The Consequences of Unobserved Heterogeneity in a Sequential Logit Model

Maarten L. Buis

Department of Sociology, Tübingen University, Wilhelmstraße 36, 72074 Tübingen, Germany

Abstract

Cameron and Heckman (1998) established that a sequential logit model is more sensitive than many other models to the possible biasing influence of unobserved heterogeneity. This article proposes a method which allows researchers to find out how large this potential problem is for their data, their model, and their hypothesis of interest. This is done by proposing a set of scenarios for this unobserved heterogeneity, and showing how to estimate the effects of interest given these scenarios. The set of results from these scenarios give an indication of how sensitive the results are to assumptions regarding unobserved heterogeneity. This sensitivity analysis has been applied to a study of educational attainment in the Netherlands, and it showed that that the finding that the effect of father's education declined over transitions is quite sensitive to the assumptions made about unobserved heterogeneity, but that the finding that the effect of father's education declined over birth cohorts is more robust than is often feared.

Keywords: sensitivity analysis, unobserved heterogeneity, Mare model, sequential response model

1. Introduction

an important model for the describing the process of educational attainment is the sequential logit model proposed by Mare (1979, 1980, 1981), which describes this process as a sequence of decisions or steps. For example: 1) whether to finish secondary education or to leave school with only primary education, and 2) whether or not to finish tertiary education given that one finished secondary education. This model also has other applications, for example O'Rand and Henretta (1982) describe the decision when to retire using the following sequence of decisions: 1) whether to retire before age 62 or later, and 2) whether to retire before age 64 or later given that one has not retired before age 62. Cragg and Uhler (1970) describe the demand for automobiles as the result of the following sequence of decisions: 1) whether or not to buy an automobile, 2) whether to add an automobile or to replace an automobile given that one decided to buy an automobile, 3) whether or not to sell an automobile or not given that one decided not to buy an automobile. The sequential logit model consists of separate logistic regression for each step or decision on the sub-sample that is 'at risk' of making that decision. This model is known under a variety of names: sequential logit model (Tutz, 1991), sequential response model (Maddala, 1983), continuation ratio logit (Agresti, 2002), model for nested dichotomies (Fox, 1997), and the Mare model (Shavit and Blossfeld, 1993).

This model has however been subject to an influential critique by Cameron and Heckman (1998). Their main point starts with the observation that the sequential logit model, like any other model, is a simplification of reality and will not include all variables that influence the probability of passing a transition. The presence of these unobserved variables is often called unobserved heterogeneity, and it will lead to biased estimates, even if these unobserved variables are not confounding variables. There are two mechanisms through which these unobserved variables will influence the results. The first mechanism, which I will call the averaging mechanism, is based on the fact that leaving a variable out of the model means that one models the probability of passing a transition averaged over the variable that was left out. This averaging causes problems because the relationship between the variable left out of the model and the probability is non-linear. As a consequence, the effect of the included variables on this average probability is not the same as the effect of these variables on the probability (Neuhaus and Jewell, 1993; Cameron and Heckman, 1998; Allison, 1999; Williams, 2009; Mood, 2010). The second mechanism, which I will call the selection mechanism, is based on the fact that even if a variable is not a confounding variable at the initial transition because it is uncorrelated with any of the observed variables, it will become a confounding variable at the higher transitions because the respondents who are at risk of passing these higher transitions form a selected sub-sample of the original sample (Mare, 1980; Cameron and Heckman, 1998).

The aim of this article is to propose a diagnostic tool that can help researchers determine how big this problem is for their data and their hypotheses. Moreover, if this tool indicates that one needs to do something about unobserved heterogeneity than it will also indicate what the key features of such a model should be. This tool is a sensitivity analysis with which one can investigate the consequences of unobserved variables in a sequential logit model. This will be done by specifying a set of plausible scenarios concerning this unobserved variability and estimating the individual-level effects within each of these scenarios, thus creating a range of plausible values for the individual-level effects. By comparing the results from different scenarios one can get an idea about whether unobserved heterogeneity is a problem, and if so, what aspect of it is most troublesome.

A typical analysis that would use this tool would start with estimating a regular sequential logit model, followed by the sensitivity analysis. If the conclusions turn out to be robust then the analysis is done. If the conclusions turn out to be sensitive, the sensitivity analysis can help the researcher with selecting an appropriate method. For example many techniques for dealing with unobserved heterogeneity rely on a random-effects type assumption that during the first transition the unobserved variable is uncorrelated with the observed variables (for example Mare, 1993; Tam, this issue; Karlson, this issue; Lucas et al., this issue; Holm and Jæger, this issue). With the sensitivity analysis proposed in this article one could investigate whether the parameters and tests of interest are sensitive to this assumption.

Any method for studying effects while controlling for unobserved heterogeneity will have to deal with the fact that it tries to control for variables that have not been observed. A common strategy in is to use information that might be available outside the data. For example one might know that a variable only has an indirect influence on the outcome variable via the main explanatory variable, in which case one can use this variable as an instrumental variable, or one might know that all variables influencing the main explanatory variable are present in the data, in which case one can use propensity score matching. An example of such a strategy that has been applied to the sequential logit model is the model by Mare (1993, 1994), who used the fact that siblings are likely to have a shared family background. If one has data on siblings, one can thus use this information for controlling for unobserved variables on the family level. Another example of this strategy is the model used by Holm and Jæger (this issue), who use instrumental variables in a sequential probit model[1] to identify individual-level effects. The strength of this strategy depends on the strength of the information outside the data that is being used to identify the model. However, such external information is often not available. In those cases, one can still use these models, except that the identification is now solely based on untestable assumptions. This implies a subtle shift in the goal of the analysis: instead of trying to obtain an empirical estimate of a causal effect, one is now trying to predict what would happen if a certain scenario were true. This is not unreasonable: these effects are often the quantity of interest, and if it is not possible to estimate them, then the results of these scenarios are the next best thing. However, the modeling challenge now changes from making the best use of some information that is present outside the data to finding the most informative comparison of scenarios. The goal of such an analysis is to find a plausible range of estimates of the causal effect and to assess how sensitive the conclusions are to changes in the assumptions. In essence, one directly manipulates the source of the problem: the degree of unobserved heterogeneity. This way one can compare how the results would change if there is a small, moderate, or large amount of unobserved heterogeneity. (Rosenbaum and Rubin, 1983; Rosenbaum, 2002; Harding, 2003; DiPrete and Gangl, 2004). In this article I will apply this general idea to the sequential logit model and propose a method of estimating these scenarios that will allow a more general set of scenarios to be estimated and a more general set of parameters and tests to be investigated for sensitivity.

This article will start with a more detailed discussion of how unobserved

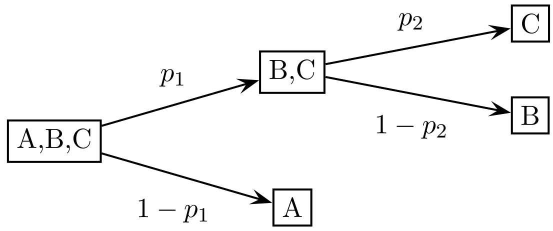
[1]The sequential probit model is similar to the sequential logit model except that the probit link function is used rather than the logit link function.

heterogeneity can cause bias in the estimates of the effect of the observed variables. I will then propose a sensitivity analysis, by specifying a series of scenarios concerning the unobserved variables. The estimation of the effects within these scenarios will be discussed next. Finally, the method will be illustrated by replicating an analysis of the effect of parental background on educational attainment in the Netherlands by De Graaf and Ganzeboom (1993) and Buis (2010a, Chapter 2), and assessing how robust their results are to changes in assumptions about unobserved heterogeneity.

2. The sequential logit model and two effects of unobserved heterogeneity

The effect of unobserved heterogeneity in a sequential logit model is best explained using an example. Figure 1 shows a hypothetical process, which is to be described using a sequential logit model. There are three levels in this process: A, B and C. This process consists of two transitions: the first transition is a choice between A on the one hand and B and C on the other. The second transition is a choice between B and C for those who have chosen B and C in first transition. Whether or not an individual passes the first and second transition is represented by two indicator variables y_1 and y_2 respectively, which receives the value 1 when an individual passes a transition and 0 when it fails that transition. Figure 1 could be a representation of both the educational attainment example and the retirement example in the introduction. In the former case, A would correspond to primary education, B would correspond to secondary education, and C would correspond to tertiary education. In the latter case, A would correspond to retire before age 62, B would correspond to retire between age 62 and 64, and C would correspond to retire after age 64.

