to account for risk attitudes and in that sense it can still be a useful control variable.

Country dummies

For the countries dummies were created. The following countries are present in the data-set: Belgium, Czech Republic, Denmark, Germany, Estonia, Greece, Spain, France, Ireland, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Norway, Iceland and the United States. The country 'United States' was chosen as base for the models. The actual coefficients of the country variables are not displayed in the model tables for the purpose of legibility because they aren't a focus of this research.

[10] The age of 6 years was calculated using the publication "Compulsory age of starting school in European countries, 2009" of Eurydice at NFER. Using the known countries present in the data-set and the compulsory age according to the aforementioned publication, as well as averaging the compulsory ages of the states of the U.S. for an average for the U.S., the average age was determined to be 6.11, rounded to 6 years.

Instrument Variables

- Social class father: This is a rough measure of the social class of the father by categorizing his occupation. It is separated into 4 dummy variables. The coding is 1 = individual is part of the category, 0 = individual is not part of the category. These dummy variables are:
    - Unemployed Father
    - Blue collar Father
    - White collar Father
    - Self-employed Father

  From the above variables, the dummy variable "White collar father" is used as the instrumental variable that is excluded from the education equation.
- Social class mother: This is a rough measure of the social class of the mother by categorizing her occupation. It is separated into 4 dummy variables. The coding is 1 = individual is part of the category, 0 = individual is not part of the category. These dummy variables are:
    - Unemployed Mother
    - Blue collar Mother
    - White collar Mother
    - Self-employed Mother

  From the above variables, the dummy variable "White collar mother" is used as instrumental variable that is excluded from the education equation.

Dependent Variables

Phase 1:
For the dependent variable the simple question *"Suppose you could choose between different kinds of jobs which one would you prefer?"* was used. This is arguably a very hypothetical way to measure the preference but nonetheless should provide information nonetheless. The following 2 variables are constructed.

- Preference for self-employment (Strict definition): This variable measures the preference for self-employment in a binary form. It is coded as 1 = preference for self-employment, 0 = preference for employee
- Preference for self-employment (Wide definition): This variable measures the preference for self-employment in a binary form. It is coded as 1 = preference for self-employment, 0 = preference for not self-employed (employee or no strict preference).

Phase 2:
For the dependent variable the simple question *"As far as your current occupation is concerned, would you say you are self-employed, in paid employment or would you say that you are without a professional activity?"* was used. The following 2 variables are constructed.

- Actual self-employment (Strict definition): This variable measures the actual occupation of the individual in a binary form, coded as 1 = self-employed, 0 = employee
- Actual self-employment (Wide definition): This variable measures the actual occupation of the individual in a binary form, coded as 1 = self-employed, 0 = not self-employed

Phase 3:
For the dependent variable the question "Have you started a business recently or are you taking steps to start one?" was used, combined with the follow up question: "How would you describe your situation".

- "No", "It never came to your mind to start a business"
- "No", "You are thinking about starting up a business
- "No", "You thought of it or had already taken steps to start a business but gave up"
- "Yes", "You are currently taking steps to start a new business"
- "Yes", "You started or took over a business in the last 3 years which is still active today"
- "Yes", "You started or took over a business more than 3 years ago and it's still active"
- "No", "You once started a business, but currently you are no longer an entrepreneur"

The answers are ordered to reflect a different increasing engagement level in entrepreneurship. This set-up for levels of entrepreneurial engagement is slightly similar to the setup in Grilo & Thurik (2008). It should be noted however that the amount of levels was reduced to five to achieve a more ordered version of engagement levels, dropping the level of "Thought about it or have taken steps but gave up." and "Once started a business but no longer an entrepreneur" as they are arguably of a different 'drop out' or 'retirement' nature and as such do not fit within an ordered setting.

This setup in five levels is coded as follows:
- 1 = "Never thought about it"
- 2 = "Thinking about it"
- 3 = "Taking Steps"
- 4 ="Young business"
- 5 = "Old business"

5. Univariate Analysis

The Kolmogorov-Smirnov test (KS-test) is used to get an indication of whether there is a significant difference in education between people who are entrepreneurs and people who are self-employed in the dataset. This test is used at first because it does not make an assumption about the distribution of data and there is reason to believe the data is not normally distributed. The Wilcoxon rank-sum (Mann-Whitney) test is used as well to find if there is a significant difference in medians between self-employed individuals and not self-employed individuals and individuals preferring self-employment and individuals not preferring self-employment. The two-sample t-test is used as well further along in these tests to obtain a more complete indication. In this case the t-test is used while drawing on the Central Limit Theorem in the sense that the amount of observations should be large enough to avoid breaking the assumption of normality which could interfere with the meaningfulness of the results.

Note: Since these tests compare between two groups only phase 1 (preference for self-employment) and phase 2 (actual employment) are looked at with these tests. Phase 3 does not have a binary division in two groups but instead an ordinal division and because of this, this phase is not included in these preliminary tests.

Phase 1

Table 1

Education with regards to Preference for Self-employment

Two-sample Kolmogorov-Smirnov test for equality of distribution functions

	Strict Definition			Wide Definition		
Group	D	P-value	Corrected	D	P-value	Corrected
No preference for self-employment:	0.018	0.057		0.021	0.019	
Preference for self-employment:	-0.021	0.023		-0.015	0.145	
Combined K-S:	0.021	0.046	**0.044**	0.021	0.038	**0.037**
Observations:	17172			17817		

With a 5% confidence level (although not a 1% confidence level) the test rejects the null-hypothesis and shows that there is a significant difference in distribution of education between individuals that prefer self-employment and individuals that do not prefer self-employment over becoming an employee. Similar to the test with actual Self-employment (which will be shown in phase two) this by itself does not mean there is an effect of education, in this case on the preference self-employment, but it does show that an effect of education is possible. This means that a relation between the preference for self-employment and education is to be expected as well, since the two groups have a significantly different distribution. It should be noted again however that the nature of the possible relationship can not be inferred from this test and also that the significance in this case is lower than in the case of actual self-employment.

Table 2

Education with regards to Preference for Self-employment

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

	Wide definition	Strict definition
Obs Employee	8522	10180
Obs Self-employed	2038	7637
Total Observations	10560	17817
z-value	-.847	-2.213
Prob > z	**0.397**	**0.027**

Ho: Education of an employee = Education of a self-employed person

For the strict definition of preference for self-employment the test shows that with a 5% confidence level that the null-hypothesis is not rejected and therefore shows that there is not a significant difference in medians of education between individuals that have a preference for self-employment and individuals that do not have a preference for self-employment. This would suggest against a significant relation between education and preference for employment, contrasting the suggestions of the Kolmogorov-Smirnov test.

For the wide definition of preference for self-employment the test shows that with a 5% confidence level that the null-hypothesis is rejected. This would suggest there is a significant relation between education and preference for employment, supporting the suggestions of the Kolmogorov-Smirnov test.

The difference between these two results can be explained by the difference in definition of a preference for self-employment. In the strict definition there is either a choice for self-employment or a choice for preferring to become an employee while preferring both or neither are not counted. In the wide definition there is either a preference for self-employment where the second category is everything else. In this case it shows that there might be a significant relation between education and whether people have a distinct preference for self-employment over all other options, but not necessarily a significant relation between education and preference of self-employment over the preference of becoming an employee.

Table 3

Education with regards to Preference for Self-employment

Variance ratio test

	Strict Definition				Wide Definition			
Group	Obs	Mean	Std. Err.	Std. Dev.	Obs	Mean	Std. Err.	Std. Dev.
Employee	9535	13.622	0.067	6.496	10180	13.508	0.064	6.477
Self-employed	7637	13.570	0.071	6.188	7637	13.570	0.071	6.188
combined	17172	13.599	0.049	6.360	17817	13.534	0.048	6.417
f-value	1.102				1.096			
p-value	**0.000**				**0.000**			

ratio = sd(0) / sd(1)

Ho: ratio = 1 Ha: ratio ≠ 1

This test determines whether the variances of the two groups (preference for self-employment and no preference for self-employment) differ significantly with respect to education. It shows that with both definitions with a 5% confidence level (as well as a 1% confidence level) the null-hypothesis can be rejected. This means the ratio is significantly different from 1 and

therefore the variances are significantly different. This in turn means that for the upcoming t-test the two-sample t test with unequal variances should be used for both definitions.

Table 4

Education with regards to Preference for Self-employment

Two-sample t test with unequal variances

	Strict Definition				Wide Definition			
Group	Obs	Mean	Std. Err.	Std. Dev.	Obs	Mean	Std. Err.	Std. Dev.
Employee	9535	13.62	0.067	6.496	10180	13.51	0.064	6.477
Self-employed	7637	13.57	0.071	6.188	7637	13.57	0.071	6.188
combined	17172	13.6	0.049	6.360	17817	13.53	0.048	6.355
t-value	0.531				0.655			
p-value	**0.596**				**0.513**			

diff = mean(Employee) - mean(Self-emp)

Ho: diff = 0 Ha: diff ≠ 1

The two tests above show a similar result for both definitions of the preference for self-employment. In both cases the null-hypothesis can not be rejected with a 5% confidence level. Therefore it shows that there is no significant difference in medians of education between individuals that have a preference for self-employment and individuals that do not have a preference for self-employment (with both types of the definition of this variable). This would suggest against a significant relation between education and actual employment. For the strict definition this supports the Wilcoxon rank-sum (Mann-Whitney) test and contradicts the Kolmogorov-Smirnov test, suggesting there is no difference between both groups with regards to education and thus less support that there would be a significant relationship. Nonetheless this contradicting result was to be expected if the previous literature regarding a relation between education and (a preference for) entrepreneurship is taken into account and as such further research is warranted. This will be done with a binary logistic regression. For the wide definition the results from this t-test contradict both the Wilcoxon rank-sum (Mann-Whitney) test and the Kolmogorov-Smirnov test, which means that two of the three tests are suggesting there is a possible relationship between education and preference for entrepreneurship in the sense that there is a difference between the two groups with regards to education.

Phase 2

Table 5

Education with regards to Actual Self-employment

Two-sample Kolmogorov-Smirnov test for equality of distribution functions

	Strict Definition			Wide Definition		
Group	D	P-value	Corrected	D	P-value	Corrected
No preference for self-employment:	0.015	0.465		0.083	0.000	
Preference for self-employment:	-0.061	0.000		-0.003	0.973	
Combined K-S:	0.061	0.000	**0.000**	0.083	0.000	**0.000**
Observations:	10560			10560		

With a 5% confidence level (as well as a 1% confidence level) the test rejects the null-hypothesis and shows that there is a significant difference in distribution of education between employees and self-employed individuals. This by itself does not mean there is an effect of education on actual self-employment but it does show that such an effect is possible. This means a relation between the education and the actual self-employment is to be expected since the two groups have a significantly different distribution, although the nature of the possible relationship cannot be inferred from this test.

Table 6

Education with regards to Actual Self-employment

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

	Wide definition	Strict definition
Obs Employee	8522	16286
Obs Self-employed	2038	2038
Total Observations	10560	18324
z-value	2.609	-7.635
Prob > z	**0.009**	**0.000**

Ho: Education of an employee = Education of a self-employed person

The test shows that with a 5% confidence level (and 1% confidence level) that the null-hypothesis is rejected and shows that there is a significant difference in medians of education between individuals that are self-employment and individuals that are not self-employed. This would suggest that there could be a significant relation between education and actual employment, confirming the suggestion from the Kolmogorov-Smirnov test. The nature of this possible relationship can not be inferred from this test either but it does make the existence of such a relationship more likely since both tests show a significant difference between self-employed cases and employees with regards to education.

Table 7

Education with regards to Actual Self-employment

Variance ratio test

	Strict Definition				Wide Definition			
Group	Obs	Mean	Std. Err.	Std. Dev.	Obs	Mean	Std. Err.	Std. Dev.
Employee	8522	14.531	0.062	5.713	16286	13.476	0.050	6.392
Self-employed	2038	14.238	0.146	6.574	2038	14.238	0.146	6.574
combined	10560	14.475	0.057	5.890	18324	13.561	0.047	6.417
f-value	0.755				0.946			
p-value	**0.000**				**0.088**			

ratio = sd(0) / sd(1)

Ho: ratio = 1 Ha: ratio ≠ 1

This test determines whether the variances of the two groups (self-employed and employees) differ significantly with respect to education. For the strict definition It shows that with a 5% confidence level the null-hypothesis can be rejected. This means the ratio is significantly different from 1 and therefore the variances are significantly different.

For the wide definition It shows that with a 5% confidence level the null-hypothesis can not be rejected. This means the ratio is not significantly different from 1 and therefore the variances are not significantly different.

This means that for the upcoming t-test the two-sample t test with unequal variances should be used for the strict definition and the two-sample t test with equal variances can be used for the wide definition of actual occupation.

Table 8

Education with regards to Actual Self-employment

Two-sample t test with unequal variances for strict definition

Two-sample t test with equal variances for wide definition

	Strict Definition				Wide Definition			
Group	Obs	Mean	Std. Err.	Std. Dev.	Obs	Mean	Std. Err.	Std. Dev.
Employee	8522	14.531	0.062	5.713	16286	13.48	0.050	6.392
Self-employed	2038	14.238	0.146	6.574	2038	14.24	0.146	6.574
combined	10560	14.475	0.057	5.890	18324	13.56	0.047	6.417
t-value	1.853				-5.058			
p-value	**0.064**				**0.000**			

diff = mean(Employee) - mean(Self-emp)

Ho: diff = 0 Ha: diff ≠ 1

For the strict definition, the null-hypothesis can not be rejected with a 5% confidence level, and this therefore does not show that there is a significant difference in medians of education between self-employed people and employees, contradicting the suggestion of the Two-sample Wilcoxon rank-sum (Mann-Whitney) test and the suggestion from the Kolmogorov-Smirnov test. However since two of the three tests do suggest a relationship by showing a significant difference in the group of employees as opposed to self-employed people with regards to education and only one test does not suggest this, it does warrant further research which will be done utilizing binary logistic regression analysis.

For the wide definition the null-hypothesis can be rejected with a 5% confidence level (as well as a 1% confidence level) and this therefore it shows that there is a significant difference in medians of education between self-employed individuals and non-self-employed people. This would suggest a significant relation between education and actual employment, confirming the suggestion of the Two-sample Wilcoxon rank-sum (Mann-Whitney) test and the suggestion from the Kolmogorov-Smirnov test.

6. Regression Results

The results of the different regression models of the three phases are presented in this part of the thesis. For each phase three different models were created of which the full model on the right is the final model chosen for the analysis. The other two models per phase are shown in the same table for comparison to support why the final full model was chosen.

Strict and wide definition

For all phases, two definitions were possible for the variable "Actual occupation" and "Preference for entrepreneurship". These definitions are a strict definition or a wide definition as mentioned previously in chapter 4. So for the models mentioned a variation in the strict form was created as well as a variation in the wide form. The strict definition models are presented below as they were found to have the most consistent results, although a comparison of the final model of the wide and strict version is shown to underline the importance of clear uniform definitions.