Figure 1: Hypothetical process



The sequential logit model models the probabilities of passing these transitions. This is done by estimating a logistic regression for each transition on the sub-sample that is at risk, as in equations (1) and (2). Equation (1) shows that the probability labelled p_1 in Figure 1 is related to two explanatory variables x and u through the function $\Lambda()$, while equation (2) shows the same for the probability labelled p_2 in Figure 1. The function $\Lambda()$ is defined such that $\Lambda(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$. This function ensures that the predicted probability always remains between 0 and 1, by modeling the effects of the explanatory variables as S-shaped curves. The coefficients of x and u (β_{11}, β_{u1}, β_{12}, and β_{u2}) can be interpreted as log odds ratios, while the constants (β_{01} and β_{02}) represent the baseline log odds of passing the first and second transitions.

$$
\begin{aligned}
p_1 = \Pr(y_1 = 1|x, u) &= \Lambda(\beta_{01} + \beta_{11}x + \beta_{u1}u) && (1)\\
p_2 = \Pr(y_2 = 1|x, u, y_1 = 1) &= \Lambda(\beta_{02} + \beta_{12}x + \beta_{u2}u) && (2)
\end{aligned}
$$

Table 1 turns Figure 1 and equations (1) and (2) into a numerical example. Panel (a) shows the counts, the probabilities of passing, the odds and log odds ratios when u is observed, while panel (b) shows what happens in this example when u is not observed. Both x and u are dichotomous (where low is coded as 0 and high as 1), and during the first transition x and u are independent, meaning that u is not a confounding variable at the first transition. The sequential logit model underlying this example is presented in equations (3) and (4).

$$
\begin{aligned}
\Pr(y_1 = 1|x, u) &= \Lambda[\log(.333) + \log(3)x + \log(3)u] && (3)\\
\Pr(y_2 = 2|x, u, y_1 = 1) &= \Lambda[\log(.333) + \log(3)x + \log(3)u] && (4)
\end{aligned}
$$

Consider the first transition in panel (a). The constant in the logistic regression equation is the log odds of passing for the group with value 0 for all explanatory variables, so the constant is in this case log(.333). The effect of x in a logistic regression equation is the log odds ratio. Within the low u group, the odds of passing for the low x group is .333 and the odds of passing for the high x group is 1, so the odds ratio is $\frac{1}{.333} = 3$, and the log odds ratio is log(3). The effect of x in the high u group is also log(3), so there is no interaction effect between x and u. The effect of u can be calculated by comparing the odds of passing for a high u and a low u individual within the low x group, which results in a log odds ratio of log(3). There is no interaction between x and u, so the log odds ratio for u within the high x group is also log(3). Panel (b) shows what happens if one only observes x and y but not u. For example, in that case $300 + 200 = 500$ low x persons are observed to have failed the first transition and $100 + 200 = 300$ low x persons are observed to have passed the first transition. The resulting counts are used to calculate the probabilities, odds, and log odds ratios. Panel (b) shows that the log odds ratios of x are smaller than those computed in panel (a). Leaving u out of the model thus resulted in an underestimation of the effect of x for both the first and the second transition, even though u was initially uncorrelated with x.

This example can be used to illustrate both mechanisms through which unobserved heterogeneity can lead to biased estimates of the individual-level effects. First, the selection mechanism can explain part of the underestimation of the effect of x at the second transition. A characteristic of the sequential logit model is that even if u is not a confounding variable during the first transition, it will become a confounding variable during later transitions (Mare, 1980; Cameron and Heckman, 1998). The example was created such that u and x are independent during the first transition, as the distribution of u is equal for both the low

Table 1: Example illustrating the consequences of not observing a non-confounding variable (u)

(a) while observing u

transition	u	x	y_k 0	1	N	Pr(pass)	odds(pass)	log odds ratio
1	low	low	300	100	400	0.25	0.333	log(3)
		high	200	200	400	0.5	1	
	high	low	200	200	400	0.5	1	log(3)
		high	100	300	400	0.75	3	
			0	1				
2	low	low	75	25	100	0.25	0.333	log(3)
		high	100	100	200	0.5	1	
	high	low	100	100	200	0.5	1	log(3)
		high	75	225	300	0.75	3	

(b) without observing u

transition	x	y_k 0	1	N	Pr	odds	log odds ratio
1	low	500	300	800	0.375	0.6	log(2.778)
	high	300	500	800	0.625	1.667	
		0	1				
2	low	175	125	300	0.417	0.714	log(2.6)
	high	175	325	500	0.65	1.857	

x group and the high x group. As a consequence, u cannot be a confounding variable during the first transition. But this is no longer true during the second transition. For the high x group, the proportion of persons with a high u is $300/500 = .6$, while for the low x group that proportion is $200/300 = .667$. The selection at the first transition has thus introduced a negative correlation between x and u, and u has become a confounding variable. If one does not observe u, and thus can not control for u, one would expect to underestimate the effect of x at the second transition. This could in part explain the underestimation of the effect of x in the second transition in panel (b) of Table 1, but not the underestimation of the effect of x in the first transition.

The averaging mechanism can explain the underestimation of the effect of x during the first transition. The models implicit in panels (a) and (b) have subtly different dependent variables: in panel (a) one is modeling the probability that an *individual* passes the transitions, while in panel (b) one models the *average* probability of passing the transitions. The two result in different estimates because the relationship between the unobserved variables and the probabilities is non-linear. This issue is discussed in terms of the sequential logit model by Cameron and Heckman (1998). It also occurs in other non-linear models, and has been discussed by Neuhaus et al. (1991), Allison (1999), Williams (2009), and Mood (2010). It is also closely related to the distinction between population average or marginal models on the one hand and mixed effects or subject specific models on the other (Fitzmaurice et al. 2004, chapter 13; Agresti 2002, chapter 12). The averaging of the probabilities can be seen in Table 1: for example the probability of passing transition 2 for low x individuals when not controlling for u is $(100 \times 0.25 + 200 \times 0.5)/300 = 0.417$. The consequence of this is that if we think that equations (1) and (2) form the true model for the probabilities of passing the transitions, then the true model for the probabilities averaged over u should be represented by equations (5) and (6), where $\mathrm{E}_u(\cdot)$ is the average of $\cdot$ over u. Instead, the model represented by equations (7) and (8) are estimated when u is not observed and u is thus left out of the model. The two models are not the equivalent because $\Lambda()$ is a non-linear transformation. Neuhaus and Jewell (1993) give an approximation of how β^*_{11} and β^*_{12} deviate from β_{11} and β_{12}: β^*_{11} and β^*_{12} will be smaller than β_{11} and β_{12}, and the difference between the estimates β^*_{11} and β^*_{12} and the estimates β_{11} and β_{12} will increase when the variances of $\beta_{21}u$ and $\beta_{22}u$ increase and when the probability of passing is closer to 50%.

$$\begin{aligned} \mathrm{E}_u(\Pr[y_1 = 1|x, z]) &= \mathrm{E}_u(\Lambda(\beta_{01} + \beta_{11}x + \beta_{u1}u)) && (5) \\ \mathrm{E}_u(\Pr[y_2 = 1|x, z, y_1 = 1]) &= \mathrm{E}_u(\Lambda(\beta_{03} + \beta_{12}x + \beta_{u2}u)) && (6) \end{aligned}$$

$$\begin{aligned} \mathrm{E}_u(\Pr[y_1 = 1|x, z]) &= \Lambda(\beta^*_{01} + \beta^*_{u1}x) && (7) \\ \mathrm{E}_u(\Pr[y_2 = 1|x, z, y_1 = 1]) &= \Lambda(\beta^*_{02} + \beta^*_{u2}x) && (8) \end{aligned}$$

2.1. Comparing effects across transitions

An important sub-question within the debate concerning the potential influence of unobserved heterogeneity on the sequential logit model is whether one can compare coefficients from different transitions (Cameron and Heckman, 1998; Mare, 2006). The concern is that the dependent variable is measured on different scales across transitions, thus making the effect at different transitions incomparable. This often called the scaling problem (for example Allison, 1999; Williams, 2009; Mood, 2010). The argument is clearest in the latent variable interpretation of logistic regression. There are two equivalent ways of deriving a logistic regression model: 1) as a non-linear model for the probability or odds of success, or 2) as a linear model for a latent propensity of success (for example Long, 1997). Within the latent variable representation of logistic regression the scale of the latent propensity is identified by constraining the residual variance to be equal to a fixed number ($\pi^2/3$). When one expects (or cannot rule out) heteroscedasticity the variance of the residual differs across the groups or transitions, which means that the scale on which the outcome variable is measured differs across groups, making the comparison of these groups or transition hard. This is particularly relevant for the sequential logit model, as this model invites the comparison of coefficients across transitions, and the selection that happens at these transitions will likely cause a difference in the variance of the unobserved variables across transitions.