Basic model, Model including age/100 squared and Full model

For all phases there were three possible models for the strict definition of the dependent variable (preference for self-employment, actual self-employment and entrepreneurial engagement levels respectively). The three possible models, inspired by previous literature are separated into:

1. A basic model including the commonly used independent variables, combining both socio-demographic variables and perceptual/preference variables but not encompassing any quadratic variables (and the associated possible non-linear effects).
2. A model that is identical to the first model but encompasses an extra variable regarding age which is the age divided by 100 and then squared to provide a useful quadratic variable that captures possible extra non-linear effects of age.
3. The full final model used which is identical to the second model but with the addition of an extra variable for education, which is education divided by 100 and then squared to provide a useful quadratic variable that captures possible extra non-linear effects of education.

After the table with the results according to the above three models another table is shown for phase one and two with a comparison of the full model of both the wide and strict definition of the dependent variable for comparison purpose to see whether the definition significantly changes the results. The strict definition however will be chosen for the final conclusion due to the fact that this definition is supported by previous literature and provides more consistent results.

6.1. Results and analysis of phase 1

Phase 1 regards the possible effects of education and the control variables on the preference for self-employment.

Table 9
Education with regards to Preference for Self-employment (Strict Definition)

	Model 1 (Basic Model)			Model 2 (Including age/100 sq.)			Model 2 (Full Model)		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Gender	.517 (.038)	0.000	13.56	.514 (.038)	0.000	13.47	.512 (.38)	0.000	13.40
Age	-.005 (.001)	0.000	-4.2	-.020 (.007)	0.004	-2.86	-.021 (.007)	0.003	-2.94
Age/100 sq.	- (-)	-	-	1.487 (.692)	0.031	2.15	1.585 (.695)	0.023	2.28
Education	-.002 (.003)	0.483	-0.7	-.002 (.003)	0.553	-0.59	.012 (.010)	0.215	1.24
Educ. /100 sq.	- (-)	-	-	- (-)	-	-	-2.519 (1.686)	0.135	-1.49
Location	.040 (.026)	0.120	1.56	.042 (.026)	0.109	1.6	.044 (.026)	0.088	1.71
S.E. Father	.222 (.048)	0.000	4.62	.218 (.048)	0.000	4.52	.217 (.048)	0.000	4.50
S.E. Mother	.149 (.066)	0.023	2.27	.151 (.066)	0.022	2.29	.154 (.066)	0.020	2.33
Lack of Fin. Sup.	-.014 (.049)	0.780	-0.28	-.017 (.049)	0.732	-0.34	-.015 (.049)	0.764	-0.30
Admin. Complex.	-.152 (.045)	0.001	-3.35	-.153 (.045)	0.001	-3.37	-.152 (.045)	0.001	-3.34
Lack of Info.	.102 (.041)	0.012	2.52	.099 (.041)	0.015	2.43	.102 (.041)	0.013	2.50
Risk Tolerance	.247 (.040)	0.000	6.26	.248 (.040)	0.000	6.29	.245 (.040)	0.000	6.17
Country Dummies	Included (Wald Chi2 = 414.20, p = 0.000)			Included (Wald Chi2 = 414.61, p = 0.000)			Included (Wald Chi2 = 415.94, p = 0.000)		
No of obs.	12522			12522			12522		
Log likelihood	-8263.272			-8260.960			-8259.799		
Prob > Chi2	0.000			0.000			0.000		
McFadden R2 (adj.)	0.046 (0.042)			0.047 (0.042)			0.047 (0.042)		
Nagelkerke R2	0.083			0.083			0.083		

Table 9 presents the results of a logit estimation where the dependent variable is the preference for self-employment, with as independent variable education and the common control variables as discussed previously.

The last column refers to the full model where the explanatory variables (including a quadratic variable for age as well as education) are used. The previous two columns correspond to a reduced form of the model, where the quadratic explanatory variable of education and age respectively are omitted from the equation.

Education

According to these estimations education (as normal or quadratic variable) has no significant impact on the preference for being self-employed. This is the case in all the three models although it is more prominent in the reduced form models compared to the full model.

Apparently the amount of education does not have a significant effect on whether or not an individual chooses to become self-employed as opposed to choosing to become an employee. The logic behind this can be that the preference to become self-employed is not a taught

aspect and it can not be created by education if it's not already present. Similarly the other way around education would not be able to negate this preference. In this light it would seem education is irrelevant when looking at the preference in the strict sense. Any attempt or policy to enhance the entrepreneurial preference of individuals would therefore be better of not being in the way of education as there would be no significant effect.

However, this does not mean that education in this context has no significance. There can be an effect on actual self-employment (which according to the table would not specifically run through preference for self-employment). This effect will be looked at in phase 2. Similarly an effect could exist where education has a significant relation with the performance of an entrepreneur. This however falls outside the scope of this thesis.

If the effect were significant there would be a quadratic relation where education would at first have a small positive effect (.012) and this increase the preference. After the cut-off point which is found to be at 6.28 years[11], this effect would be decrease by the quadratic negative effect which is a lot stronger (-2.519) after which the negative effect would prevail. The resulting effect would be an inverted u-shape with the cut-off point of 6.28 being the highest point of the relationship's effect. Regarding the relevant range, the negative quadratic effect is the important effect.
A possible reason for this type of relationship would be that the negative quadratic effect could point out that a large amount of education would sway the preference away from self-employment because other potential factors are taken into account such as the rate of return on investment in education or a better view of the potential possibilities and a better understanding of one's own capacities and what would be best suited (which may or may not be a misguided view). There could be a better rate of return in employment (or at least this image could exist) leading to the negative effect.

The results from other studies where education is used as an explanatory variable for the preference for self-employment are not always comparable to these, since in a part of the previous research the variable education was constructed with 3 dummy levels as opposed to using linear years of education.

As stated before, the variable education and the squared variable of education are not significant, indicating that the amount of education does not significantly change the preference for entrepreneurship. This preference stems from other factors, including the significant control variables in the results such as gender, age (and a quadratic variable of age), the self-employment status of father and mother as well as perceptual variables such as the perceived lack of information and risk tolerance.

Control variables

The control variables are discussed below in a similar order as they are found in table 9, apart from the country variables which are only shortly mentioned as first item.

[11] The turning point is 6.28 years of age when full-time education was finished (not counting the first six years as stated in the data discussion section regarding educaton.). This was calculated by the odds ratio of education divided by 2 x the odds ratio of education^2 within the context of setting the first order derivative of preference for self-employment with regards to education to 0.

Country variables
Country dummies were also found to be significant although the specific coefficients per country were omitted in the results-table as they are not an important focus of this thesis.)

Gender
Gender is found to have a significant positive effect (.512) on the probability of preferring to be self-employed. Apparently male individuals have a higher probability to prefer to be self-employed when compared to their female counterparts. If this difference does not remain similar when the actual occupational choice is being looked at then the discrepancy between the preference and the actual status would be stronger for men than for women.

Age
Age as a linear variable is found to have a significant negative effect (-.021), indicating that when the individual gets older the preference for entrepreneurship decreases. The significant positive effect (1.585) of the quadratic variable of age indicates the relationship to be a more complex u-shape. A cut-off point is found to be at 0.1 years[12], indicating that the negative linear effect is irrelevant for the relevant range and age has an increasing positive effect on preference for entrepreneurship.

Location
Location shows a small positive effect (.044) which is significant (at a 90% confidence level but not at a 95% confidence level). Since this variable increases the further the individual is located in a rural, less-populated area and decreases the further an individual lives in a metropolitan area, this positive effect indicates that individuals in a rural area are more inclined to prefer self-employment as opposed to people in a metropolitan area.[13]
A logical intuitive reasoning could be that the sense of freedom could play a part for both the choice of location and the preference for entrepreneurship and as such the psychology of the individual plays a role in this relationship.

Self-Employment of father and mother
Not surprisingly both the self-employment status of the father as well as the self-employment status of the mother both have a significant positive effect (.217 and .154 respectively) on the preference for self-employment. This indicates that if one or both of the parents are self-employed the preference for self-employment increases. Considering the influence of the parents on a child in the context of the nature versus nurture-theme, it is generally accepted that the upbringing from parents has a significant effect on the preferences and personality of the child.

Lack of financial support
It is found that lack of financial support as a perceptual variable has no significant relation with the preference for self-employment. This may perhaps be indicating that financial constraints are either seen as not-present when an individual considers the hypothetical case

[12] The turning point of 0.1 years of age was calculated by the odds ratio of age divided by 2 x the odds ratio of Age^2 within the context of setting the first order derivative of preference for self-employment with regards to age to 0.
[13] It must be noted that agricultural occupations and such are logically more commonly found in rural areas and these occupations are categorized as self-employed. Someone who has a preference for such a profession would therefore be more likely to be located in a more rural area. This might have influenced the results if a significant part of the self-employment-preferring individuals in the sample have a preference for the agricultural sector.

of becoming an entrepreneur and gauges their own preference, is overlooked or is seen as something that can be overcome and therefore discarded as irrelevant.

Perception of administrative complexity
Perception of administrative complexity is found to have a significant negative effect (-.152) on the preference for self-employment. The negative effect of perception of administrative complexity can be deduced to stem from the possibility that the individuals either see self-employment as inherently linked with administrative complexity or from self-employment truly being administratively complex.

Perceived lack of information
Concerning the difficulty of finding information related to entering self-employment, the positive effect of this (.102) on the preference for self-employment indicates that when a lack of information is perceived, the preference for self-employment rises. This could mean that the apparent lack of information is seen as a positive signal that not much information might be required or it might indicate that individuals preferring self-employment have a higher need for information and as such as more likely to find the available information lacking.

Risk tolerance
Risk tolerance is found to have a significant positive effect (.245) on the preference for self-employment. This indicates that the more risk tolerant the individual is, the more inclined they will be to have a preference for self-employment which fits the intuitive notion that self-employment has inherently more risks than employment and therefore more risk tolerant individuals prefer this career choice.

In essence for policy the results above for the variables are most relevant if preference for self-employment is chosen as the measure for any policy and that would assume that the actual choice for self-employment significantly depends on this which is shown in phase 2.

Strict & Wide definition

Since both definitions have a logical reasoning behind them and could provide information about the relationship between the preference for self-employment and education they are both shown in table 10. The full model is chosen as the model to compare between the two definitions since this model was earlier found to be the best-fitting of the three possible models. It should be mentioned that the number of observations between the two definitions differs slightly (356 observations difference) and as such any conclusions from a comparison should be looked at critically. Nonetheless a comparison remains useful since the difference is only 2.8% when the total amount of observations is taken into account.

Table 10

Education with regards to Preference for Self-employment (Strict Definition & Wide Definition)

	Strict Definition (Full Model)			Wide Definition (Full Model)		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Gender	.512 (.38)	0.000	13.40	.501 (.038)	0.000	13.29
Age	-.021 (.007)	0.003	-2.94	-.019 (.007)	0.007	-2.68
Age/100 sq.	1.585 (.695)	0.023	2.28	1.216 (.683)	0.075	1.78
Education	.012 (.010)	0.215	1.24	.015 (.010)	0.129	1.52
Educ. /100 sq.	-2.519 (1.686)	0.135	-1.49	-2.728 (1.677)	0.104	-1.63
Location	.044 (.026)	0.088	1.71	.040 (.026)	0.123	1.54
S.E. Father	.217 (.048)	0.000	4.50	.217 (.047)	0.000	4.57
S.E. Mother	.154 (.066)	0.020	2.33	.138 (.065)	0.033	2.14
Lack of Fin. Sup.	-.015 (.049)	0.764	-0.30	.004 (.048)	0.935	0.08
Admin. Complex.	-.152 (.045)	0.001	-3.34	-.153 (.045)	0.001	-3.41
Lack of Info.	.102 (.041)	0.013	2.50	.087 (.040)	0.030	2.16
Risk Tolerance	.245 (.040)	0.000	6.17	.245 (.039)	0.000	6.27
Country Dummies	Included (Wald Chi2 = 415.94, p = 0.000)			Included (Wald Chi2 = 411.78, p = 0.000)		
No of obs.	12522			12878		
Log likelihood	-8259.799			-8478.480		
Prob > Chi2	0.000			0.000		
McFadden R2 (adj.)	0.047 (0.042)			0.046 (0.042)		
Nagelkerke R2	0.083			0.083		

Table 10 presents the results of a logit estimation where the dependent variable is the preference for self-employment, with as independent variable education and the common control variables as discussed previously.

Both columns refer to the full model where the explanatory variables (including a quadratic variable for age as well as education) are used. The first column is identical to the right column of table 9 and corresponds to the model with a strict definition of preference for self-employment. The right column corresponds to the model with a wide definition of preference for self-employment. This to ease a comparison between the two definitions for the case of preference for self-employment.

When comparing the strict definition to the wide definition the following differences can be observed, taking the strict definition as baseline:

- The effects of gender, age in the linear form, administrative complexity, lack of info and risk tolerance barely change. (Slight increases or decreases per variable, but no large differences that stand out.)
- The effect of the quadratic form of age becomes less significant, below the 95% confidence level.
- Education becomes slightly more significant although still below the 90% confidence level for both the linear and quadratic form of the variable of education.
- Location becomes less significant, now below the 90% confidence level.
- Self-employment of father and mother become slightly less strong (in the case of the mother) or remain the same (in the case of the father).

- Perceived lack of financial support becomes even less significant, despite already having been very insignificant and the sign of the effect changes. However due to the extreme p-value, the coefficient most likely doesn't add any useful information to the analysis.
- It should be noted that the McFadden R2 is slightly higher in the case of the strict definition but due to the small change in number of observations this might not be a significant change.

In the wide definition the preference for self-employment is seen as not merely the choice between a 'preference for self-employment' or a 'preference for employment' but the choice between 'a preference for self-employment' or 'a preference for employment, a preference for both or for neither'. With that in mind the above would seem to indicate that with the wide definition, the quadratic effect of age becomes less pronounced, in other words, it is less likely to find older individuals having a lower preference than with the strict definition. This may be because of a change of preference for either both or neither which is stronger at an older age. Education becoming slightly more significant indicates that perhaps the amount of education matters more in the context of preferring self-employment over all other options.
In any case the comparison offers that the definition chosen can influence the results more than marginally and as such comparisons to previous studies should take this into account, meaning that the definitions used per study should be compared to avoid wrongful comparisons. It should be noted that in most studies a strict definition was followed for as far as the articles were clear about their used definitions and terminology.

6.2. Results and analysis of phase 2

Phase 2 regards the possible effects of education and the control variables on actual self-employment in the strict definition.

Table 11

Education with regards to Actual Self-employment (Strict Definition)
Using Preference for self-employment in the strict definition as control variable.