The scaling problem and the averaging mechanism are equivalent in the sense that the mechanism through which unobserved variables influence the results are equivalent. However, there is an important difference. On the one hand, the scaling problem persists even after one controls for all observed and unobserved variables one wants to control for. After one has controlled for all these variables there will still be variation between respondents, which we might call 'idiosyncratic error' and which we may conceptually think of as 'luck'. In the latent variable interpretation the scale of the dependent variable is still dependent on the variance of this idiosyncratic error and this variance is still likely to differ across transitions, thus making comparisons of effects across transitions difficult. On the other hand, the averaging problem disappears as soon as one has controlled for all observed and unobserved variable one wants to control for. Differences in probability or odds can still be due to differences in the variance of the remaining idiosyncratic error, but the scale of the outcome of interest is fixed as the expected proportion of successes (probability) or the expected number of successes per failure (odds). In essence, the problem is defined away by defining the outcome of interest as the probability or odds after controlling for a given set of variables. This is the logic behind the suggestion by Angrist and Pischke (2008) and Mood (2010), to solve the scaling problem by focussing on the probability instead of the latent variable. This article will follow the same logic, except that the effects will be interpreted as odds ratios instead of marginal effects[2]. The main reason for preferring odds ratios over

[2]Mood (2010) advocates the use of marginal effects over odds ratios, but the key element

marginal effects is that the interpretation of interaction effects is much easier in the former metric compared to the latter metric (Compare (Buis, 2010b) with (Ai and Norton, 2003)).

Interpreting effects in terms of odds or probabilities does not solve all problems and arbitrariness. The dependent variable changes when add or remove explanatory variables, as was shown above when discussing the averaging mechanism. However, this is consistent with what the dual nature of what a probability or an odds is supposed to measure: how likely an event is and the degree of uncertainty. Such uncertainty can be thought of as coming from all observed and unobserved variables that were not included in the model. So the dependent variable is in a sense defined by what one chose *not* to control for. It is the modeling decision what variables to control for and by implication what variables not to control for, that makes the scale of the probability or odds meaningfully comparable across groups or transitions.

3. A sensitivity analysis

The previous section discussed what kind of problems unobserved variables might cause. The difficulty with finding a solution for these problems is that it is challenging to control for something that has not been observed. One possible solution is to perform a sensitivity analysis: specify a number of plausible scenarios concerning the unobserved variables, and estimate the effects within each scenario. The aim of this type of analysis is not to obtain an empirical estimate of the effect *per se*, but to assess how important assumptions are for the estimated effect and to obtain a feeling for the range of plausible values for the effect.

A key step in creating such scenarios is to create a set of reasonable scenarios concerning the unobserved variable u. When creating the scenarios, it is more useful to think about u as not being a single unobserved variable but as a (weighted) sum of all the unobserved variables that one wants to control for. There are two equivalent ways of thinking about the scale of this compound unobserved variable. It is sometimes convenient to think of the resulting variable as being standardized, such that its mean is 0 and the standard deviation is 1. This way the 'effect' of u during transition k — β_{uk} — can be compared with the effects of standardized observed variables to get an impression of the range of reasonable values of this 'effect'. Alternatively, it is possible to think of the composite unobserved variable as just being an unstandardized random variable or error term. In this case, the standard deviation of this random variable is the same as β_{uk}. The standardized unobserved variable will be referred to as u, while the unstandardized unobserved variable will be referred to as ν_k in order to distinguish between the two. The two are related in the following way: $\beta_{uk} u = \nu_k$.

of the solution is that probabilities have a known and interpretable scale and the same applies to odds.

All scenarios are variations on the following basic scenario, which is introduced in equations (9) and (10). In this example there are two transitions, with the probabilities of passing these transitions influenced by two variables x and u, where u is as defined above. The observed dependent variables are the probabilities of passing the two transitions averaged over u. So by estimating models (9) and (10), one can recover the true effects of x. To estimate it, all one needs to know is the distribution of $\beta_{uk}u$ $(= \nu_k)$ and to integrate over this distribution. In this article I will consider scenarios where ν_k follows either a normal (Gaussian) distribution, a uniform distribution, or a discrete distribution. The mean of ν_1 will be set at 0 and the effect of u in transition k is set equal to β_{uk}, which are *a priori* fixed in the scenario. This means that a person's value on u will not change over the transitions, but that the effect β_{uk} can change over transitions. Finally, this article will consider scenarios where the correlation between u and x during the initial transition is non-zero.

$$\begin{aligned}
\mathrm{E}_{\nu_1}(\Pr[y_1 = 1|x, \nu_1]) &= \mathrm{E}_{\nu_1}(\Lambda(\beta_{01} + \beta_{11}x + \underbrace{\beta_{u1}u}_{\nu_1})) && (9)\\
&= \int \Lambda(\beta_{01} + \beta_{11}x + \nu_1) f(\nu_1) d\nu_1 \\
\mathrm{E}_{\nu_2}(\Pr[y_2 = 1|x, \nu_2, y_1 = 1]) &= \mathrm{E}_{\nu_2}(\Lambda(\beta_{02} + \beta_{12}x + \underbrace{\beta_{u2}u}_{\nu_2})) && (10)\\
&= \int \Lambda(\beta_{02} + \beta_{12}x + \nu_2) \\
& \quad f(\nu_2) d\nu_2 \\
&= \int \Lambda(\beta_{02} + \beta_{12}x + \beta_{u2}u) \\
& \quad f(\beta_{u2}u|y_1 = 1) du
\end{aligned}$$

The effects in each scenario are estimated using maximum likelihood. Referring back to Figure 1, the likelihood function for an individual i can be written as equation (11), that is, the probability of observing someone with value A equals the probability of failing the first transition, the probability of observing someone with value B equals the probability of passing the first transition and failing the second transition, and the probability of observing someone with value C equals the probability of passing both transitions.

$$L_i = \begin{cases} 1 - p_{1i} & \text{if } y_i = A \\ p_{1i} \times (1 - p_{2i}) & \text{if } y_i = B \\ p_{1i} \times p_{2i} & \text{if } y_i = C \end{cases} \tag{11}$$

By replacing p_{1i} with equation (9) and p_{2i} with equation (10), one gets a function that gives the probability of an observation, given the parameters β. This probability can be computed for each observation and the product of these form the probability of observing the data, given a set of parameters. Maximizing this function with respect to the parameters gives the maximum likelihood estimates. These estimates include the true effects of the variable of interest x assuming that the model for the unobserved heterogeneity is correct.

The difficulty with this likelihood is that there are no closed form solutions for the integrals in equations (9) and (10) if ν_k follows a normal distribution[3]. This can be resolved by numerically approximating these integrals using maximum simulated likelihood (Train, 2003). Maximum simulated likelihood uses the fact that the integral is only there to compute a mean probability. This mean can be approximated by drawing at random many values for ν_k from the distribution of ν_k, computing the probability of passing a transition assuming that this randomly drawn value is the true value of ν_k, and then computing the average of these probabilities. This approach can be further refined by realizing that using true random draws is somewhat inefficient as these tend to cluster. Increasing the efficiency is important as these integrals need to be computed for each observation, meaning that these simulations need to be repeated for each observation. One can cover the entire distribution with less draws if one can use a more regular sequence of numbers. An example of a more regular sequence of numbers is a Halton (1960) sequence. A Halton sequence will result in a more regular series of quasi-random draws from a uniform distribution. These quasi-random draws can be transformed into quasi-draws from a normal distribution or a discrete distribution by applying the inverse cumulative distribution function. These are then used to compute the average probability of passing the first transition, as is shown in equation (12), where m represents the number of draws from the distribution of ν_1. At the second transition, the distribution of ν_2 no longer follows the original distribution, but is now conditional on being at risk. The integral over this distribution is computed by drawing ν_2 from the original distribution as before, but then computing a weighted mean whereby each draw is given a weight equal to the probability of being at risk assuming that that draw was the true ν_2, as is shown in equation (13). In the appendix I show that this is a special case of importance sampling (Robert and Casella, 2004, 90–107). This procedure is implemented in the `seqlogit` package (Buis, 2010c) in Stata (StataCorp, 2009), using the facilities for generating Halton sequences discussed by Drukker and Gates (2006).

$$\mathrm{E}_\varepsilon(\Pr(y_1 = 1|x, u)) \approx \frac{1}{m}\sum_{j=1}^{m} \Lambda(\beta_{01} + \beta_{11}x + \beta_{u1}u) \tag{12}$$

$$\mathrm{E}_\varepsilon(\Pr(y_2 = 1|x, u, y_1 = 1) \approx \frac{\sum_{j=1}^{m}\left[\Pr(y_1=1|x,u_j)\Lambda(\beta_{02}+\beta_{12}x+\beta_{u2}u_j)\right]}{\sum_{j=1}^{m}\Pr(y_1=1|x,u_j)} \tag{13}$$

[3]A closed form solution does exist when ν_k follows a discrete distribution, but the implementation of this method in the `seqlogit` package (Buis, 2010c) will not use this analytical solution for technical reasons.

4. An example: The effect of family background on educational attainment in the Netherlands

An important application for the sequential logit model is the study of the influence of family background on educational attainment (for recent reviews see: Breen and Jonsson, 2005; Hout and DiPrete, 2006). The potential problems that unobserved variables can cause were recognized from the time that the sequential logit model was introduced in this literature (Mare, 1979, 1980, 1981), but interest in this issue has been revived by the critique from Cameron and Heckman (1998). However, only a limited number of empirical studies have tried to actually account for unobserved heterogeneity (for exceptions see: Mare, 1993; Rijken, 1999; Chevalier and Lanot, 2002; Lauer, 2003; Arends-Kuenning and Duryea, 2006; Colding, 2006; Lucas et al., this issue; Holm and Jæger, this issue; Karlson, this issue; Tam, this issue). The method proposed in this paper will be illustrated by replicating an analysis that does not control for unobserved heterogeneity by De Graaf and Ganzeboom (1993) and Buis (2010a, Chapter 2). These studies estimated the effect of father's occupational status and education on transition probabilities between educational levels in the Netherlands. The original study by De Graaf and Ganzeboom (1993) was part of an influential international comparison of the effect of family background on educational attainment (Shavit and Blossfeld, 1993). It used 10 Dutch surveys that were post-harmonized as part of the International Stratification and Mobility File [ISMF] (Ganzeboom and Treiman, 2009). Buis (2010a, Chapter 2) updated this analysis by using an additional 33 Dutch surveys that have since been added to the ISMF.

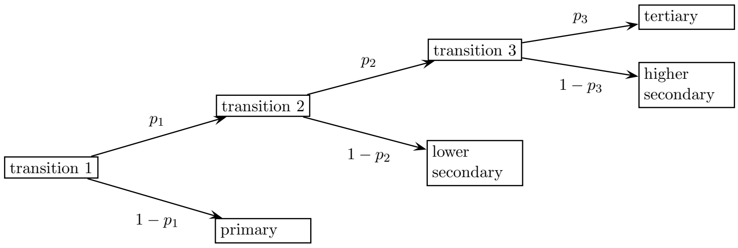
4.1. The data

The total of 43 surveys are all part of the ISMF and were held between 1958 and 2006. The individual surveys are described in detail in the appendix of (Buis, 2010a, chapter 2). Only male respondents older than 25 are used in the analysis. These surveys contain 35,829 men with valid information on all variables used in the model[4]. Family background is measured as father's occupational status and father's highest achieved level of education. Time is measured by 10-year birth cohorts covering the cohorts that were born between 1891 and 1980. The main effect of birth cohort is added as a set of dummies, while the effects of family background variables is allowed to change linearly over cohorts.

Father's occupational status was measured using the International Socio-Economic Index (ISEI) of occupational status (Ganzeboom and Treiman, 2003), which originally ranged between 10 and 90 and was recoded to range between 0 and 8. In concordance with De Graaf and Ganzeboom (1993) and Buis (2010a, Chapter 2), education of both, the father and the respondent, was measured

[4]This count deviates slightly from Buis (2010a, Chapter 2) due to a coding error in the original article involving 63 observations.

Figure 2: Simplified model of the Dutch education system



in four categories: primary education (LO), lower second secondary education (LBO and MAVO), higher secondary education (HAVO, MBO, and VWO), and tertiary education (HBO and WO). The transitions that were studied by De Graaf and Ganzeboom (1993) and Buis (2010a, Chapter 2) are: 1) from primary education or less to a diploma in secondary or tertiary education; 2) from a diploma in lower secondary education to a diploma in higher secondary or tertiary education; and 3) from a diploma in higher secondary education to completed tertiary education. These transitions are displayed in Figure 2.

4.2. The scenarios

In the sensitivity analysis I focuss on the effect of father's education, and in particular on two hypotheses: 1) the effect of father's education decreases over birth cohorts ,and 2) the effect of father's education decreases over transitions. The first hypothesis was chosen as an example because it played such a central role in the study of educational attainment (for example Mare, 1981; Shavit and Blossfeld, 1993; Breen et al., 2009). The second hypothesis was chosen because it is suspected of being particularly sensitive to unobserved heterogeneity (for example Mare, 1980; Cameron and Heckman, 1998).

The sensitivity analysis is broken up in the following five sets of scenarios, each exploring different ways in which unobserved heterogeneity could influence the results.

The first set of scenarios is used to investigate the impact of the amount of unobserved heterogeneity (β_{uk}) on the statistics of interest. The aim is find out how extreme a scenario needs to be before conclusions change. In order to figure out how extreme a scenario is, one needs to have an idea about what a reasonable value for β_{uk} might be. Remember that u is a standardized variable and that β_{uk} is its effect in terms of log odds ratios on the odds of passing transition k. Looking in the literature at the effects of standardized variables that are likely to have very large effects can help to pin down a reasonable upper bound for β_{uk}. A good candidate is for instance prior academic performance because we can expect this variable to have a large effect and because recent empirical estimates of its effect exist in the literature. The size of this effect appears to be approximately 2.5, both in the Netherlands (Kloosterman et al., 2009) and in the United Kingdom (Erikson et al., 2005). This is a very large effect, meaning that a standard deviation increase leads to an increase in the odds of passing of approximately a factor 12 ($e^{2.5}$). The replication study by Buis (2010a, p. 148) found that the effects of standardized father's education and standardized father's occupation are .82 and 1.45 respectively. So 2.5 seems like a very high value that could still occur in real data. In order to make sure that the set of scenarios include very extreme scenarios, models will be estimated where β_{uk} will be fixed at 0 , 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5 respectively. Within this set of scenarios, the values of β_{uk} will be assumed to be constant over transitions, the correlation between u and father's education during the first transitions is assumed to be 0, and u is assumed to be normally distributed.

The second set of scenarios is used to investigate the impact of changes in β_{uk} over transitions on the statistics of interest. Empirically we find that the effects of most variables decrease over transitions, so the assumption that the effect of u remains constant over the transitions is probably not realistic. Moreover, the test of the second hypothesis — the effect of father's education decreases over transitions — may be particularly sensitive to deviations from this assumption. If the size of the effect of father's education is influenced by the amount of unobserved heterogeneity (β_{uk}) and if this amount of unobserved heterogeneity changes over transitions, then this will influence the trend in the effect of father's education over transitions. For these scenarios it would be reasonable to let β_{uk} decrease over the transitions, just as most other variables. However, it would also be reasonable to have scenarios in which β_{uk} increases over transitions. The unobserved variable u is often thought of as some sort of ability or school readiness (for example Cameron and Heckman, 1998), and we might expect selection on ability to increase over transitions if only because teachers and other school officials might find it harder to observe that ability in younger children. In this set of scenarios β_{u1}, the β_u at the first transition, will be fixed at 2.5, and at each subsequent transition it will be increased by a fixed number. This set of scenarios consists of 7 such increments: -1.2, -0.8, -0.4, 0, 0.4, 0.8, 1.2. These factors were chosen because this way the scenario with most extreme decrease (-1.2) implies that β_u in the finale transition is virtually 0 (0.1), while the scenario with the most extreme increase (1.2) implies that β_u in the final transition is close to 5, the maximum in the previous set of scenarios. The intermediate scenarios represent changes over transitions that could plausibly occur in real data. Otherwise, the scenarios are equal to the first set of scenarios.