	Model 1 (Basic Model)			Model 2 (Including age/100 sq.)			Model 3 (Full Model)		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Preference (Strict)	1.602 (.069)	0.000	23.25	1.607 (.069)	0.000	23.31	1.614 (.069)	0.000	23.36
Gender	0.448 (.063)	0.000	7.16	0.454 (.063)	0.000	7.24	.457 (.063)	0.000	7.28
Age	.030 (.003)	0.000	11.08	.062 (.016)	0.000	3.94	.067 (.016)	0.000	4.20
Age/100 sq.	- (-)	-	-	-3.515 (1.692)	0.038	-2.08	-4.187 (1.713)	0.015	-2.44
Education	-.009 (.005)	0.083	-1.73	-.009 (.005)	0.079	-1.76	-.053 (.017)	0.002	-3.17
Educ. /100 sq.	- (-)	-	-	- (-)	-	-	7.853 (2.855)	0.006	2.75
Location	.131 (.043)	0.002	3.03	.128 (.043)	0.003	2.96	.119 (.043)	0.006	2.75
S.E. Father	.570 (.074)	0.000	7.68	.579 (.074)	0.000	7.78	.579 (.074)	0.000	7.78
S.E. Mother	.203 (.097)	0.037	2.09	.200 (.097)	0.040	2.06	.190 (.098)	0.051	1.95
Lack of Fin. Sup.	-.092 (.076)	0.227	-1.21	-.092 (.076)	0.229	-1.20	-.101 (.076)	0.186	-1.32
Admin. Complex.	-.191 (.072)	0.008	-2.64	-.186 (.072)	0.010	-2.57	-.189 (.072)	0.009	-2.62
Lack of Info.	.147 (.068)	0.031	2.16	.153 (.068)	0.026	2.23	.144 (.068)	0.035	2.11
Risk Tolerance	.133 (.066)	0.043	2.02	.130 (.066)	0.048	1.98	.143 (.066)	0.031	2.16
Country Dummies	Included (Wald Chi2 = 145.06, p = 0.000)			Included (Wald Chi2 = 145.05 p = 0.000)			Included (Wald Chi2 = 144.95, p = 0.000)		
No of obs.	7536			7536			7536		
Log likelihood	-3320.044			-3317.852			-3314.051		
Prob > Chi2	0.000			0.000			0.000		
McFadden R2 (adj.)	0.158 (0.148)			0.159 (0.148)			0.159 (0.149)		
Nagelkerke R2	0.235			0.236			0.237		

Table 11 presents the results of a logit estimation where the dependent variable is the actual self-employment, with as independent variables education, preference for entrepreneurship and the common control variables as discussed previously.

The last column refers to the full model where the explanatory variables (including a quadratic variable for age as well as education) are used. The previous two columns correspond to a reduced form of the model, where the quadratic explanatory variable of education and age respectively are omitted from the equation.

Education

According to the results in the table, education (as a normal and as a quadratic variable) has a significant impact on the actual self-employment of individuals. This is the case in all three models, although the effects are strongest in the full model when compared to the reduced form models. This is entirely different from the insignificant results for this independent variable when looking at the preference for self-employment.

The linear variable for amount of education has a significant small negative effect (-.053) indicating that when an individual has more years of education the probability of being self-employed decreases. The logic behind this can be that perhaps the return to education could be bigger for employment than for self-employment at higher levels of education, assuming that the "employment status choice" of the individual is based on the returns to education. This assumption comes from the intuitive logic that a person prefers higher earnings over lower earnings (if risks are more or less equal in their eyes or not relevant). It could also indicate what Van der Sluis et al. (2008) call the "Bill Gates effect" indicating that it might be common for a nascent entrepreneur to drop out of full-time education. Alternatively, the aforementioned article indicates it could be that screening in the wage sector pushes low educated individuals into entrepreneurship more as opposed to highly educated individuals. This assumes that the small negative linear effect occurs within the relevant range, which is shown to be untrue in the next paragraph.

Taking the quadratic variable of education into account, it has a significant and strong positive effect (7.853) which indicates that the relation between education and actual self-employment in its totality is of a quadratic nature where the small negative effect (-.053) is negated by the much stronger positive effect (7.853) after a certain cut-off point which is found to be at 0 years[14]of age (not including the first six years, as stated in the description of the education-variable). This means that only the positive effect of the quadratic variable is meaningful in the relevant range, indicating that the possible reasoning in the previous paragraph such as the "Bill Gates effect" are not proven or found in this result.

The resulting total effect of education is a u-shape with the cut-off point of 0 years being the lowest point of the relationship's effect. After the cut-off point the positive effect would prevail creating a situation where more education strongly increases the probability of being self-employed, which happens at a slightly higher rate as an individual has more education.

A possible reason for the positive effect could be that at there may actually be a higher return to education for self-employment and therefore this option becomes more appealing with higher education. Another possibility is that at higher levels of education, self-employment indicates a highly educated individuals such as doctors etc. starting their own practice, not necessarily just because of a higher return but for the purpose of status or because it's common practice. This would explain why higher levels of education have a positive effect on entrepreneurship.

When taking the results from phase 1 into account again it seems that the effect of education is overall more positive in the relevant area (the levels of education that are most observed which lie above the cut-off point) and thus increases the probability of self-employment. This positive effect of education does not come through the preference for self-employment as this preference isn't significantly affected by education, so it is a direct effect of education of the probability of self-employment.

[14] The turning point is 0 years of age when full-time education was finished (not counting the first six years as stated in the data discussion section regarding educaton.). This was calculated by the odds ratio of education divided by 2 x the odds ratio of education^2 within the context of setting the first order derivative of preference for self-employment with regards to education to 0.

Control variables

The control variables are discussed below in a similar order as they are found in table 11, apart from the country variables which are only shortly mentioned as first item.

Country variables
Country dummies were also found to be significant although the specific coefficients per country were omitted in the results-table as they are not an important focus of this thesis.

Preference for self-employment
The preference for self-employment is very significant in explaining the actual self-employment of an individual. The effect is strong and positive (1.614) and indicates that the more a person prefers self-employment, the higher the probability that that person will actually become self-employed. The strong positive effect of preference for self-employment on actual self-employment intuitively makes intuitive sense.

Gender
Gender is found to have a significant positive effect (.457) on the probability being self-employed . male individuals have a higher probability to be self-employed when compared to their female counterparts. This difference seems similar to the results of phase 1 (the preference for self-employment). This indicates that the discrepancy between the preference and the actual status is not per definition stronger for men than for women.

Age
Similar to the positive effect on the preference for self-employment in the relevant range, the linear effect of age on actual self-employment is positive (.067). This indicates that as an individual gets older the probability of being self-employed becomes larger.
Apart from the linear age-variable, there is also a significant negative effect (-4.187) of the quadratic variable of age which makes this relationship a more complicated inverted u-shape. The cut-off point for this inverted u-shape is found to be at 35.19 years[15], indicating that before this age the effect is positive, after this age the effect turns negative. An intuitive reasoning for this might be that after that age there could be practical inabilities to self-employment (a higher probability to have fixed costs and more responsibilities).

Location
Location shows a significant positive effect (.119) which is slightly stronger than the effect of location of preference for self-employment. The positive effect of this variable indicates that the further an individual is located in a rural, less-populated area the higher the probability that the individual is self-employed. This relationship could be one of both ways: it could be that individuals who become entrepreneurs tend to move to more rural areas or it could be that individuals who live in rural areas are more probable to be self-employed.[16]

[15] The turning point of 35.19 years of age was calculated by the odds ratio of age divided by 2 x the odds ratio of Age^2 within the context of setting the first order derivative of preference for self-employment with regards to age to 0.
[16] It should be noted, just as in phase 1, that agricultural occupations such as farmers etc are logically more likely to be found in rural areas. Therefore the positive effect of location could be, in part, due to the agricultural and similar professions which classify as self-employed in the data-set.

Self-Employment of father and mother
Both the self-employment status of the father as well as the self-employment status of the mother both have a significant positive effect (.217 and .154 respectively) on actual self-employment. This indicates that if one or both of the parents are self-employed the probability that their child will be self-employed increases. It should be noted that the effect of the mother's employment status was less significant then the effect of the father's employment status: 90% confidence level as opposed to a 95% confidence level.
An intuitive idea behind the above results could be that there is logically a relatively strong influence of parents on their children in the context of the nature versus nurture-theme. It is generally accepted that the upbringing from parents has a significant effect on the preferences and personality of the child.

Lack of financial support
Lack of financial support as a perceptual variable is found to have no significant effect on actual self-employment. This could indicate that financial constraints are not present when an individual chooses to be self-employed as there is no significant effect in a positive or negative sense.

Perception of administrative complexity
When it comes to the perception of administrative complexity, this is found to be significant and to have a negative effect (-.189). This negative effect could indicate that actual self-employment is inherently linked with a large amount of administration of a complex nature or that self-employment is seen as being linked in that way with administrative complexity which influences the probability that an individual chooses to become self-employed because of a dislike for administrative complexity.

Perceived lack of information
When it comes to the difficulty of finding information related to entering self-employment, in other words perceived lack of information, there is a significant positive effect (.144) on the probability of actual self-employment. This indicates that when a lack of information is perceived by an individual , the probability of this person being self-employment rises. This could mean the people who become self-employed have a higher need for information about self-employment than the general population. Due to this heightened informational need they would be more likely to perceive the available information as lacking when compared to their counterparts who do not have this informational need as strongly. This then perceived lack of information would be more present in individuals who are more probable to become self-employed and this would thus explain the positive significant effect.

Risk tolerance
The results show that risk tolerance has a significant positive effect (.143) on actual self-employment. This effect is smaller than the effect on the preference for self-employment. The fact that the effect is positive indicates that the more risk tolerant the individual is, the more probable it is that they are self-employed. This fits the intuitive notion that self-employment has inherently more risks than employment which is the classic view that self-employed individuals/entrepreneurs are risk-takers. With this view it would make sense that individuals that are more risk tolerant would therefore be more inclined to be(come) self-employed. Previous studies in the literature such as Brown et al. (2008) find similar effects for risk tolerance, confirming this result.

Strict & Wide definition

Since both definitions have a logical reasoning behind them and could provide information about the relationship between the preference for self-employment and education they are both shown in table 12. The full model is chosen as the model to compare between the two definitions since this model was earlier found to be the best-fitting of the three possible models. It should be mentioned that the number of observations between the two definitions differs to a large extent (4846 observations difference) and as such any conclusions from a comparison should be looked at critically.

Table 12

Education with regards to Actual Self-employment (Strict Definition & Wide Definition)

	Strict Definition (Full Model)			Wide Definition (Full Model)		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Preference (Strict)	1.614 (.069)	0.000	23.36	1.469 (.066)	0.000	22.33
Gender	.457 (.063)	0.000	7.28	.741 (.059)	0.000	12.65
Age	.067 (.016)	0.000	4.2	.216 (.015)	0.000	13.98
Age/100 sq.	-4.187 (1.713)	0.015	-2.44	-24.354 (1.622)	0.000	-15
Education	-.053 (.017)	0.002	-3.17	-.032 (.015)	0.035	-2.11
Educ. /100 sq.	7.853 (2.855)	0.006	2.75	7.177 (2.515)	0.004	2.85
Location	.119 (.043)	0.006	2.75	.057 (.040)	0.156	1.42
S.E. Father	.579 (.074)	0.000	7.78	.549 (.069)	0.000	7.96
S.E. Mother	.190 (.098)	0.051	1.95	.191 (.089)	0.033	2.13
Lack of Fin. Sup.	-.101 (.076)	0.186	-1.32	-.131 (.072)	0.067	-1.83
Admin. Complex.	-.189 (.072)	0.009	-2.62	-.166 (.068)	0.014	-2.45
Lack of Info.	.144 (.068)	0.035	2.11	.097 (.064)	0.129	1.52
Risk Tolerance	.143 (.066)	0.031	2.16	.192 (.062)	0.002	3.13
Country Dummies	Included (Wald Chi2 = 144.95, p = 0.000)			Included (Wald Chi2 = 114.65, p = 0.000)		
No of obs.	7536			12382		
Log likelihood	-3314.051			-4056.7089		
Prob > Chi2	0.000			0.000		
McFadden R2 (adj.)	0.159 (0.149)			0.161 (0.152)		
Nagelkerke R2	0.237			0.218		

Table 12 presents the results of a logit estimation where the dependent variable is the actual self-employment, with as independent variable education and the common control variables as discussed previously.

Both columns refer to the full model where the explanatory variables (including a quadratic variable for age as well as education) are used. The first column is identical to the right column of table 11 and corresponds to the model with a strict definition of preference for self-employment. The right column corresponds to the model with a wide definition of preference for self-employment. This to ease a comparison between the two definitions for the case of actual self-employment.

When comparing the strict definition to the wide definition the following differences can be observed, taking the strict definition as baseline:

- Education becomes less significant but still significant at 95% confidence level with a slightly weaker effect. The effect of the quadratic form of education also becomes slightly weaker.
- Gender and age have slight increases in their effects, where-as the quadratic form of age has a very strong increase in its effect.
- The effects of preference for entrepreneurship and the self-employment status of the father are slightly less strong.
- Perceived lack of financial support has a stronger effect which is now significant at a 90% confidence level as opposed to not being significant.
- The self-employment status of the mother as well as the perceived administrative complexity have slightly less strong effects and become less significant although still significant at a 95% confidence level.
- The effects of location and perceived lack of information are less strong and no longer significant
- Risk tolerance becomes more significant and has a stronger effect.
- It should be noted that the McFadden R2 is slightly smaller in the case of the strict definition but due to the large difference in number of observations between the two models this might not be significant. Similarly this large difference in number of observations impacts the whole comparison displayed here, and could mean that any conclusions drawn from this should be viewed with caution.

In the wide definition the actual self-employment is seen as not merely the choice between a 'self-employment' or a 'employee' but the choice between 'self-employment' or 'any other employment status choice apart from self-employment'. In the strict definition however the choice is seen as being between 'self-employment' and 'employee'.

With that in mind the above would seem to indicate that with the wide definition, education becomes less significant because of a less strictly defined separation of the two groups. For the same reason the effects would be slightly less strong. The strong increase in the negative effect of the quadratic form of age might be because in the wide definition categories such as "Looking after the home" and "Retired" are also incorporated which are logically more often found with older people.

In any case the comparison offers that the definition chosen can influence the results more than marginally and as such comparisons to previous studies should take this into account, meaning that the definitions used per study should be compared to avoid wrongful comparisons. It should be noted that in most studies a strict definition was followed for as far as the articles were clear about their used definitions and terminology.

6.3. Results and analysis of phase 3

Phase 3 regards the possible effects of education and the control variables on entrepreneurial engagement levels.

Table 13
Education with regards to Entrepreneurial Engagement Levels (Ordinal Regression)
Using Preference for self-employment in the strict definition as control variable.

	Model 1 (Basic Model)			Model 2 (Including age/100 sq.)			Model 3 (Full Model)		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Preference	1.664 (.050)	0.000	33.01	1.699 (.051)	0.000	33.34	1.699 (.051)	0.000	33.34
Gender	.637 (.048)	0.000	13.31	.680 (.048)	0.000	14.07	.6749 (.048)	0.000	14.06
Age	-.026 (.002)	0.000	-16.1	.122 (.010)	0.000	12.11	.122 (.010)	0.000	12.06
Age/100 sq.	- (-)	-	-	-15.948 (1.088)	0.000	-14.7	-15.920 (1.095)	0.000	-14.5
Education	.033 (.004)	0.000	8.17	.029 (.004)	0.000	6.95	.032 (.014)	0.024	2.26
Educ. /100 sq.	- (-)	-	-	- (-)	-	-	-0.564 (2.531)	0.824	-0.22
Location	.033 (.032)	0.308	1.02	.025 (.033)	0.447	0.76	.025 (.033)	0.439	0.77
S.E. Father	.198 (.060)	0.001	3.33	.243 (.060)	0.000	4.04	.243 (.060)	0.000	4.04
S.E. Mother	.117 (.080)	0.141	1.47	.114 (.080)	0.154	1.43	.115 (.080)	0.153	1.43
Lack of Fin. Sup.	-.138 (.059)	0.020	-2.33	-.113 (.060)	0.059	-1.89	-.113 (.060)	0.060	-1.88
Admin. Complex.	-.281 (.055)	0.000	-5.09	-.262 (.056)	0.000	-4.71	-.262 (.056)	0.000	-4.71
Lack of Info.	-.065 (.051)	0.201	-1.28	-.044 (.052)	0.397	-0.85	-.043 (.052)	0.403	-0.84
Risk Tolerance	.267 (.050)	0.000	5.36	.258 (.050)	0.000	5.14	.257 (.050)	0.000	5.11
Country Dummies	Included (Wald Chi2 = 230.24, p = 0.000)			Included (Wald Chi2 = 238.92, p = 0.000)			Included (Wald Chi2 = 237.11, p = 0.000)		
/cut1	.477 (.172)			3.563 (.272)			3.581 (.283)		
/cut2	1.424 (.173)			4.531 (.274)			4.548 (.284)		
/cut3	1.897 (.174)			5.017 (.275)			5.034 (.286)		
/cut4	2.419 (.175)			5.552 (.277)			5.569 (.287)		
No of obs.	8643			8643			8643		
Log likelihood	-8755.877			-8632.897			-8632.872		
Prob > Chi2	0.000			0.000			0.000		
McFadden R2 (adj.)	0.129 (0.125)			0.141 (0.137)			0.141 (0.137)		
Nagelkerke R2	0.287			0.310			0.310		

Table 13 presents the results of an ordinal logit estimation where the dependent variable are the entrepreneurial engagement levels, with as independent variables education, preference for entrepreneurship and the common control variables as discussed previously.

The last column refers to the full model where the explanatory variables (including a quadratic variable for age as well as education) are used. The previous two columns correspond to a reduced form of the model, where the quadratic explanatory variable of education and age respectively are omitted from the equation.

Parallel Regression Assumption

One of the assumptions underlying this ordinal logistic (and ordinal probit) regression is that the relationship between the entrepreneurial engagement levels is the same. This means that the coefficients describing the relationship between the stage "Never thought about it" and the other higher engagement levels are the same as between "Thinking about it" and the other higher engagement levels etc. concluding that the relationship between all pairs of engagement levels is the same. To test this the Brant parallel regression assumption test was performed and can be found in appendix C.

Since the parallel regression assumption is violated for preference, gender, age and age squared, the interpretation of these variables is slightly more difficult and more general. The reason for this is because the coefficient from the results-table does not apply equally to the differences between the five entrepreneurial engagement levels. It gives a general idea however whether an increase in a variable causes an individual to be more likely to be at a lower engagement level or to go to a higher engagement level.
For the variables education, education squared, location, self-employment father, self-employment mother, lack of financial support, administrative complexity, lack of information and risk tolerance the assumption is not violated as the Brant parallel regression assumption table in the appendix C does not show a significant result for these variables at a 95% confidence level.

Education

In the ordinal regression education (as a normal variable) is found to have a significant positive effect on the entrepreneurial engagement levels. This is the case in all three models, although the effects are strongest in the full model when compared to the reduced form models. The variable education in a quadratic form is however found to be extremely non-significant, indicating a more linear, non-quadratic relationship between education and the engagement levels. This is different from the relationship with the binary dependent variable of actual occupation which did have a significant effect from the quadratic education variable.
It should be noted that for education the parallel regression assumption is not violated, meaning that the coefficient is the same across the five engagement levels.

If the quadratic variable of education were significant with its negative effect (-.564) the relation between education and the engagement level would be an inverted u-shape for the total effect, taking both the quadratic and normal variable of education into account.
Since the quadratic variable of education is not significant there is an easier linear positive total effect. However it should be noted that the turning point of this inverted u-shape relationship is at 0.907 years[17],meaning that if the quadratic variable were significant, the relevant range would encompass the negative effect of the quadratic variable.

The linear variable for amount of education has a significant small positive effect (.032) indicating that when an individual has more years of education, he/she is more likely to be in a higher entrepreneurial engagement level. A possible logic for this could be that there are

[17] The turning point is 0.907 years of age when full-time education was finished (not counting the first six years as stated in the data discussion section regarding educaton.). This was calculated by the odds ratio of education divided by 2 x the odds ratio of education^2 within the context of setting the first order derivative of preference for self-employment with regards to education to 0.

certain barriers to moving to a higher engagement level for which a higher education are required to overcome them. Barriers such as this could be that it is more difficult and requires more knowledge to go to the next engagement level. For example: to take actual steps to make a business as opposed to merely thinking about it and having a young business as opposed to taking steps, etc.

When taking the results from the previous two phases into account it seems these results are similar to the result for actual occupation and confirm that education has a positive effect on entrepreneurship. This in the sense that a person with higher education is more likely to be in a higher engagement level and similarly more likely to become an entrepreneur, despite this education not significantly impacting the individual's preference.

Control variables

The control variables are discussed below in a similar order as they are found in table 13, apart from the country variables which are only shortly mentioned as first item.

Country variables
Country dummies were also found to be significant although the specific coefficients per country were omitted in the results-table as they are not an important focus of this thesis.

Preference for self-employment
The preference for self-employment is found to be significant in explaining the engagement levels. The effect is strong and positive (1.699), indicating that the more a person prefers self-employment, the higher the probability that that person will go to a higher entrepreneurial engagement level. It should be noted that the parallel regression assumption was violated for this variable indicating that the effect is not the same between all engagement levels. It can be assumed that its effect would be stronger between the lower engagement levels such as between "Never thought about it", "Thinking about it" and "Taking Steps" and less in the higher engagement levels such as between "Young business" and "Old business".

Gender
Gender is found to have a significant positive effect (.675) on the entrepreneurial engagement levels. Male individuals have a higher probability to be in a higher engagement level when compared to their female counterparts. This difference seems similar to the results of phase 1 (the preference for self-employment) and phase 2 (actual self-employment), confirming that males are more likely to become entrepreneurs and be on a higher engagement level.
It should be noted that for gender the parallel regression assumption is violated, meaning that the coefficient is not the same across the 5 engagement levels.

Age
Age as a linear variable has a significant positive effect (.122) on the entrepreneurial engagement levels. The quadratic effect of age on the entrepreneurial engagement levels is also significant and is a very strong negative effect (-15.920) opposing the earlier linear effect. This shows that the total effect is an inverse u-shape where the decline from the

quadratic effect is very sharp after a cut-off point located at 0 years[18], which marks the highest point of the inverse u-shape and indicates that in the relevant range there is a negative effect. An intuitive rationale for this could be that the increase in responsibilities (financial and otherwise) would limit the individual strongly from going to a higher entrepreneurial engagement level. It should be noted however that for age (in both linear and quadratic form) the parallel regression assumption is violated, meaning that the coefficient is not the same across the 5 engagement levels and could be different between levels.

Location
Location has been found that have a non-significant effect on the entrepreneurial engagement level of an individual. This indicates that there is no significant relation between whether an individual lives in a rural or metropolitan area and whether this individual is on a higher or lower entrepreneurial engagement level. The fact that is it not significant seems contradictory to the significant positive effect of location on actual self-employment in phase 2.

Self-Employment of father and mother
The self-employment status of the father has been found to have a significant positive effect (.243) on the entrepreneurial engagement level, where-as surprisingly the self-employment status of the mother does not have a significant effect. It seems odd that the effect of the mother's self-employment status is not significant while the father's self-employment status is, but in phase 2 a hint at this was already shown since the mother's self-employment status was less significant than the father's in those results. The positive effect of the father's self-employment status indicates that if the father is self-employed the child will be more likely to be at a higher entrepreneurial engagement level.

Lack of financial support
Lack of financial support as a perceptual variable is found to be significant only at a 90% confidence level and to have a negative effect (-.113) on the entrepreneurial engagement level. This could indicate that financial constraints impedes an individual from moving to a higher engagement level. A logical intuition behind this would be that when actually taking steps and/or starting a young business the financial support is required, where-as when only thinking about it there is not yet an immediate need for financial support in that phase.

Perception of administrative complexity
When looking at the perception of administrative complexity, this has been found to have a significant negative effect (-.262) on the entrepreneurial engagement level. This negative effect could indicate that for a higher level of entrepreneurial engagement this is inherently linked with or seen as being linked with an increasing amount of administration of a complex nature.

Perceived lack of information
The perceived lack of information related to (entering) self-employment is found to have a non-significant negative effect (-.043). This result indicates that the lack of information as perceived by the individual has a negative effect on that individuals entrepreneurial engagement level. Someone who perceives there to be a lack of information is less likely to take steps to start a business or actually start a business as opposed to merely thinking about it or not thinking about it at all.

[18] The turning point of 0 years of age was calculated by the odds ratio of age divided by 2 x the odds ratio of Age^2 within the context of setting the first order derivative of preference for self-employment with regards to age to 0.

Risk tolerance

It is found that risk tolerance has a significant and positive effect (.257) on the entrepreneurial engagement level. This effect shows that the more risk tolerant an individual is, the higher the probability that he/she is at a higher engagement level. This seems similar to phase 2 where a higher risk tolerance was found to be linked to a higher probability of self-employment. These results fit the intuitive notion that self-employment has more inherent risks and as such an individual with higher risk tolerance is more likely to move along the engagement levels.

7. Instrumental variables method

The method used to correct for the endogeneity will be the instrumental variables (IV) method. After the logistic regressions of the preference for self-employment and the actual self-employment as well as the entrepreneurial engagement levels, the instrumental variable method will be utilized to correct the possible bias that exists in the estimates. Due to constraints in time and complexity the entrepreneurial engagement levels will not be investigated with the IV-method.

The instrumental variables (IV) technique is used to correct the estimated coefficients of a relationship between education and self-employment (in the form of preference for self-employment, actual self-employment and entrepreneurial engagement levels respectively).
As OLS (Ordinary least squares) regression is ill-equipped to handle endogeneity and this concept of endogeneity is found to be relevant for this research, the IV method is a proven way to handle the endogenous nature of education.

The way the IV method works will be explained further. The endogenous nature of the variable education means that it is in some way correlated to the error term in the original logistic regression equation. This endogeneity would bias the results from an OLS regression.
The IV method utilizes instrumental variables, additional variables, which are used to isolate the part of the original endogenous variable that is uncorrelated with the error term.

There are two conditions for a valid instrument according to Stock & Watson (2007) and Wooldridge (2006 Chapter 15) which are:

- Relevance.
  This means that the variation in the instrumental variable should be related to variation in the original endogenous variable. This is tested by examining the correlation between the instrumental variable and the endogenous variable which should be not equal to zero.
- Exogeneity.
  This means that the part of the variation of the original endogenous variable that is captured by the instrumental variable should be exogenous, thus the correlation between the instrumental variable and the error term of the original equation should be zero.

If the coefficients are exactly identified, an exact statistic test for the exogeneity of the instrumental variable is not possible according to Stock & Watson (2007; Chapter 12) and Verbeek (2004; Chapter 5), because of the exact identification. This is not the case in this research since there are two instrumental variables and one instrumented variable.

IV is an adequate method to deal with various problems of endogeneity and as such the method is valuable since the instruments ensure that the problems do not produce inaccurate results. However as an article from The Economist (2009; August 13) pointedly illustrates, while citing Leamer (1983) that the choice of variables can drastically alter the results from an analysis and even more so in the case of IV-analysis. Moreover caution is warranted due to the flaws inherent to an IV-approach as the scope is sufficiently narrowed that the goal of the research should not be missed.

To deal with at least one of the problems of the IV-method the Wald-test for exogeneity was conducted after each IV-model to test, with valid instruments, if endogeneity really exists in the relationship between the education and preference for self-employment or actual self-

employment. With a significant test this would indicate that the use of IV-models is warranted, but if the test shows as insignificant then it indicates that no endogeneity exists and it would be better to use a regular model as this would be likely to have smaller standard errors. Similarly the Amemiya-Lee-Newey minimum chi-square test was used as an over identification test to check the validity of the instruments where the H0 states that the instruments used are valid and as such can be used for the IV-method and the alternative hypothesis states the instruments are not valid and thus would conclude these weak instruments are better replaced or improved to avoid bias.

Ashenfelter and Rouse (1998) and Card (1998) as well as many other previous regarding returns to education which used instrumental variables used family background variables such as the education of the parents as an instrument. Van der Sluis et al. (2008) state that family backgrounds variables have been used as IV's as well as quarter of birth or changes in compulsory schooling laws. Since those latter two however are absent in the data-set used in this thesis, the focus for IV's is on family background. As Card (1999) notes the interest in these types of variables comes from the fact that children's choices with regards to education are usually highly correlated with the characteristics of their parents. Since there are limits in the data-set used in this thesis another similar variable in family background had to be used as instrument. The variable regarding social class of the parents was investigated and ultimately used..

Using an intuitive approach and since previous literature seem to support utilizing social class of the parents (using a separate variable for the father and the mother, making the assumption that there can be a distinctly different influence from the father and the mother), this is the type of instrumental variable used in the IV-models. The logic behind this would be that the social class influences the choice for education as supported by research such as Connor et al. (2001). According to Connor et al. (2001) a higher social class raises the probability of achieving a higher education. As such this logically implies a higher amount of years of education followed. Blackburn and Neumark (1993) investigated the endogeneity of schooling and experience and find that instrumenting for these proxies reduces the estimated return to human capital. They find that family background variables are suitable to be used to instrument, one of these being family status. According to Wolf (2004) when using international comparisons the wealth of people is positively correlated with the education of their children. Assuming this wealth can be measured by social class this would indicate that the social class of the parents would be positively correlated with the education of their children and as such provides another indication that this is indeed the right choice to use as an instrument.

As shown in the chapter 4 the variables used as instrumental variables are "White collar father" and "White collar mother" which are dummies composed from a variable for the father and a variable for the mother These dummies are part of the set of four dummies for the father and four for the mother respectively, constructed from the social class variable which indicates the occupation of the father and mother respectively. The other options such as the unemployed dummy, blue collar dummy and self-employed dummy were tested as instruments as well but were found to produce errors, unwanted numerical oddities and mostly did not pass the Amemiya-Lee-Newey minimum chi-square test as an over identification test or the Wald-test for exogeneity. The results of these tests for the chosen instrument variables are shown at the bottom of the tables of the respective IV-models.