The third set of scenarios is used to investigate the impact of changes in β_{uk} over cohorts. Just as we expect the results regarding the second hypothesis to be particularly sensitive to changes in β_{uk} over transitions, we might expect the results regarding the first hypothesis to be particularly sensitive to changes in β_{uk} over birth cohorts. If we think that u is primarily ability, then it would be reasonable to have scenarios where β_{uk} increases over birth cohorts as we might believe or hope that selection on family background has gradually been replaced by selection on ability. It would also be reasonable to include scenarios in which β_{uk} decreases over birth cohorts as we might be more sceptical and believe that schools have just become less selective over time, because they face increasing pressure to let more and more students pass. In this set of scenarios, I assume that the β_{uk} is in the first cohort 2.5, and than decreases or increases linearly over birth cohorts. The scenarios let β_{uk} change with the following amounts per decade: -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3. These values were chosen because they have similar consequences as the second set of scenarios: in the most extreme decreasing scenario β_{uk} ends up with a value of 0.1 and in the most extreme increasing scenario, β_{uk} ends up with a value of 4.9.

The fourth set of scenarios is used to investigate the impact of the degree to which u is a confounding variable on the statistics of interest. If we think of u as consisting mainly of ability then it is unlikely that u is completely independent of father's education. I will fix the correlation between u and father's education at -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, and β_{uk} at 2.5. Otherwise, the scenarios are equal to the first set of scenarios.

The final set of scenarios is used to investigate the impact of the distribution of u on the statistics of interest. Cameron and Heckman (1998) make the case that assumptions about the distribution of u may influence the results. In this set of scenarios I will consider two deviations form normality. First, I will assume that u follows a uniform distribution. This is a less exotic assumption than it may appear at first glance. Assume that u consists mainly of an ability score as observed by teachers. In that case it is reasonable to assume that teachers find it easier to determine a rank ordering of children than some absolute ability measure. The distribution of a rank order is the uniform distribution. This assumption will be implemented in two ways: First I will force u to follow a continuous uniform distribution, and second I will approximate this distribution by a discrete distribution where four mass points each receive a probability of 1/4 and the locations of the mass points are chosen such that the average is 0, the standard deviation is 1, and the mass points are equally spaced (leading to the following locations for the mass points -1.34, -0.45, 0.45, 1.34). The discrete distribution is much more flexible and allows us to approximate a wide range of distributions (Lindsay, 1983; Heckman and Singer, 1984)[5]. I used both methods here in order to get an idea how reasonable this approximation works in this context. As the second type of deviation from normality, I consider skewed distributions. Right skewed distributions are approximated by discrete distributions with 4 mass points that receive the probabilities 0.1, 0.6, 0.2, 0.1 (with locations -1.66, -0.38, 0.90, 2.18) and 5 mass points that receive the probabilities 0.1, 0.6, 0.2, 0, 1 (with locations -1.37, -0.39, 0.59, 1.57, 2.55). By adding the mass point with 0 probability, the right tail of the distribution was moved further out, thus increasing the skewness. Left skewed distributions were approximated with discrete distributions contain 4 mass points with probabilities 0.1, 0.2, 0.6, 0.1 (with locations -2.18, -0.90, 0.38, 1.66), and discrete distributions with 5 mass points with 0.1, 0, 0.2, 0.6, 0.1 (with locations -2.55, -1.57, -0.59, 0.39, 1.37). The amount of unobserved heterogeneity is fixed at

[5](Lindsay, 1983; Heckman and Singer, 1984) go further and actually try to estimate this distribution, while in this article the discrete distribution is only used to approximate different distributions to create different scenarios.

Table 2: Description of the different sets of scenarios

baseline	
variable	$\beta_u \in \{0, 2.5\}$
fixed	corr(u, fed) $= 0$, $\Delta_{\text{cohort}} = 0$, $\Delta_{\text{transition}} = 0$, $u \sim N(0, 1)$
Changing quantity of unobserved heterogeneity, super-set I	
variable	$\beta_u \in \{0.0, 0.5, \cdots, 5\}$
fixed	corr(u, fed) $= 0$, $\Delta_{\text{cohort}} = 0$, $\Delta_{\text{transition}} = 0$, $u \sim N(0, 1)$
Unobserved heterogeneity changing over transitions, super-set II	
variable	$\Delta_{\text{transition}} \in \{-1.2, -0.8, \cdots, 1.2\}$
fixed	$\beta_{u1} = 2.5$, corr(u, fed) $= 0$, $\Delta_{\text{cohort}} = 0$, $u \sim N(0, 1)$
Unobserved heterogeneity changing over cohorts, super-set II	
variable	$\Delta_{\text{cohort}} \in \{-0.3, -0.2, \cdots, 0.3\}$
fixed	$\beta_{u1} = 2.5$, corr(u, fed) $= 0$, $\Delta_{\text{transition}} = 0$, $u \sim N(0, 1)$
Changing correlation between unobserved variable and father's education, super-set II	
variable	corr(u, fed) $\in \{-0.6, -0.4, \cdots, 0.6\}$
fixed	$\beta_u = 2.5$, $\Delta_{\text{cohort}} = 0$, $\Delta_{\text{transition}} = 0$, $u \sim N(0, 1)$
Changing distribution of unobserved variable, super-set II	
variable	$u \sim N(0, 1)$ $u \sim U(-1.73, 1.73)$ $u \sim D(-1.34, 0.25; -0.45, 0.25; 0.45, 0.25; 1.34, 0.25)$ $u \sim D(-1.66, 0.1; -0.38, 0.6; 0.90, 0.2; 2.18, 0.1)$ $u \sim D(-1.37, 0.1; -.039, 0.6; 0.59, 0.2; 1.57, 0; 2.55, 0.1)$ $u \sim D(-2.18, 0.1; -0.90, 0.2; 0.38, 0.6; 1.66, 0.1)$ $u \sim D(-2.55, 0.1; -1.57, 0; -0.59, 0.2; 0.39, 0.6; 1.37, 0.1)$
fixed	$\beta_u = 2.5$, corr(u, fed) $= 0$, $\Delta_{\text{cohort}} = 0$, $\Delta_{\text{transition}} = 0$

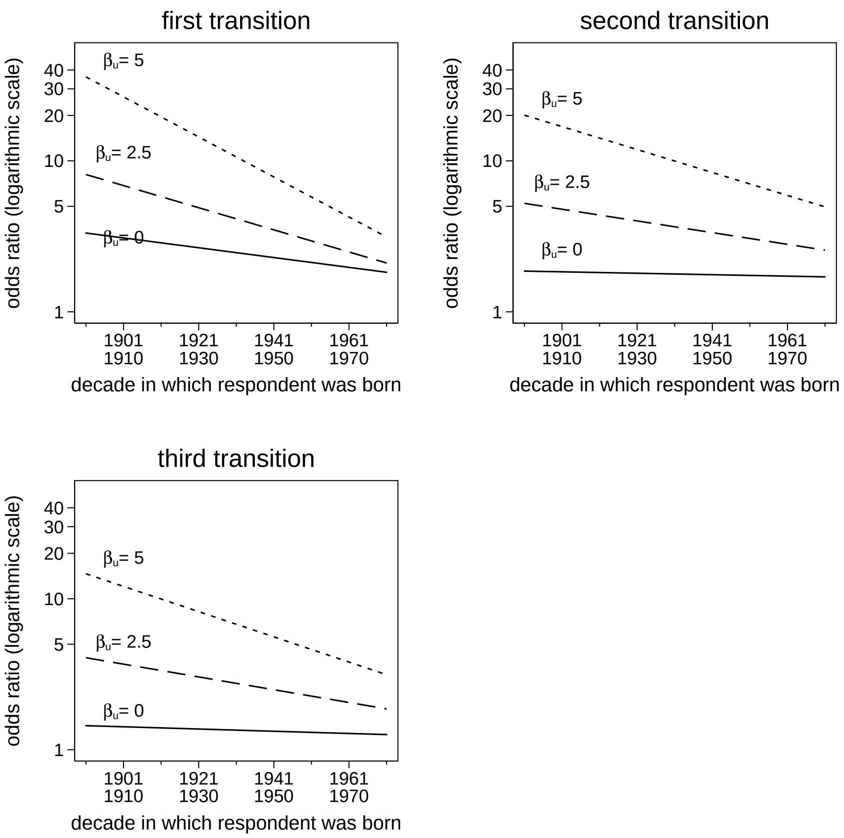
2.5, otherwise the scenarios are the same as the first set of scenarios.