7.1. Results and Analysis of IV Phase-1

IV-Phase 1 regards the possible effects of education and the control variables on the preference for self-employment whilst using the IV-method to control for endogeneity problems.

Table 14

Education with regards to Preference for self-employment in the strict definition (Probit regression & IV-Probit regression)

	Probit Model (Including age/100 sq.)			IV-Probit Model (Including age/100 sq.)					
				Preference for self-employment equation			Education equation		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Gender	.318 (.024)	0.000	13.47	.310 (.026)	0.000	12.12	-	-	-
Age	-.012 (.004)	0.005	-2.81	-.014 (.005)	0.003	-3.00	-	-	-
Age/100 sq.	.891 (.427)	0.037	2.09	1.065 (.460)	0.020	2.32	-	-	-
Education	-.001 (.002)	0.566	-0.57	.015 (.017)	0.363	0.91	-	-	-
Location	.025 (.016)	0.126	1.53	.042 (.024)	0.077	1.77	-	-	-
Lack of Fin. Sup.	-.011 (.030)	0.721	-0.36	-.003 (.031)	0.914	-0.11	-	-	-
Admin. Complex.	-.095 (.028)	0.001	-3.37	-.091 (.028)	0.001	-3.22	-	-	-
Lack of Info.	.059 (.025)	0.018	2.36	.062 (.025)	0.014	2.46	-	-	-
Risk Tolerance	.154 (.024)	0.000	6.32	.141 (.029)	0.000	4.92	-	-	-
Social class									
Father unemployed	.107 (.054)	0.049	1.97	.105 (.054)	0.055	1.92	-	-	-
Father self-employed	.142 (.030)	0.000	4.70	.138 (.030)	0.000	4.53	-	-	-
Mother unemployed	.015 (.028)	0.594	0.53	.024 (.029)	0.412	0.82	-	-	-
Mother self-employed	.102 (.044)	0.022	2.30	.111 (.045)	0.014	2.46	-	-	-
Father white-collar	-	-	-	-	-	-	1.191 (.142)	0.000	8.36
Mother white-collar	-	-	-	-	-	-	1.164 (.186)	0.000	6.27
/athrho	-	-	-	-	-	-	-.098 (.101)	0.329	-0.98
/lnsigma	-	-	-	-	-	-	1.791 (.006)	0.000	283.46
rho	-	-	-	-	-	-	-.098 (.100)	-	-
sigma	-	-	-	-	-	-	5.997 (.038)	-	-
Country Dummies	Included (Wald Chi2 = 411.79, p = 0.000)			Included (Wald Chi2 = 1345.45 p = 0.000)			-		
No of obs.	12522			12522					
Log likelihood	-8258.657			48455.607					
Prob > Chi2	0.000			0.000					
McFadden R2 (adj.)	0.047 (0.042)			-					
Nagelkerke R2	0.084			-					
Wald test of exogeneity	-			Chi2 = 0.59 p = 0.329					
Amemiya-Lee-Newey minimum chi-sq.	-			Chi2 = 1.444 p = 0.230					
Instruments[19]	-			Father white-collar, Mother white-collar					

[19] It should be noted that Stata's Ivprobit module utilizes all the independent variables as instruments in the education equation due to the nature of the ivprobit command. The instruments mentioned are the main intended instruments. For this reason two tests were performed. A Wald test for "Father white-collar" and "Mother white-collar" which shows Chi2 = 164.56 and p = 0.000 as well as a Wald test for all instruments as used by Stata's Ivprobit which shows Chi2 = 2583.92 and p = 0.000.

Table 14 presents the results of a probit estimation where the dependent variable is the preference for self-employment, compared to an IV-model with the same dependent variable.

The last column refers to the IV model, where-as the first column refers to the comparable non-IV model. For comparison purposes instead of using the earlier shown logistic model, a probit model was calculated using the same variables as the earlier logistic model. Also for comparison purposes the quadratic explanatory variable of education was omitted and social class variables were added.

Exogeneity and validity

The Wald test of exogeneity is shown in the table to have a Chi2 of 0.59 with a p-value of 0.329. This test has the H0 that the instrumented variable (education) is exogenous as opposed to endogenous. This means that if the test is significant, the IV-approach is necessary and the variable is suggested to be endogenous. The test however is non-significant and as such this suggest the variable education is not endogenous enough with respect to preference for self-employment and the IV-approach for this case is less appropriate. In this light the below analysis should be seen as more of an indication than a conclusive argument in favour of the IV-method. It is plausible that this test-result is related from the fact that in the original probit (and logit) regression education was found to be insignificant.

The Amemiya-Lee-Newey minimum Chi-sq. test as shown at the bottom of the table is found to have a Chi2 of 1.444 and to be non-significant (0.230). The H0 of this test is that the instruments used are valid. Since the test is found to be non significant, the H0 is not rejected and it suggests that the instruments “Father white-collar” and Mother white-collar” are valid instruments in the IV-probit-regression for the preference for self-employment..

Education

According to the IV-estimations education has no significant impact on the preference for being self-employed. This is similar to the case where IV is not used. It becomes a bit more significant but still well outside of any acceptable range of significance. Because of this an interpretation of the effect itself would not hold useful information. Nonetheless this will be mentioned, be it merely because the change that occurs due to the IV approach could bear merit in its own right.

As shown in the table the amount of education does not have a significant effect on whether or not an individual chooses to become self-employed as opposed to choosing to become an employee. This lack of significance is similar to the model where the instrumental variable method is not used. In both cases the preference does not have a significant effect which suggests that the lack of significance is not a result of endogeneity issues, but a more valid reason. A possible reason for this was discussed earlier and is still a valid theory: it could be that the preference to become self-employed is not a taught aspect and it can not be created by education if it's not already present. Similarly the other way around education would not be able to negate this preference. In this light it would seem education is irrelevant when looking at the preference in the strict sense regardless of whether endogeneity is accounted for or not.

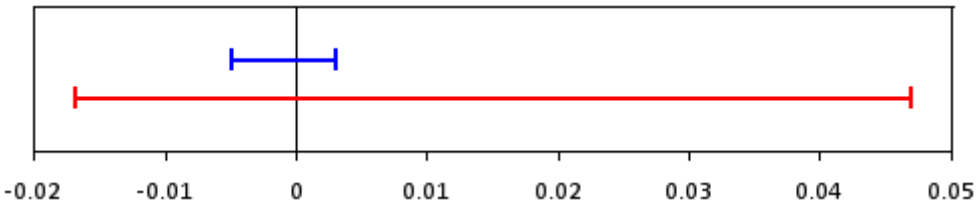
If the effect were significant one could note a change in the sign of the effect from negative (-.001) to positive (.015) as well as an increase in the strength of the effect by a factor 15. This would suggest endogeneity in the original model causes a downwards bias and with the usage of IV this can be at least partly corrected. It should be emphasized as stated before, the variable education in the IV-model is not significant and as such the effect itself does not lend itself to further useful interpretation. For illustrative purposes however, the confidence interval for the education variable in both models is shown below.

Preference for self-employment	Probit model		IV-probit model	
95% Confidence interval for education	-.005	.003	-.017	.047

To better display the difference for education between the probit and the IV-probit model the above table shows the 95% confidence intervals for both. The interval in the IV-probit model becomes wider by a factor eight when compared to the probit model from 0.008 to 0.064 and despite a slight shift (when comparing the centre of the interval) the IV-probit interval encompasses the probit interval. This suggest that the bias as found by the coefficients themselves is not necessarily true, since the intervals overlap and the true value could be in both intervals at the same time. This is illustrated by graph 1 below, where the blue line represents the interval of the probit regression and the red line the interval of the IV-probit regression.

Graph 1

95% Confidence interval of Education with regards to Preference for self-employment in the strict definition Probit regression (blue) & IV-Probit regression (red)



Control variables

The control variables are discussed below in a similar order as they are found in table 14, apart from the country variables which are only shortly mentioned as first item.

Country variables

Country dummies were also found to be significant although the specific coefficients per country were omitted in the results-table as they are not an important focus of this thesis.

Gender

Gender is found to remain significant in the IV model when compared to the original model. Its effect similarly remains positive and only has a slight decrease from .318 to .310 when IV is applied. This suggests that the ordinary probit-regression has an upwards bias, which is corrected downwards by the IV-regression. A reason could be self-selection by gender to be present through education influencing self-employment preference which is filtered out by the IV-method.

Age
Age as a linear variable is found to remain significant when the IV-method is used. The effect itself remains negative and becomes slightly stronger from -.012 to -.014, indicating a downwards bias from the probit regression with regards to age. Age as a quadratic variable remains significant and shows a strong increase in effect from .891 to 1.065. This suggest that there was a downwards bias in the probit regression for this effect as well. The combined effect is found to be positive in the relevant range, similar to the original regression in phase one.

Location
Location in the original probit regression is insignificant and it remains insignificant when IV is used although slightly less insignificant to the point where at a 90% confidence level it would be significant. The effect itself remains positive and becomes stronger from .025 to 0.42 indicating that the probit regression has a downwards bias, although this might be due to the insignificance of the variable in the probit regression.

Lack of financial support
It is found that lack of financial support which is insignificant in the original probit regression becomes more insignificant when IV is applied. The effect itself, despite not being useful to interpret due to the lack of significance, decreases from -.011 to -.003.

Perception of administrative complexity
Perception of administrative complexity is found to remain a significant negative effect when IV is applied, although the effect is slightly smaller, going from -.095 to -.091. This suggest that the probit regression has an upwards bias for this perception variable.

Perceived lack of information
Concerning the difficulty of finding information related to entering self-employment, the positive effect of this become slightly stronger from .059 to .062 while remaining significant when IV is applied. This suggests a downwards bias in the original probit regression.

Risk tolerance
Risk tolerance remains significant when IV is applied, the positive effect becomes less strong from .154 to .141 which suggests there is an upwards bias in the original probit regression.

Unemployment of father and mother
Utilizing the IV-method the unemployment status of the father becomes just insignificant, where-as the significance of the unemployment stats of the mother remains very insignificant. The effect of the father changes from a positive .107 to a slightly less strong positive .105 which suggests an upwards bias from the probit model. The effect of the mother increases in size from .015 to 0.24 indicating a downwards bias but due to the high insignificance this cannot be further analyzed.

Self-Employment of father and mother
When IV is used the self-employment status of the father as well as the self-employment status of the mother both remain significant, which supports the importance of social class variables. The effects remain positive for both, decreasing from .142 to .138 for the father and increasing from .102 to .111 for the mother. These changes are contradicting to each-other indicating that the effects of the parents are inherently different. It does show however that

self-employment of the parents as opposed to regular employment stimulates preference for self-employment.

The bias of probit vs IV-probit

In conclusion out of the 13 variables compared above, there are six variables indicating an upwards bias from the probit model when compared to the IV-probit model. It should be noted however that one of these is insignificant and as such the coefficient can not be interpreted meaningfully. The remaining seven variables indicate a downwards bias of the probit model when compared to the IV-probit model. It should be noted that two of these are insignificant (and a third one only significant at 90% confidence level), so that the coefficients of these might not be proper estimations.

The current results indicate that it varies per variable to such a degree that it is unclear whether as a whole there is a downwards or upwards bias. Further research might be useful to determine if this is the case or whether it is indeed a differing bias per (type of) variable.

7.2. Results and Analysis of IV Phase-2

IV-Phase 2 regards the possible effects of education and the control variables on actual self-employment whilst using the IV-method to control for endogeneity problems.

Table 15

Education with regards to Actual self-employment in the strict definition
(Probit regression & IV-Probit regression)
Including Preference for self-employment in the strict definition as control variable.

	Probit Model (Including age/100 sq.)			IV-Probit Model (Including age/100 sq.)					
				Actual self-employment equation			Education equation		
	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z	Coef. (Std.Err)	P>z	z
Preference	.912 (.038)	0.000	24.25	.714 (.082)	0.000	8.68	-	-	-
Gender	.259 (.036)	0.000	7.22	.211 (.038)	0.000	5.59	-	-	-
Age	.032 (.009)	0.000	3.56	.020 (.009)	0.024	2.26	-	-	-
Age/100 sq.	-1.679 (.973)	0.084	-1.73	-.767 (.903)	0.396	-0.85	-	-	-
Education	-.005 (.003)	0.093	-1.68	.103 (.019)	0.000	5.31	-	-	-
Location	.075 (.025)	0.003	3.02	.179 (.027)	0.000	6.71	-	-	-
Lack of Fin. Sup.	-.052 (.044)	0.235	-1.19	-.018 (.040)	0.658	-0.44	-	-	-
Admin. Complex.	-.100 (.042)	0.017	-2.39	-.057 (.039)	0.146	-1.45	-	-	-
Lack of Info.	.086 (.039)	0.027	2.21	.101 (.035)	0.004	2.89	-	-	-
Risk Tolerance	.074 (.038)	0.048	1.98	-.028 (.040)	0.482	-0.7	-	-	-
Social class									
Father unemployed	.009 (.089)	0.921	0.1	-.004 (.079)	0.957	-0.05	-	-	-
Father self-employed	.331 (.044)	0.000	7.5	.241 (.050)	0.000	4.82	-	-	-
Mother unemployed	.039 (.043)	0.364	0.91	.085 (.039)	0.030	2.18	-	-	-
Mother self-employed	.151 (.063)	0.016	2.4	.192 (.057)	0.001	3.39	-	-	-
Father white-collar	-	-	-	-	-	-	0.855 (.155)	0.000	5.53
Mother white-collar	-	-	-	-	-	-	1.024 (.205)	0.000	5.00
/athrho	-	-	-	-	-	-	-.699 (.167)	0.000	-4.19
/lnsigma	-	-	-	-	-	-	1.20 (.008)	0.000	211.13
rho	-	-	-	-	-	-	-.604 (.106)	-	-
sigma	-	-	-	-	-	-	5.584 (.045)	-	-
Country Dummies	Included (Wald Chi2 = 148.21 p = 0.000)			Included (Wald Chi2 = 633.65 p = 0.000)			-		
No of obs.	7536			7536					
Log likelihood	-3314.200			-26959.719					
Prob > Chi2	0.000			0.000					
McFadden R2 (adj.)	0.159 (0.149)			-					
Nagelkerke R2	0.237			-					
Wald test of exogeneity	-			Chi2 = 17.59 p = 0.000					
Amemiya-Lee-Newey minimum chi-sq.	-			Chi2 = 2.012 p = 0.156					
Instruments[20]	-			Father white-collar, Mother white-collar					

[20] It should be noted that Stata's Ivprobit module utilizes all the independent variables as instruments in the education equation due to the nature of the ivprobit command. The instruments mentioned are the main intended instruments. For this reason two tests were performed. A Wald test for "Father white-collar" and "Mother white-collar" which shows Chi2 = 75.45 and p = 0.000 as well as a Wald test for all instruments as used by Stata's Ivprobit which shows Chi2 = 1345.59 and p = 0.000.