These five sets of scenarios are clustered into two super-sets that differ regarding to the relevant baseline model: The first super-set uses the model without unobserved heterogeneity as the baseline and consists of the first set of scenarios. The second super-set uses the model that fixes β_{uk} at 2.5 (constant across transitions and birth cohorts) and fixes u to be normally distributed and uncorrelated with any of the observed variables as its baseline model. It consists of the remaining sets of scenarios. The second super-set checks whether the conclusions of the sensitivity analysis conducted using the first super-set are robust. All scenarios are summarized in Table 2.

4.3. The results

Figure 3 gives a first idea of what unobserved heterogeneity does to the estimates of the effect of father's education. The lines labeled with $\beta_u = 0$ represent what one would find using a regular sequential logit model. The lines labeled $\beta_u = 2.5$ and $\beta_u = 5$ represent the effects we would have found if the true amount of unobserved heterogeneity equaled 2.5 and 5 respectively and we

Figure 3: Influence of β_u on the effect of father's education



correctly controlled for that. This figure clearly shows that a regular sequential logit model in all likelihood underestimates the effect of father's education. It also shows that such a model underestimates the trend in that effect over birth cohorts, but whether it underestimates or overestimates the trend in the effect over transitions is less clear.

A more detailed description of the effects in the different scenarios is given in tables 4 and 5. The results of the baseline models with which these scenarios are compared are reported in table 3. The baseline model for the first super-set of scenarios is the model without unobserved heterogeneity, that is, $\beta_u = 0$, while the baseline model of the second super-set is the model that fixes β_u at 2.5. In the first baseline model the trends in the effect of father's education over birth cohorts and across transitions are all negative and the one-sided hypothesis that the trends are 0 or positive can be rejected at the 5% level for all trends except

Table 3: Results for baseline models

H_0	coefficient[a]		p-value	
	$\beta_u = 0$	$\beta_u = 2.5$	$\beta_u = 0$	$\beta_u = 2.5$
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-7.2	-15.5	0.000	0.000
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-1.1	-8.5	0.078	0.000
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-1.7	-9.3	0.016	0.000
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-17.9	3.8	0.000	0.812
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-25.2	-26.0	0.000	0.000

[a] % change in odds ratio of fathers education for a decade change in cohort or between transitions, i.e. $(\exp[\beta_{\text{fed}\times\text{coh}k}] - 1) \times 100\%$ and $(\exp[\beta_{\text{fed}k+1} - \beta_{\text{fed}k}] - 1) \times 100\%$, respectively.

for the trend over cohorts in the second transition[6]. In the second baseline model most of the trends are stronger than in the first baseline model and the trend over cohorts in the second transition is now also significantly negative. However, now the trend in effect of father's education from the first to the second transition has become positive and non-significant.

The first two columns of Table 4 shows that changing the amount of unobserved heterogeneity can lead to substantial variation in the size of the trend. The last three columns show that there are two deviations from the baseline model: The non-significant trend over cohorts in the second transitions already turns significant with a small amount of unobserved heterogeneity. This would indicate that the trend is in all likelihood really negative. The second deviation from the baseline model is the trend from the first to the second transition. This shows more worrying pattern in that the variation in coefficient and significance level is large and the change occur only after β_u reached the relatively high but still reasonable value of 2. This means that there are reasonable scenarios where the trend is negative and other equally reasonable scenarios where the trend is not negative, which means that the conclusion concerning this trend is not robust. The test results concerning the remaining trends are not influenced by the amount of unobserved heterogeneity.

An interesting finding is that the trend in effect from the first to the second transition not only changes in significance but also the parameter changes from negative to positive. This is consistent with the two mechanisms through which unobserved variables could influence the results. Both the averaging and selection mechanism tend to lead to an underestimation of effects in a regular sequential logit model, so corrections tend to increase the effect. However, the selection mechanism plays no role during the first transition as no selection has taken place yet. Because of that one can expect that corrections for unobserved

[6] The coefficients are shown in terms of % change in odds, while the test concerns the asymptotically equivalent null hypothesis that the change in log(odds ratio) is larger than or equal to 0. The reason for this difference is that sampling distribution of the log(odds ratio) is more likely to be normally distributed (Sribney and Wiggins, 2010).

Table 4: Results for scenarios, super-set I

H_0	coefficient[a]		p-value		non-robust scenarios[b]
	min	max	min	max	
Changing quantity of unobserved heterogeneity					
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-26.3	-7.2	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-16.1	-1.1	0.000	0.078	$\beta_u \geq 0.5$
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-17.5	-1.7	0.000	0.016	-
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-17.9	22.0	0.000	0.999	$\beta_u \geq 2$
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-34.4	-24.0	0.000	0.000	-

[a] % change in odds ratio of fathers education for a decade change in cohort or between transitions, i.e. $(\exp[\beta_{\text{fed}\times\text{coh}k}] - 1) \times 100\%$ and $(\exp[\beta_{\text{fed}k+1} - \beta_{\text{fed}k}] - 1) \times 100\%$, respectively.
[b] deviate from baseline model with $\beta_u = 0$

heterogeneity tend to lead to a larger increase in effect during the second transition than during the first transition, which in turn can lead to the reversal of the trend across transitions that was observed[7].

The scenarios in Table 4 makes rather strong assumptions: β_u is constant across birth cohorts and transitions, and the unobserved variable u is normally distributed and uncorrelated with any of the observed variables. The scenarios in Table 5 relax in turn each of these assumptions. The first set of scenarios in Table 5 shows that, as expected, allowing β_u to change across transitions only influences the test results concerning the trend across transitions. The trend form the first to the second transition appears to be particularly sensitive, while the trend from the second to third transition is only influenced by the most extreme scenario. That would indicate that the results concerning the latter trend is still fairly robust.

The second set of scenarios in Table 5 shows, against expectation, that changes in β_u across cohorts has fairly little effect on the conclusions concerning the trend in effect of father's education across cohorts. The only exception being the trend during the second transition, but than only in the most extreme scenario. So the trend across cohorts seems to be fairly robust against changes in the amount of unobserved heterogeneity across cohorts. However, these changes do seem to have an influence on the trend from the first to the second transition.

The third set of of scenarios in Table 5 show that correlation between the

[7] However, this is not the only possible outcome because the selection and averaging mechanisms interact with one another. Selection tends to decrease the variance of the unobserved variable at higher transitions which in turn tends to decrease the averaging mechanism at higher transitions. This might become an issue when the amount of selection is extreme and the distribution of the unobserved variable is bounded, as in that case the expected negative correlation between the unobserved variable and the observed variables is likely to be less or even positive (Cameron and Heckman, 1998), and the reduction in the variance of the unobserved variable at higher transition is likely to be stronger. Such a scenario is however unlikely to occur in most educational systems.

unobserved variable and father's education at the first transition has no effect on any of the test results. This is good news as many of the techniques for dealing with unobserved heterogeneity discussed in this special issue rely on the assumption that the initial correlation between the observed and unobserved variables is 0. This result would indicate that such models could be used for this data and these hypotheses.

The final set of the scenarios showed that the assumption of normality for the unobserved variable is not quite harmless. In particular allowing for skewed distributions seems to be important when one tries to control for unobserved heterogeneity as these assumptions does change some the of the conclusion.