Table 15 presents the results of a logit estimation where the dependent variable is the actual self-employment, compared to an IV-model with the same dependent variable.

The last column refers to the IV model, where-as the first column refers to the comparable non-IV model. For comparison purposes instead of using the earlier shown logistic model, a probit model was calculated using the same variables as the earlier logistic model. Also for comparison purposes the quadratic explanatory variable of education was omitted and social class variables were added.

Exogeneity and validity

The Wald test of exogeneity is shown in the table to have a Chi2 of 17.59 with a p of 0.000. This test has the H0 that the instrumented variable (education) is exogenous as opposed to endogenous. This means that if the test is significant, the IV-approach is necessary and the variable is suggested to be endogenous. The test is indeed significant and as such this suggest the variable education is endogenous enough with respect to actual self-employment and the IV-approach for this case is appropriate to correct for a bias in the original regression.

The Amemiya-Lee-Newey minimum Chi-sq. test as shown at the bottom of the table is found to have a Chi2 of 2.012 and to be non-significant (0.156). The H0 of this test is that the instruments used are valid. Since the test is found to be non significant, the H0 is not rejected and it suggests that the instruments "Father white-collar" and Mother white-collar" are valid instruments in the IV-probit-regression for actual self-employment.

Education

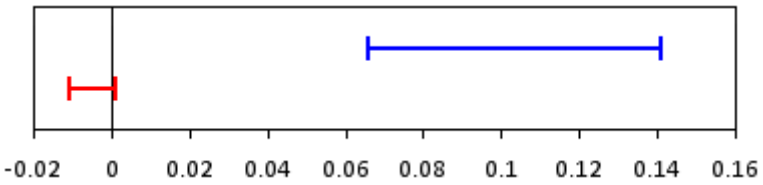
According to the IV-estimations the variable education becomes more significant, going from being significant at a 90% confidence level to being significant at a 95% level. This by itself could be a sign that education is more relevant than previously assumed by the estimations of the original probit or logit model. More notable is the impact on the probability of being self-employed when the IV-method is used. The effect changes from a small negative effect (-.005) to a stronger positive effect (.103). This would indicate that more education increases the probability of becoming self-employed as opposed to decreasing it, when endogeneity is taken into account. This sheds a different light on results previously found in this thesis as well as previous literature in the sense that the earlier negative effect found was merely due to endogeneity and endogeneity similarly could be the cause for the conflicting results in previous literature related to education and self-employment.
The positive effect of education which is now found could have several reasons. Intuitively it could mean the return to education for self-employment is bigger then the return to education for employment or at least perceived as bigger by the individual. This however assumes that the individual bases the choice of education on the return to education. This assumption comes from the intuitive logic that a person prefers higher earnings over lower earnings (if risks are more or less equal in their eyes or not relevant).
Another possibility is that at higher levels of education, self-employment indicates a highly educated individuals such as doctors and lawyers would start their own practice, not necessarily due to the higher return to education but for the purpose of status or because it's common practice in their field.

Actual self-employment	Probit model		IV-probit model	
95% Confidence interval for education	-.011	.001	.065	.141

To better display the difference for education between the probit and the IV-probit model the above table shows the 95% confidence intervals for both. The interval in the IV-probit model becomes wider by a little more than a factor six when compared to the probit model from 0.012 to 0.076 and has a clear shift towards a positive effect to such an extent that the IV-probit interval does not contain the probit interval. This suggest that the bias as found by the coefficients themselves is rather large, since the intervals do not have an overlap and the 'true value' as approximated by the models has to be different for both at a 95% confidence level. This is illustrated by graph 2 below, where the blue line represents the interval of the probit regression and the red line the interval of the IV-probit regression.

Graph 2

95% Confidence interval of Education with regards to Actual self-employment in the strict definition
Probit regression (blue) & IV-Probit regression (red)



Control variables

The control variables are discussed below in a similar order as they are found in table 15, apart from the country variables which are only shortly mentioned as first item.

Country variables
Country dummies were also found to be significant although the specific coefficients per country were omitted in the results-table as they are not an important focus of this thesis.)

Preference for self-employment
The preference for self-employment remains significant in explaining the actual self-employment of an individual when the IV-method is used. The effect becomes weaker from .912 to .714 and remains positive. This indicates that the more a person prefers self-employment, the higher the probability that that person will actually become self-employed even when endogeneity of education is taken into account. The change in effect suggests that the probit regression had an upwards bias.

Gender
Gender is found to have an equally significant positive effect on the probability of being self-employed when IV is applied. The effect does diminish from .259 to .211 and thus becomes less pronounced, indicating a downwards bias from the previous probit regression which could be caused by gender-selection through education.

Age
Age as a linear variable is found to become slightly less significant although remaining significant. Its effect does become slightly less strong from .032 to 0.20 indicating there might have been an upwards bias in the original probit regression.

The quadratic variable of age becomes much weaker in effect from -1.679 to -.767 and becomes insignificant from a previous significance at a 90% confidence level, indicating that the quadratic effect and resulting u-shape total effect are perhaps not the proper approximation when endogeneity of education is taken into account. The resulting effect shows an upward bias of the original probit regression.

Location
Location remains significant when IV is used and its effect increases strongly from .075 to .179, indicating that location has more of an impact than earlier estimated when endogeneity of education is taken into account. Similarly it indicates a downwards bias of the original probit regression.

Lack of financial support
It is found that lack of financial support becomes even more insignificant when IV is applied. The effect itself becomes weaker from -.052 to -.018 but is insignificant to such an extent that further interpretation would not yield much useful information. In both models the variable is insignificant, indicating that this is not altered by the endogeneity of education.

Perception of administrative complexity
Perception of administrative complexity is found to turn from a significant negative effect of -.100 to a slightly less strong but insignificant effect of -.057 when IV is applied. It would seem to indicate that the original probit regression has an upwards bias with regards to this variable, due to the insignificance however this could be misguided.

Perceived lack of information
Concerning the difficulty of finding information related to entering self-employment, the positive effect of this become slightly less pronounced and turns from positive to negative from .074 to -.028. However in the IV-model it loses its significance and becomes insignificant so the change in effect might not be an upwards bias from the probit model but instead a result of the insignificance.

Risk tolerance
Risk tolerance becomes highly insignificant when IV is applied, the effect itself turning from a positive effect of .074 to a less strong negative effect of -.004 which could be an upwards bias from the probit model, but due to the high insignificance an interpretation of this would not be prudent.

Unemployment of father and mother
Utilizing the IV-method the unemployment status of the father remains extremely insignificant, where-as the significance of the unemployment stats of the mother increases and the effect becomes significant. The effect of the father changes from a positive .009 to a negative -.004 which would be notable, except the high insignificance could be the reason for this change, as opposed to an upwards bias from the probit model. The effect of the mother increases in size from .039 to 0.85 indicating a downwards bias.

Self-Employment of father and mother
When IV is used the self-employment status of the father as well as the self-employment status of the mother remain significant. The self-employment status of the father decreases from .331 to .241 indicating an upwards bias of the probit regression. This bias however is not

seen in the self-employment status of the mother where there seems to be a downward bias as the effect rises from .151 to .192.

The bias of probit vs IV-probit

In conclusion out of the 14 variables compared above, there are nine variables indicating an upwards bias from the probit model when compared to the IV-probit model. It should be noted however that five of these are insignificant and as such the coefficient can not be interpreted meaningfully. The remaining five variables indicate a downwards bias of the probit model when compared to the IV-probit model. It should be noted that with one of these the coefficient of the probit model is insignificant, so that the comparison of this specific coefficient might not be useful since it might not be a proper estimation.
The current results indicate that it varies per variable to such a degree that it is unclear whether as a whole there is a downwards or upwards bias. Further research might be useful to determine if this is the case or whether it is indeed a differing bias per (type of) variable.

8. Conclusion

As a conclusion the results for education from the previous regression analyses are again briefly reviewed for the three phases that did not include IV-methodology to deal with endogeneity. For the first two phases these results are directly contrasted with the results from the IV-models to emphasize the difference endogeneity makes. Following this the issues of endogeneity are shortly reviewed and how the IV-methods is a useful tool to correct for these issues, albeit a tool whose potential problems have to be taken into account. Concluding suggestions are made for further research whilst emphasizing the importance of taking endogeneity into account. This being noted the following aspects are discussed:

- The effect of education on the preference for self-employment with and without endogeneity being taken into account
- The effect of education on the propensity for self-employment with and without endogeneity being taken into account
- The effect of education on the different levels of entrepreneurial engagement without endogeneity being taken into account
- Endogeneity, the IV-method and suggestions for further research

The effect of education on the preference for self-employment

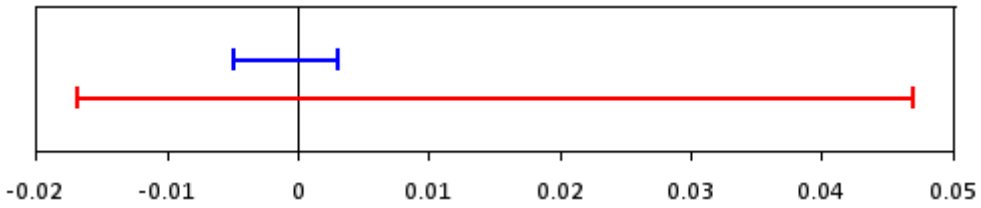
According to the estimates, the effects of education on the preference for self-employment when endogeneity is not taken into account are not significant. The amount of education apparently has no significant effect on an individuals preference for self-employment and as such education apparently does not play a role in the desire for entrepreneurship. When the IV-method is used and endogeneity is accounted for the effect remains insignificant. Because of this an interpretation of the effect itself would not hold useful information. The negative effect in the relative range therefore can not be seen as a valid negative effect.
The logic behind the insignificance seems intuitively to be that preference for self-employment is not a taught aspect but possible more-so an aspect of personality on which education does not have a noticeable effect. This preference stems from other factors, including the significant control variables in the results such as gender, age, the self-employment status of father and mother as well as perceptual variables.
In this light it would seem education is irrelevant when looking at the preference in the context of using policy to raise the preference for entrepreneurship (which by itself has a significant positive effect on actual entrepreneurship).

The confidence interval for education is shown in a table as well as a graph below. It shows that the interval when endogeneity is taken into account becomes wider and shifts slightly to the right, despite encompassing the original interval. This suggest that for the preference for self-employment, a bias of the original model is not necessarily true. Taking the results of the tests into account, endogeneity of education seems to not be an issue for the preference for self-employment.

Preference for self-employment	Probit model		IV-probit model	
95% Confidence interval for education	-.005	.003	-.017	.047

Graph 3

95% Confidence interval of Education with regards to Preference for self-employment in the strict definition
Probit regression (blue) & IV-Probit regression (red)



The effect of education on the propensity for self-employment

According to the estimates, the effects of education on the actual propensity for self-employment when endogeneity is not taken into account is significant. The amount of education has a significant quadratic relationship with the probability of an individual becoming self-employed. This effect is positive in the relevant range. When the IV-method is used and endogeneity taken into account, the effect becomes more significant which underlines its importance but more notably: the size and sign of the effect changes.

The formerly small negative (-.005) turns out to be a strong positive effect (.103) when endogeneity is accounted for. This shows that endogeneity creates a significant downwards bias in the estimates when a model is used that does not take endogeneity into account. This could explain the lack of consensus on the effect of education on the probability for self-employment. Considering that most previous literature does not correct for the endogeneity, the estimates could be biased which could cause the positive effect to be misinterpreted as being negative or non significant.

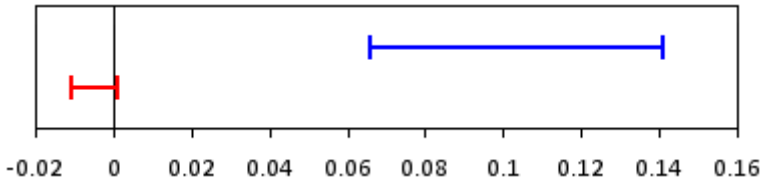
The positive effect as established could be due to various reasons such as a potentially higher return to education for self-employment and therefore this option would become increasingly more appealing with higher education. Another possibility is that at higher levels of education, self-employment indicates a highly educated individuals start their own practice due to status-purposes or common practice.

The confidence interval for education is shown in a table as well as a graph below. It shows that the interval when endogeneity is taken into account becomes wider and shifts to a very large amount to the right, eliminating any overlap with the original interval. As stated before this suggests that for the actual choice for self-employment, there is a significant downwards bias of the original model. Taking the results of the tests into account, endogeneity of education seems to be a severe issue which needs to be taken seriously.

Actual self-employment	Probit model		IV-probit model	
95% Confidence interval for education	-.011	.001	.065	.141

Graph 4

95% Confidence interval of Education with regards to Actual self-employment in the strict definition
Probit regression (blue) & IV-Probit regression (red)



The effect of education on the different levels of entrepreneurial engagement

The estimates of the ordinal regression show that education has a linear aspect that is significant and positive and a quadratic aspect that is not significant and negative. Due to the non-significance of the quadratic effect, the overall effect is harder to determine and likely to be positive in the relevant range. This indicates that the higher the education of the individual, the higher the engagement level on which the individual will be. Intuitively the reasoning can be used that there are certain barriers to moving to a higher engagement level for which a higher education are required. The positive effect is similar to the effect of education on the propensity for self-employment and in that respect consistent.

Endogeneity, the IV-method and suggestions for further research

As noted above, endogeneity is significantly present in education, at the very least with regards to actual self-employment choice. The endogeneity perceived could be a likely cause for the contrasting results for the significance, direction and strength of the relationship of education with self-employment in previous literature.

Apart from this the lack of uniformity in approaches and definitions makes comparison of studies difficult and the mere choice of variable or definitions could cause a bias in the results. This by itself could be remedied by choosing stricter definitions and coming to a standard approach and terminology for the different areas of research to avoid that an activity such as choosing variables has an overbearing effect on the final results.

Nonetheless, the uniformity that most studies do not correct for endogeneity suggest that this is the first problem to tackle. In that sense it would be prudent to look at the various sources for endogeneity and utilize the IV-method or an improved method to eliminate bias as much as possible.

Various sources of this endogeneity could be possible such as:

- Omitted variables such as ability
- Misreporting of educational attainment
- Self-selection with regards to education and self-employment
- Simultaneous causality between education and self-employment
- Option value of education if viewed in a dynamic way where further education creates more valuable options

These sources can cumulate to form a strong endogeneity of education causing a bias in research estimates and understating the effects of education on self-employment-choice
To minimize bias in further research it seems prudent to suggest the usage of the IV-method or at least in some way take account of the endogeneity when education is present as a variable. This suggestion is a direct result from the results found that when endogeneity is taken into account the estimates for education's effect change dramatically. Looking at previous literature Grilo and Thurik (2008) already noted that endogeneity should be warded for as education does suffer from this risk.