Table 5: Results for scenarios, super-set II

H_0	coefficient[a] min	coefficient[a] max	p-value min	p-value max	non-robust scenarios[b]
Unobserved heterogeneity changing over transitions					
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-15.6	-15.5	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-11.9	-5.1	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-17.0	-2.0	0.000	0.007	-
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-26.4	53.6	0.000	1.000	$\Delta_{\text{transition}} < 0$
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-39.0	-1.9	0.000	0.301	$\Delta_{\text{transition}} = 1.2$
Unobserved heterogeneity changing over cohorts					
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-15.5	-13.3	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-10.6	-2.6	0.000	0.074	$\Delta_{\text{cohort}} = 0.3$
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-10.8	-5.3	0.000	0.001	-
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-13.9	18.7	0.000	0.998	$\Delta_{\text{cohort}} \leq -0.2$
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-32.1	-23.8	0.000	0.000	-
Changing correlation between unobserved variable and father's education					
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-15.5	-13.4	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-8.5	-6.9	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-9.3	-7.6	0.000	0.000	-
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-0.6	3.8	0.433	0.812	-
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-26.0	-24.9	0.000	0.000	-
Changing distribution of unobserved variable					
$\beta_{\text{fed}\times\text{coh1}} \geq 0$	-22.4	-10.7	0.000	0.000	-
$\beta_{\text{fed}\times\text{coh2}} \geq 0$	-11.7	1.0	0.000	0.770	left skewed
$\beta_{\text{fed}\times\text{coh3}} \geq 0$	-19.3	-5.4	0.000	0.000	-
$\beta_{\text{fed2}} - \beta_{\text{fed1}} \geq 0$	-25.5	34.0	0.000	1.000	extreme left skewed
$\beta_{\text{fed3}} - \beta_{\text{fed2}} \geq 0$	-44.7	-3.3	0.000	0.102	right skewed

[a] % change in odds ratio of fathers education for a decade change in cohort or between transitions, i.e. $(\exp[\beta_{\text{fed}\times\text{coh}k}] - 1) \times 100\%$ and $(\exp[\beta_{\text{fed}k+1} - \beta_{\text{fed}k}] - 1) \times 100\%$, respectively.

[b] deviate from baseline model with $\beta_u = 2.5$

5. Conclusion and discussion

The aim of this article is to present a sensitivity analysis that can be used to investigate the consequences of unobserved variables in a sequential logit model. The potential bias that these unobserved variables cause are the result of two mechanisms: First, the averaging mechanism is based on the fact that when a variable is left out of the model, one models the probability of passing the transitions averaged over the variable that is left out. As a consequence, estimates of effects of the observed variables on the probability of passing the transitions are effects on the average probability rather than the effects on the probability of passing. These two are different because the unobserved variable is related to the probabilities through a non-linear function. Second, the selection mechanism is based on the fact that a variable that is not a confounding variable at the first transition is likely to become a confounding variable at later transitions. The reason for this is that the process of selection at earlier transitions will introduce correlation between observed and unobserved variables.

The method proposed in this article to investigate the consequences of unobserved heterogeneity is to perform a sensitivity analysis by specifying scenarios regarding unobserved heterogeneity, and estimating the effects of the observed variables given those scenarios. This will not give an empirical estimate of the effects of interest, but it does give an idea about the sensitivity of the estimates to assumptions about unobserved heterogeneity, the direction of the bias, the size of the bias, and the range of likely values of the effect. It is often useful to organize a sensitivity analysis by specifying different sets of scenarios, each exploring different ways in which unobserved heterogeneity could influence the results. In the empirical example I discussed five such sets of scenarios: The first set explores the effect of the amount of unobserved heterogeneity. The second set explores the effect of changes in the amount of unobserved heterogeneity over transitions. The third set of scenarios explores the effect of changes in the amount of unobserved heterogeneity over cohorts. The fourth set explores the effect of correlation between the unobserved variables and the observed variables during the first transition. The fifth set explores the effect of different distributions of the unobserved variable. The effects of the observed variables within each scenario are estimated by maximum likelihood. The likelihood is defined by integrating over the unobserved variable, which is done using Maximum Simulated Likelihood (Train, 2003).

This method was illustrated by replicating a study by De Graaf and Ganzeboom (1993) and Buis (2010a, Chapter 2) on the effect of the father's occupational status and education on the offspring's educational attainment. The sensitivity analysis showes that the test whether the effect of father's education decreased over birth cohorts are rather robust, but that the test of whether the effect of father's education decreased over transitions is rather sensitive. Moreover, the effect of both father's education and its trend over birth cohorts are likely to be underestimated, as these effects are stronger in scenarios with more unobserved heterogeneity.

To sum up, Cameron and Heckman's (1998) finding that unobserved heterogeneity might be more of a problem in sequential logit models than in other models is useful, but in order to have practical implications on the way we do research, we also need to have an idea of how large this problem is. A sequential logit model that assumes that there is no unobserved heterogeneity is a simplification of reality. However, that is in itself not a problem, the whole purpose of a model is that it simplifies reality. The thing to worry about is that these simplifications have such a strong influence on the results that they rather than the observations influence the conclusions. Whether unobserved heterogeneity has a noticeable influence on the conclusions depends not only on the unobserved variables but also on the exact hypothesis being tested and the observed data. The sensitivity analysis proposed in this article can help researchers determine whether they need to worry about unobserved heterogeneity in their data given the hypotheses that they want to test, and if so, what aspects of unobserved heterogeneity they need to pay particular attention to.

Appendix: Sampling from the distribution of ν_k conditional on having passed the previous transitions

One method of sampling from a distribution is importance sampling (Robert and Casella, 2004, 90–107). This appendix shows that the method used in this article is a special case of importance sampling. The idea behind importance sampling is that instead of sampling from the distribution of interest $f(u)$ one draws samples from another distribution $g(u)$, and computes the mean by weighting each draw by $\frac{f(u_j)}{g(u_j)}$, so one could approximate $\mathrm{E}_u[\Lambda(\beta_{02}+\beta_{12}x+\beta_{u2}u)]$ with equation (14).

$$\mathrm{E}_u[\Lambda(\beta_{02} + \beta_{12}x + \beta_{u2}u)] \approx \frac{1}{m}\sum_{j=1}^{m} \frac{f(u_j)}{g(u_j)}\Lambda(\beta_{02} + \beta_{12}x + \beta_{u2}u_j) \tag{14}$$

In this article the distribution of interest is the distribution conditional on being at risk, while the other distribution is the distribution not conditional on being at risk. These distributions are independent of x, so the conditioning on x in equation (15) is superfluous, but this will prove useful later on.

$$\mathrm{E}_u[\Lambda(\beta_{02}+\beta_{12}x+\beta_{u2}u)] \approx \frac{1}{m}\sum_{j=1}^{m} \frac{f(\beta_{u2}u_j|x, y_1 = 1)}{f(\beta_{u2}u_j|x)}\Lambda(\beta_{02}+\beta_{12}x+\beta_{u2}u_j) \tag{15}$$

Instead of using equation (15) directly, the integral is computed using equation (16). The aim of this appendix is to show that these two are equivalent.

$$\mathrm{E}_\varepsilon[\Lambda(\beta_{02} + \beta_{12}x + \varepsilon)] \approx \frac{\sum_{j=1}^{m} \left[\Pr(y \in \{B, C\}|x, \varepsilon_j)\Lambda(\beta_{02} + \beta_{12}x + \varepsilon)\right]}{\sum_{j=1}^{m} \Pr(y \in \{B, C\}|x, \varepsilon_j)} \tag{16}$$

The denominator of equation (16) can be rewritten as in equation (17), which leads to equation (18)

$$\begin{aligned}\sum_{j=1}^{m} \Pr(y_1 = 1|x, u_j) &= m\frac{\sum_{j=1}^{m} \Pr(y_1 = 1|x, u_j)}{m} \\ &\approx m\Pr(y_1 = 1|x) \end{aligned} \tag{17}$$

$$\mathrm{E}_u[\Lambda(\beta_{02}+\beta_{12}x+\beta - u2u)] \approx \frac{1}{m}\sum_{j=1}^{m} \frac{\Pr(y_1 = 1|x, u_j)}{\Pr(y_1 = 1|x)}\Lambda(\beta_{02}+\beta_{12}x+\beta_{u2}u_j) \tag{18}$$

Comparing equations (15) and (18) indicates that the problem can be simplified to showing that equation (19) is true.

$$\frac{f(u_j|x, y_1 = 1)}{f(u_j|x)} = \frac{\Pr(y_1 = 1|x, u_j)}{\Pr(y_1 = 1|x)} \tag{19}$$

Equation (19) can be rewritten as equation (20). Using Bayes' theorem, equation (20) can be rewritten as equation (21). Equation (21) is true, thus showing that equations (15) and (16) are equivalent. Notice, however, that this is based on the approximation in equation (17), which will improve as the number of samples m increases.

$$f(u_j|x, y_1 = 1)\Pr(y_1 = 1|x) = \Pr(y_1 = 1|x, u_j)f(u_j|x) \tag{20}$$

$$f(u_j \cap y_1 = 1|x) = f(y_1 = 1 \cap u_j|x) \tag{21}$$

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. Wiley-Interscience, Hoboken, NJ.

Ai, C., Norton, E. C., 2003. Interaction terms in logit and probit models. Economics letters 80 (1), 123–129.

Allison, P. D., November 1999. Comparing logit and probit coefficients across groups. Sociological Methods & Research 28 (2), 186–208.

Angrist, J. D., Pischke, J.-S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Arends-Kuenning, M., Duryea, S., 2006. The effect of parental presence, parents' education, and household headship on adolescents' schooling and work in Latin America. Journal of Family and Economic Issues 27 (2), 263–286.

Breen, R., Jonsson, J. O., 2005. Inequality of opportunity in comparative perspective: Recent research on educational and social mobility. Annual Review of Sociology 31, 223–243.

Breen, R., Luijkx, R., Müller, W., Pollak, R., 2009. Nonpersistent inequality in educational attainment: Evidence from eight european countries. American Journal of Sociology 114 (5), 1475–1521.

Buis, M. L., 2010a. Inequality of educational outcome and inequality of educational opportunity in the netherlands during the 20^{th} century. Ph.D. thesis, Faculty of Social Sciences, VU-University Amsterdam.

Buis, M. L., 2010b. Stata tip 87: Interpretation of interactions in non-linear models. The Stata Journal 10 (2), 305–308.

Buis, M. L., 2010c. `seqlogit`: Stata module to fit a sequential logit model, version 1.1.12, `http://econpapers.repec.org/software/bocbocode/S456843.htm`.

Cameron, S. V., Heckman, J. J., 1998. Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. The Journal of Political Economy 106 (2), 262–333.

Chevalier, A., Lanot, G., 2002. The relative effect of family characteristics and financial situation on educational achievement. Education Economics 10 (2), 165–181.

Colding, B., 2006. A dynamic analysis of educational progression of children of immigrants. Labour Economics 13 (4), 479–492.

Cragg, J. G., Uhler, R. S., 1970. The demand for automobiles. The Canadian Journal of Economics 3 (3), 386–406.

De Graaf, P. M., Ganzeboom, H. B., 1993. Family background and educational attainment in the Netherlands for the 1891-1960 birth cohorts. In: Shavit, Y., Blossfeld, H.-P. (Eds.), Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder, Ch. 4, pp. 75–99.

DiPrete, T. A., Gangl, M., 2004. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimates and instrumental variables with imperfect instruments. Sociological Methodology 34, 271–310.

Drukker, D. M., Gates, R., 2006. Generating Halton sequences using Mata. The Stata Journal 6 (2), 214–228.

Erikson, R., Goldthorpe, J. H., Jackson, M., Yaish, M., Cox, D. R., 2005. On class differentials in educational attainment. Proceedings of the National Academy of Science 102 (27), 9730–9733.

Fitzmaurice, G. M., Laird, N. M., Ware, J. H., 2004. Applied Longitudinal Analysis. John Wiley & Sons., Inc., Hoboken, NJ.

Fox, J., 1997. Applied Regression Analysis, Linear Models, and Related Methods. Sage, Thousand Oaks.

Ganzeboom, H. B., Treiman, D. J., 2003. Three internationally standardised measures for comparative research on occupational status. In: Hoffmeyer-Zlotnik, J. H., Wolf, C. (Eds.), Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables. Kluwer Academic Press, New York, Ch. 9, pp. 159–193.

Ganzeboom, H. B., Treiman, D. J., 2009. International stratification and mobility file, `http://home.fsw.vu.nl/hbg.ganzeboom/ISMF/index.htm`, last accessed on 7/9/2009.

Halton, J. H., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik 2 (1), 84–90.

Harding, D. J., 2003. Counterfactual models of neighbourhood effects: The effect of neighbourhood poverty on dropping out and teenage pregnancy. The American Journal of Sociology 109 (3), 676–719.

Heckman, J. J., Singer, B., March 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica 52 (2), 271–320.

Holm, A., Jæger, M. M., this issue. Dealing with selection bias in educational transition models: The bivariate probit selection model. Research in Social Stratification and Mobility.

Hout, M., DiPrete, T. A., 2006. What we have learned: RC 28's contributions to knowledge about social stratification. Research in Social Stratification and Mobility 24 (1), 1–20.

Karlson, K. B., this issue. Multiple paths in educational transitions: A multinomial transition model with unobserved heterogeneity. Research in Social Stratification and Mobility.

Kloosterman, R., Ruiter, S., De Graaf, P. M., Kraaykamp, G., June 2009. Parental education, children's performance and the transition to higher secondary education: trends in primary and secondary effects over five dutch school cohorts (1965–99). The British Journal of Sociology 60 (2), 377–398.

Lauer, C., 2003. Family background, cohort and education: A french-German comparison based on a multivariate ordered probit model of educational attainment. Labour Economics 10 (2), 231–251.

Lindsay, B. G., 1983. The geometry of mixture likelihoods: A general theory. The Annals of Statistics 11 (1), 86–94.

Long, J. S., 1997. Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences, 7. Sage Publications, Thousand Oaks.

Lucas, S. R., Fucella, P. N., Berends, M., this issue. A neo-classical education transitions approach: A corrected tale for three cohorts. Research in Social Stratification and Mobility.

Maddala, G. S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Mare, R. D., February 1979. Social background composition and educational growth. Demography 16 (1), 55–71.

Mare, R. D., 1980. Social background and school continuation decisions. Journal of the American Statistical Association 75 (370), 295–305.

Mare, R. D., February 1981. Change and stability in educational stratification. American Sociological Review 46 (1), 72–87.

Mare, R. D., 1993. Educational stratification on observed and unobserved components of family background. In: Shavit, Y., Blossfeld, H.-P. (Eds.), Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder, Ch. 15, pp. 351–376.

Mare, R. D., 1994. Discrete-time bivariate hazards with unobserved heterogeneity: A partially observed contingency table approach. Sociological Methodology 24, 341–383.

Mare, R. D., 2006. Response: Statistical models of educational stratification — hauser and andrew's models for school transitions. Sociological Methodology 11, 27–37.

Mood, C., 2010. Logistic regression: Why we cannot do what we think we can do, and what we can do about it. European Sociological Review 26 (1), 67–82.

Neuhaus, J. M., Jewell, N. P., December 1993. A geometric approach to assess bias due to omited covariates in generalized linear models. Biometrika 80 (4), 807–815.

Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W., 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. International Statistical Review 59 (1), 25–35.

O'Rand, A. M., Henretta, J. C., 1982. Delayed career entry, industrial pension structure, and early retirement in a cohort of unmarried women. American Sociological Review 47 (3), 365–373.

Rijken, S., 1999. Education expansion and status attainment: A cross-national and over-time comparison. Ph.D. thesis, ICS, Utrecht University, Utrecht.

Robert, C. P., Casella, G., 2004. Monte Carlo Statistical Methods, 2nd Edition. Springer texts in statistics. Springer-Verlag, New York.

Rosenbaum, P. R., 2002. Observational Studies, 2nd Edition. Springer, New York.

Rosenbaum, P. R., Rubin, D. B., 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society. Series B 45 (2), 212–218.

Shavit, Y., Blossfeld, H.-P., 1993. Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Westview Press, Boulder.

Sribney, W., Wiggins, V. L., 2010. Standard errors, confidence intervals, and significance tests for ORs, HRs, IRRs, and RRRs, `http://www.stata.com/support/faqs/stat/2deltameth.html`, accessed on August 18th, 2010.

StataCorp, 2009. Stata Statistical Software: Release 11. StataCorp LP, College Station, TX.

Tam, T., this issue. Accounting for dynamic selection bias in educational transitions: The cameron-heckman latent class estimator and its generalizations. Research in Social Stratification and Mobility.

Train, K. E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge.

Tutz, G., 1991. Sequential models in categorical regression. Computational Statistics & Data Analysis 11 (3), 275–295.

Williams, R., 2009. Using heterogenous choice models to compare logit and probit coefficients across groups. Sociological Methods & Research 37 (4), 531–559.