The IV-method can be used to correct for the endogeneity as long as the potential pitfalls are avoided and instruments are carefully chosen to ensure:

- Relevance: the variation in the instrument should relate to the variation in education.
- Exogeneity: part of the variation of education that is captured by the instrument should be exogenous so the instrument doesn't suffer the same problem as education.

Similarly the Wald-test for exogeneity and the Amemiya-Lee-Newey minimum chi-square test should be performed at least to test for the validity of the instruments and to confirm the appropriateness of the IV-method. Since using the IV-method when not appropriate or with weak instruments can create a bias by itself and the 'cure' in that case would only deteriorate the problem instead of alleviating it. In other words, the emphasis should remain on *"thinking about how and why things work"*[21] as opposed to becoming the subject of the description: *"Like elaborately plumed birds…we preen and strut and display our t-values."*[22]

[21] Quoted from Angus Deaton as taken from the article "Cause and Defect" from The Economist (2009)
[22] Quoted from Edward Leamer, as taken from Leamer (1983) from the article "Cause and Defect" from The Economist (2009)

9. References

- Acs, Z., (2006), "How is entrepreneurship good for economic growth?", Innovations: Technology, Governance, Globalization. MIT Press

- Ashenfelter, O. and Rouse, C., (1998), "Income, Schooling and Ability: Evidence from a New Sample of Identical Twins" , Quarterly Journal of Economics 113, no.1, 253-284

- Bascle, G., (2008), "Controlling for endogeneity with instrumental variables in strategic management research", Strategic Organization Vol 6(3): 285–327

- Bates, T. (1990), "Entrepreneur Human-Capital Inputs and Small Business Longevity." The Review of Economics and Statistics, 72, (4), 551-559

- Becker, G. S., (1962) "Investment in Human Capital: A Theoretical Analysis," Journal of Political Economy LXX: 9 - 49

- Becker, G.S., (1975), "Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education", 2nd ed. NBER

- Becker, G.S. and Chiswick, B.R., (1966), "Education and the Distribution of Earnings", The American Economic Review, Vol. 56, No. 1/2, pp. 358-369

- Blackburn, M.L. and Neumark, D., (1993), "Are OLS estimates of the return to schooling biased downward? Another Look.", NBER

- Blanchflower, D. G., (2004), "Self-Employment: More May Not Be Better", NBER Working Paper No. 10286

- Blanchflower et al. (2001). "Latent entrepreneurship across nations.", European Economic Review, 45, 680-691

- Brown et al. (2008) "Self-Employment and Risk Preference", Working paper, Department of Economics, University of Sheffield

- Carree, M.A. and Thurik, A.R., (2003), "The impact of entrepreneurship on economic growth", in "Handbook of Entrepreneurship Research", Audretsch, D.B. and Acs, Z.J. (eds), Kluwer Academic Publishers, Boston/Dordrecht

- Carree, M.A. and Thurik, A.R., (2006), "Understanding the role of entrepreneurship for economic growth", in "The Handbook Entrepreneurship and Economic Growth", Carree, M.A. and Thurik, R., (eds), Edward Elgar Publishing Limited, Cheltenham, UK

- Card, David, (1998), "The causal effect of education." Working paper, University of California, Berkeley

- Connor, H. et al., (2001), "Social class and higher education issues affecting decisions on participation by lower social class groups", Institute for employment studies, HMSO

- Davidsson, P., (2006), "Nascent entrepreneurship: empirical studies and developments", Foundations and Trends in Entrepreneurship Research, 2(1), p. 1–76

- European Commission Flash Barometer 192 Report (2007), The Gallup Organisation, url: http://ec.europa.eu/public_opinion/flash/fl_192_en.pdf

- Eurydice, (2009), "Compulsory age of starting school in European countries, 2009", NFER, url: http://www.nfer.ac.uk/eurydice/briefingseurope/school-starting-ages.cfm

- Evans, D.S. and Leighton, L.S., (1989), "Some empirical aspects of entrepreneurship", American Economic Review 79(3), 519-535

- Evans, D.S. and Leighton, L.S., (1990), "Small Business Formation by Unemployed and Employed Workers," Small Business Economics, 2(4), 319-330

- Grilo, I. and Irigoyen J.M., (2006), "Entrepreneurship in the EU: to wish and not to be", Small Business Economics 26 (4), 305-318

- Grilo, I. and Thurik, A.R., (2006), "Entrepreneurship in the old and new Europe", in: "Entrepreneurship, Growth, and Innovation: the Dynamics of Firms and Industries, International Studies in Entrepreneurship", Santarelli, E. (ed.), Berlin: Springer Science

- Grilo, I. and A.R. Thurik, (2008), "Determinants of entrepreneurial engagement levels in Europe and the US", Industrial and Corporate Change, forthcoming

- Heckman et al (2005), "Earnings functions, rates of return and treatment effects: The Mincer equation and beyond", IZA Discussion Paper No. 1700

- Katz, J.A. (1992), "A Psychosocial Cognitive Model of Employment Status Choice", Entrepreneurship: Theory and Practice, Vol. 17

- Knight, F.H., (1921), "Risk, Uncertainty and Profit", Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Co

- Kolstad, I and Wiig, A., (2009), "Education and entrepreneurship" Chr. Michelsen Institute, CEIC-CMI annual conference, Luanda, Angola

- Kolvereid, L., (1996), "Organizational Employment versus Self-Employment: Reasons for Career Choice Intentions.", Entrepreneurship: Theory and Practice, Vol. 20

- Le, A.T., (1999), "Empirical studies of self-employment", Journal of Economic Surveys 13(4), 381-416

- Leamer, E., (1983), "Let's take the con out of econometrics", American Economic Review 73(1)

- Lucas, R.E. Jr. (1978), "On the size distribution of business firms", Bell Journal of Economics, 9, 508-523

- Mincer, J.A., (1970), "The Distribution of Labor Incomes: A Survey With Special Reference to the Human Capital Approach", Journal of Economic Literature, Vol. 8, No. 1, p. 1-26

- Mincer, J.A., (1974), "Schooling, Experience, and Earnings", NBER Books, National Bureau of Economic Research

- Murray, M. P., (2006), "Econometrics : a modern introduction", Pearson Addison-Wesley

- Parker, S.C., (2004), “The Economics of Self-employment and Entrepreneurship”, Cambridge: Cambridge University Press

- Parker, S. and Van Praag, M., (2006), “Schooling, capital constraints and entrepreneurial performance: The endogenous triangle”, Journal of Business and Economic Statistics 24(4), 416-431

- Robinson, P. B. and Sexton, E.A., (1994), “The effect of education and experience on self-employment success”, Journal of Business Venturing, Elsevier, vol. 9(2), p. 141-156

- Schumpeter, J., (1934), “The Theory of Economic Development”, Harvard University Press, Boston

- Stock, J.H., Watson, M.W., (2007), “Introduction to econometrics”, 2nd edition, Pearson Education

- The Economist, (2009), “Cause and defect”, The Economist Aug 13, The Economist Newspaper Limited London

- Thurik et al., (2007), "Modelling latent and actual entrepreneurship", Scales Research Reports H200719, EIM Business and Policy Research

- Van der Sluis et al., (2004), “Education and Entrepreneurship in Industrialized Countries: A Meta-Analysis” Tinbergen Institute Discussion Paper, 2003- 046/3. Revised Sept. 2004

- Van der Sluis et al., (2008), “Education and entrepreneurship selection and performance: a review of the empirical literature”, Journal of Economic Surveys 22:5, 795-841

- Verbeek, M., (2004), “A guide to modern econometrics”, 2nd edition, John Wiley & Sons

- Wennekers, S. and Thurik, R., (1999), “Linking Entrepreneurship and Economic Growth”, Small Business Economics 13: 27–55

- Wolf, A., (2004), “Education and economic performance: simplistic theories and their policy consequences” Oxford review of economic policy, Vol 20, No 2

- Wooldridge, J.M., (2006), “Introductory econometrics : a modern approach”, 3rd edition, Thomson South-Western

- Zwan et al., (2008), “The entrepreneurial ladder and its determinants”, Applied Economics, forthcoming

10. Appendices

Appendix A Tables supporting Regression Phase 1

Strict Definition

Table A1
Goodness of fit tests
Preference for Self-employment (Strict Definition)

	Model 1 (Basic Model)	Model 2 (Including age/100 sq.)	Model 2 (Full Model)
No of obs.	12522	12522	12522
No of covariate patterns	12366	12366	12366
Pearson chi2(d.o.f.)	12388.57 (12382)	12388.70 (12327)	12387.37 (12326)
Pearson Prob > chi2	0.349	0.346	0.347
Hosmer-Lemeshow Chi2 (d.o.f.)	5.69 (8)	6.69 (8)	6.35 (8)
Hosmer-Lemeshow Prob > Chi2	0.682	0.570	0.608

Table A2
Classification Table
Preference for Self-employment (Strict Definition)

	Model 1 (Basic Model)	Model 2 (Including age/100 sq.)	Model 2 (Q1Wide Full Model)
Sensitivity Pr(+ D)	53.17%	53.29%	53.18%
Specificity Pr(-~D)	67.01%	66.92%	66.78%
Positive predictive value Pr(D +)	59.34%	59.33%	59.18%
Negative predictive value Pr(~D -)	61.24%	61.27%	61.17%
False + rate for true ~D Pr(+~D)	32.99%	33.08%	33.22%
False - rate for true D Pr(- D)	46.83%	46.71%	46.82%
False + rate for class. + Pr(~D +)	40.66%	40.67%	40.82%
False - rate for class. - Pr(D -)	38.76%	38.73%	38.83%
Correctly classified	60.43%	60.44%	60.32%

Classified + if predicted Pr(D) >= .5

Table A3

Multicollinearity diagnostics (VIF) Table
Preference for Self-employment (Strict Definition)

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 2 (Full Model)	
Variable	VIF	Tolerance	VIF	Tolerance	VIF	Tolerance
Preference	1.07	0.938	1.07	0.9372	1.07	0.937
Gender	1.05	0.955	1.05	0.9532	1.05	0.952
Age	1.07	0.938	33.66	0.0297	33.78	0.030
Age/100 sq.	-	-	33.77	0.0296	34.11	0.029
Education	1.13	0.889	1.13	0.8867	10.36	0.097
Educ. /100 sq.	-	-	-	-	9.96	0.100
Location	1.09	0.919	1.09	0.9182	1.09	0.914
S.E. Father	1.32	0.758	1.32	0.7569	1.32	0.757
S.E. Mother	1.25	0.803	1.25	0.8025	1.25	0.802
Lack of Fin. Sup.	1.17	0.854	1.17	0.8529	1.17	0.852
Admin. Complex.	1.16	0.860	1.16	0.8597	1.16	0.859
Lack of Info.	1.2	0.831	1.21	0.8291	1.21	0.827
Risk Tolerance	1.15	0.870	1.15	0.8697	1.15	0.866
Country Dummies	Included		Included		Included	
Mean VIF	1.46		3.13		3.54	

Table A4

Specification Diagnostics
Preference for Self-employment (Strict Definition)

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 2 (Full Model)	
	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z
_hat	.999 (.038)	0.000	1.003 (.038)	0.000	1.004 (.038)	0.000
_hatsq	-.007 (.058)	0.911	.019 (.058)	0.742	.025 (.058)	0.662
_cons	.002 (.024)	0.945	-.005 (.024)	0.839	-.006 (.024)	0.789
No of obs.	12522		12522		12522	
Log likelihood	-8258.248		-8257.053		-8256.150	
LR Chi2 (d.o.f.)	811.97 (2)		814.36 (2)		816.17 (2)	
Prob > Chi2	0.000		0.000		0.000	
Pseudo R2	0.0469		0.0470		0.0471	

Strict Definition & Wide Definition

The following tables show the full model for both the strict definition and the wide definition of phase one for comparison purposes.

Table A5

Goodness of fit tests
Preference for Self-employment (Strict Definition & Wide Definition)

	Model 1 (Strict Full Model)	Model 2 (Wide Full Model)
No of obs.	12522	12878
No of covariate patterns	12366	12713
Pearson chi2(d.o.f.)	12387.37 (12326)	12736.47 (12673)
Pearson Prob > chi2	0.347	0.344
Hosmer-Lemeshow Chi2 (d.o.f.)	6.35	12.72
Hosmer-Lemeshow Prob > Chi2	0.608	0.122

Table A6

Classification Table
Education with regards to Preference for Self-employment (Strict Definition & Wide Definition)

	Model 1 (Strict Full Model)	Model 2 (Wide Full Model)
Sensitivity Pr(+ D)	53.18%	49.25%
Specificity Pr(-~D)	66.78%	70.36%
Positive predictive value Pr(D +)	59.18%	58.81%
Negative predictive value Pr(~D -)	61.17%	61.74%
False + rate for true ~D Pr(+~D)	33.22%	29.64%
False - rate for true D Pr(- D)	46.82%	50.75%
False + rate for class. + Pr(~D +)	40.82%	41.19%
False - rate for class. - Pr(D -)	38.83%	38.26%
Correctly classified	60.32%	60.61%

Classified + if predicted Pr(D) >= .5

Table A7

Multicollinearity diagnostics (VIF) Table
Preference for Self-employment (Strict Definition & Wide Definition)

	Model 1 (Strict Full Model)		Model 2 (Wide Full Model)	
Variable	VIF	Tolerance	VIF	Tolerance
Preference	1.07	0.937	1.07	0.938
Gender	1.05	0.952	1.05	0.953
Age	33.78	0.030	33.97	0.029
Age/100 sq.	34.11	0.029	34.3	0.029
Education	10.36	0.097	10.3	0.097
Educ. /100 sq.	9.96	0.100	9.89	0.101
Location	1.09	0.914	1.09	0.915
S.E. Father	1.32	0.757	1.32	0.757
S.E. Mother	1.25	0.802	1.25	0.802
Lack of Fin. Sup.	1.17	0.852	1.17	0.852
Admin. Complex.	1.16	0.859	1.16	0.859
Lack of Info.	1.21	0.827	1.21	0.826
Risk Tolerance	1.15	0.866	1.16	0.865
Country Dummies	Included		Included	
Mean VIF	3.54		3.56	

Table A8

Specification Diagnostics
Preference for Self-employment (Strict Definition & Wide Definition)

	Model 1 (Strict Full Model)		Model 2 (Wide Full Model)	
	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z
_hat	1.004 (.038)	0.000	1.00 (.0391)	0.000
_hatsq	.025 (.058)	0.662	.018 (.057)	0.752
_cons	-.006 (.024)	0.789	-.004 (.023)	0.854
No of obs.	12522		12878	
Log likelihood	-8256.150		-847834	
LR Chi2 (d.o.f.)	816.17 (2)		826.99 (2)	
Prob > Chi2	0.000		0.000	
Pseudo R2	0.0471		0.0465	

Appendix B Tables supporting Regression Phase 2

Strict Definition

Table B1

Goodness of fit tests
Education with regards to Actual Self-employment (Strict Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)	Model 2 (Including age/100 sq.)	Model 3 (Full Model)
No of obs.	7536	7536	7536
No of covariate patterns	7495	7495	7495
Pearson chi2(d.o.f.)	7333.6 (7456)	7369.10 (7455)	7372.74 (7454
Pearson Prob > chi2	0.815	0.758	0.746
Hosmer-Lemeshow Chi2 (d.o.f.)	28.02 (8)	32.02 (8)	28.22 (8)
Hosmer-Lemeshow Prob > Chi2	0.001	0.000	0.000

Table B2

Classification Table
Education with regards to Actual Self-employment (Strict Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)	Model 2 (Including age/100 sq.)	Model 3 (Full Model)
Sensitivity Pr(+ D)	21.33%	21.52%	20.97%
Specificity Pr(-~D)	95.42%	95.46%	95.56%
Positive predictive value Pr(D +)	56.38%	56.77%	56.69%
Negative predictive value Pr(~D -)	81.39%	81.43%	81.34%
False + rate for true ~D Pr(+~D)	4.58%	4.54%	4.44%
False - rate for true D Pr(- D)	78.67%	78.48%	79.03%
False + rate for class. + Pr(~D +)	43.62%	43.23%	43.31%
False - rate for class. - Pr(D -)	18.61%	18.57%	18.66%
Correctly classified	79.34%	79.41%	79.37%

Classified + if predicted Pr(D) >= .5

Table B3

Multicollinearity diagnostics (VIF) Table
Education with regards to Actual Self-employment (Strict Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 3 (Full Model)	
Variable	VIF	Tolerance	VIF	Tolerance	VIF	Tolerance
Actual S.E.	1.19	0.844	1.19	0.844	1.19	0.843
Preference	1.17	0.858	1.17	0.856	1.17	0.856
Gender	1.06	0.946	1.06	0.945	1.06	0.945
Age	1.07	0.939	35.31	0.028	35.64	0.028
Age/100 sq.	-	-	35.33	0.028	35.95	0.028
Education	1.12	0.896	1.12	0.896	12.08	0.083
Educ. /100 sq.	-	-	-	-	11.89	0.084
Location	1.11	0.905	1.11	0.904	1.11	0.900
S.E. Father	1.32	0.758	1.32	0.757	1.32	0.757
S.E. Mother	1.23	0.811	1.23	0.811	1.23	0.810
Lack of Fin. Sup.	1.17	0.852	1.17	0.852	1.18	0.850
Admin. Complex.	1.17	0.853	1.17	0.852	1.17	0.852
Lack of Info.	1.23	0.815	1.23	0.814	1.23	0.813
Risk Tolerance	1.13	0.886	1.13	0.886	1.13	0.882
Country Dummies	Included		Included		Included	
Mean VIF	1.46		3.16		3.66	

Table B4

Specification Diagnostics
Education with regards to Actual Self-employment (Strict Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 3 (Full Model)	
	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z
_hat	.810 (.071)	0.000	.813 (.071)	0.000	.872 (.040)	0.000
_hatsq	-.082 (.028)	0.003	-.081 (.028)	0.004	-.056 (.013)	0.000
_cons	-.035 (.046)	0.451	-.034 (.046)	0.459	-.022 (.044)	0.619
No of obs.	7536		7536		7536	
Log likelihood	-3307.998		-3307.580		-3,305.702	
LR Chi2 (d.o.f.)	1269.70 (2)		1270.54 (2)		1274.29	
Prob > Chi2	0.000		0.000		0.000	
Pseudo R2	0.161		0.161		0.162	

Strict Definition & Wide Definition

The following tables show the full model for both the strict definition and the wide definition of phase two for comparison purposes.

Table B5

Goodness of fit tests
Education with regards to Actual Self-employment (Strict Definition & Wide Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Strict Full Model)	Model 2 (Wide Full Model)
No of obs.	7536	12382
No of covariate patterns	7495	12288
Pearson chi2(d.o.f.)	7372.74 (7454)	13084.89 (12247)
Pearson Prob > chi2	0.746	0.000
Hosmer-Lemeshow Chi2 (d.o.f.)	28.22 (8)	21.20
Hosmer-Lemeshow Prob > Chi2	0.004	0.007

Table B6

Classification Table
Education with regards to Actual Self-employment (Strict Definition & Wide Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Strict Full Model)	Model 2 (Wide Full Model)
Sensitivity Pr(+ D)	20.97%	7.58%
Specificity Pr(-~D)	95.56%	99.24%
Positive predictive value Pr(D +)	56.69%	60.19%
Negative predictive value Pr(~D -)	81.34%	87.58%
False + rate for true ~D Pr(+~D)	4.44%	0.76%
False - rate for true D Pr(- D)	79.03%	92.42%
False + rate for class. + Pr(~D +)	43.31%	39.81%
False - rate for class. - Pr(D -)	18.66%	12.42%
Correctly classified	79.37%	87.13%

Classified + if predicted Pr(D) >= .5

Table B7

Multicollinearity diagnostics (VIF) Table
Education with regards to Actual Self-employment (Strict Definition & Wide Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Strict Full Model)		Model 2 (Wide Full Model)	
Variable	VIF	Tolerance	VIF	Tolerance
Actual Occupation	1.19	0.843	1.13	0.887
Preference	1.17	0.856	1.12	0.896
Gender	1.06	0.945	1.07	0.939
Age	35.64	0.028	35.36	0.028
Age/100 sq.	35.95	0.028	35.75	0.028
Education	12.08	0.083	10.53	0.095
Educ. /100 sq.	11.89	0.084	10.12	0.099
Location	1.11	0.900	1.10	0.912
S.E. Father	1.32	0.757	1.33	0.752
S.E. Mother	1.23	0.810	1.25	0.801
Lack of Fin. Sup.	1.18	0.850	1.18	0.851
Admin. Complex.	1.17	0.852	1.17	0.858
Lack of Info.	1.23	0.813	1.21	0.825
Risk Tolerance	1.13	0.882	1.16	0.865
Country Dummies	Included		Included	
Mean VIF	3.66		3.57	

Table B8

Specification Diagnostics
Education with regards to Actual Self-employment (Strict Definition & Wide Definition)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (D4Strict Full Model)		Model 2 (D4Wide Full Model)	
	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z
_hat	.872 (.040)	0.000	.794 (.071)	0.000
_hatsq	-.056 (.013)	0.000	-.089 (.028)	0.001
_cons	-.022 (.044)	0.619	-.037 (.046)	0.415
No of obs.	7536		7536	
Log likelihood	-3305.702		-3,308.788	
LR Chi2 (d.o.f.)	1274.29 (2)		1268.12 (2)	
Prob > Chi2	0.000		0.000	
Pseudo R2	0.1616		0.161	

Appendix C Tables supporting Regression Phase 3

Table C1

Multicollinearity diagnostics (VIF) Table
Education with regards to Entrepreneurial Engagement Levels (Ordinal Regression)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 3 (Full Model)	
Variable	VIF	Tolerance	VIF	Tolerance	VIF	Tolerance
Engagement Level	1.29	0.778	1.31	0.761	1.31	0.761
Preference	1.23	0.810	1.24	0.808	1.24	0.808
Gender	1.08	0.924	1.09	0.920	1.09	0.920
Age	1.09	0.916	32.94	0.030	33	0.030
Age/100 sq.	-	-	33.24	0.030	33.45	0.030
Education	1.14	0.873	1.15	0.871	11.75	0.085
Educ. /100 sq.	-	-	-	-	11.13	0.090
Location	1.08	0.923	1.08	0.923	1.09	0.918
S.E. Father	1.32	0.756	1.33	0.754	1.33	0.754
S.E. Mother	1.25	0.799	1.25	0.799	1.25	0.799
Lack of Fin. Sup.	1.19	0.841	1.19	0.840	1.19	0.840
Admin. Complex.	1.18	0.845	1.18	0.844	1.18	0.844
Lack of Info.	1.21	0.829	1.21	0.828	1.21	0.826
Risk Tolerance	1.16	0.862	1.16	0.862	1.16	0.859
Country Dummies	Included		Included		Included	
Mean VIF	1.48		3.07		3.53	

Table C2

Specification Diagnostics
Education with regards to Entrepreneurial Engagement Levels (Ordinal Regression)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 3 (Full Model)	
	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z	Coef. (Std.Err)	P>z
_hat	1.076 (.034)	0.000	1.054 (.031)	0.000	1.275 (.110)	0.000
_hatsq	-.052 (.017)	0.002	-.040 (.016)	0.013	-.041 (.016)	0.010
/cut1	.964 (.032)		.900 (.032)		3.986 (.179)	
/cut2	1.924 (.038)		1.864 (.038)		4.950 (.181)	
/cut3	2.402 (.042)		2.4348 (.042)		5.434 (.182)	
/cut4	2.930 (.047)		2.881 (.047)		5.967 (.183)	
No of obs.	8643		8643		8643	
Log likelihood	-8652.401		-8629.443		-8629.439	
LR Chi2 (d.o.f.)	2802.36 (2)		2848.27 (2)		2848.28 (2)	
Prob > Chi2	0.000		0.000		0.000	
Pseudo R2	0.1394		0.1417		0.1417	

Table C3

Brant Parallel Regression Assumption Test
Education with regards to Entrepreneurial Engagement Levels (Ordinal Regression)
Using Preference for self-employment in the strict definition as control variable

	Model 1 (Basic Model)		Model 2 (Including age/100 sq.)		Model 3 (Full Model)	
Variable	Chi2 (d.o.f.)	p > Chi2	Chi2 (d.o.f.)	p > Chi2	Chi2 (d.o.f.)	p > Chi2
All	1792.36 (114)	0.000	1534.96 (117)	0.000	1515.31 (120)	0.000
Preference	49.49 (3)	0.000	48.3 (3)	0.000	48.41 (3)	0.000
Gender	30.21 (3)	0.000	35.47 (3)	0.000	34.64 (3)	0.000
Age	688.58 (3)	0.000	135.93 (3)	0.000	134.55 (3)	0.000
Age/100 sq.	-	-	73.45 (3)	0.000	71.38 (3)	0.000
Education	29.79 (3)	0.001	33.24 (3)	0.000	1.93 (3)	0.588
Educ. /100 sq.	-	-	-	-	3.53 (3)	0.317
Location	4.32 (3)	0.229	4.04 (3)	0.257	4.13 (3)	0.248
S.E. Father	3.14 (3)	0.370	5.41 (3)	0.144	5.4 (3)	0.145
S.E. Mother	6.53 (3)	0.088	6.41 (3)	0.093	6.5 (3)	0.090
Lack of Fin. Sup.	7.12 (3)	0.068	6.98 (3)	0.072	6.73 (3)	0.081
Admin. Complex.	3.95 (3)	0.266	3.98 (3)	0.264	3.92 (3)	0.270
Lack of Info.	1.23 (3)	0.745	0.79 (3)	0.851	0.83 (3)	0.843
Risk Tolerance	5.81 (3)	0.121	6.13 (3)	0.105	6.6 (3)	0.086
Country Dummies	Included		Included		Included	

Appendix D Tables supporting IV-regression Phase 1 & 2

Table D1

Specification Diagnostics
Education with regards to Preference for Self-employment (Strict Definition)

	Probit Model (Including age/100 sq.)	IV-probit Model (Including age/100 sq.)
Sensitivity Pr(+ D)	53.10%	52.93%
Specificity Pr(-~D)	66.96%	66.61%
Positive predictive value Pr(D +)	59.28%	58.94%
Negative predictive value Pr(~D -)	61.19%	60.98%
False + rate for true ~D Pr(+~D)	33.04%	33.39%
False - rate for true D Pr(- D)	46.90%	47.07%
False + rate for class. + Pr(~D +)	40.72%	41.06%
False - rate for class. - Pr(D -)	38.81%	39.02%
Correctly classified	60.37%	60.11%

Classified + if predicted Pr(D) >= .5

Table D2

Multicollinearity diagnostics (VIF) Table
Education with regards to Preference for Self-employment (Strict Definition)

	Probit Model (Including age/100 sq.)		IV-probit Model (Including age/100 sq.)	
Variable	VIF	Tolerance	VIF	Tolerance
Preference	1.07	0.937	1.07	0.937
Gender	1.05	0.953	1.05	0.953
Age	33.83	0.030	33.81	0.030
Age/100 sq.	33.81	0.030	33.74	0.030
Education	1.13	0.885	-	-
Location	1.09	0.917	1.08	0.922
Lack of Fin. Sup.	1.17	0.853	1.17	0.852
Admin. Complex.	1.16	0.859	1.16	0.860
Lack of Info.	1.21	0.829	1.21	0.827
Risk Tolerance	1.15	0.869	1.15	0.871
Father unemployed	1.08	0.923	1.16	0.865
Father self-employed	1.36	0.738	1.72	0.582
Mother unemployed	1.47	0.680	2.38	0.420
Mother self-employed	1.48	0.677	1.88	0.532
Country Dummies	Included		Included	
Mean VIF	3.06		3.08	

Table D3

Specification Diagnostics
Education with regards to Actual Self-employment (Strict Definition)

	Probit Model (Including age/100 sq.)	IV-probit Model (Including age/100 sq.)
Sensitivity Pr(+ D)	19.87%	24.63%
Specificity Pr(-~D)	95.73%	89.12%
Positive predictive value Pr(D +)	56.33%	38.56%
Negative predictive value Pr(~D -)	81.16%	81.00%
False + rate for true ~D Pr(+~D)	4.27%	10.88%
False - rate for true D Pr(- D)	80.13%	75.37%
False + rate for class. + Pr(~D +)	43.67%	61.44%
False - rate for class. - Pr(D -)	18.81%	19.00%
Correctly classified	79.26%	75.12%

Classified + if predicted Pr(D) >= .5

Table D4

Multicollinearity diagnostics (VIF) Table
Education with regards to Actual Self-employment (Strict Definition)

	Probit Model (Including age/100 sq.)		IV-probit Model (Including age/100 sq.)	
Variable	VIF	Tolerance	VIF	Tolerance
Actual S.E.	1.19	0.844	1.19	0.842
Preference	1.17	0.855	1.17	0.855
Gender	1.06	0.943	1.06	0.942
Age	35.44	0.028	35.46	0.028
Age/100 sq.	35.35	0.028	35.36	0.028
Education	1.12	0.894	-	-
Location	1.11	0.903	1.1	0.912
Lack of Fin. Sup.	1.17	0.851	1.18	0.850
Admin. Complex.	1.17	0.852	1.17	0.852
Lack of Info.	1.23	0.814	1.23	0.814
Risk Tolerance	1.13	0.884	1.13	0.887
Father unemployed	1.09	0.916	1.16	0.863
Father self-employed	1.35	0.739	1.74	0.576
Mother unemployed	1.45	0.689	2.36	0.423
Mother self-employed	1.43	0.700	1.83	0.545
Country Dummies	Included		Included	
Mean VIF	3.08		3.11	