Not all transitions are equal: the relationship between effects on passing steps in a sequential process and effects on the final outcome

Maarten L. Buis
University of Konstanz
Department of History and Sociology
Box 40
78457 Konstanz, Germany
maarten.buis@uni-konstanz.de

Abstract

This article deals with a model for describing a sequence of events, for example: education is typically attained by a set of transitions from one level of education to the next. In particular this article tries to reconcile measures describing the effect of a variable on each of these transitions with measures describing the effect of this variable on the final outcome of that process. Such a relationship has been known to exist within a sequential logit model, but it has hardly been used in empirical research, mainly because of an absence of a practical way of giving it a substantive interpretation. This article tries to provide such an interpretation, by showing that the effect on the final outcome is a weighted sum of the effects on each transition, such that a transition gets more weight if more people are at risk of passing that transition, passing the transition is more differentiating, and people gain more from passing.

1   Introduction

Some processes are best described as a sequence of steps or choices. For example, the choice to have many children is not a single event, but rather a sequence of choices to get yet another child (ignoring twins, triplets, etc.). Similarly, the process of attaining education can be seen as sequence of transitions between one school type and the next. Yet another example would be a process where someone first enters the military or not, and given that one has entered the military, one is exposed to combat or not. In order to describe such a process it is often useful to look at these individual steps or transitions, but also at the end result of that process. The purpose of this article is to present a technique that relates these two. This technique will be based on a sequential model and especially, but not exclusively, a sequential logit model (Fienberg, 1977; Fullerton, 2009)[1]. The sequential logit model estimates the relationship

between explanatory variables and the odds of passing each transition. This model will then be used to also derive the effect of an explanatory variable on the final outcome, which turns out to be a weighted sum of the effects on passing transitions.

This was already established by Mare (1981). However, he framed this result as indicating that models focussing solely on the final outcome — for example, a linear regression model explaining years of education — conflate changes in the marginal distribution of the outcome and changes in effects on the different transitions (Mare, 1981, 83). As a consequence, this result has mainly been understood, at least within the discipline of social stratification research, as an argument against models that try to study effects on the final outcome and not as a possibility to study the relationship between the effects on transitions and effects on the final outcome.

However, Mare did make two attempts at interpreting this relationship. First, by deriving how changes in the transition probabilities influence the effect of an explanatory variable on the final outcome, which turned out to be 'too complex' to be of practical use (Mare, 1981, 78). Consequently, this method has not been picked up at all by the discipline. Second, by comparing a set of scenarios that showed how the effects on the final outcome changed when the marginal distribution changes (Mare, 1981, 82). This method has only been used twice (Smith and Cheung, 1986; Nieuwbeerta and Rijken, 1996). A reason for this may be that it is hard to choose which scenarios to compare: it is hard to find the right balance between scenarios different enough to be interesting but not so different that the amount of extrapolation becomes uncomfortable.

This article tries to add to this literature by showing that one can interpret the relationship between effects on passing transitions and effects on the final outcome. For example, if one studied the influence of family background on educational attainment, then the log odds ratios of passing transitions returned by a sequential logit gives a good description of the influence of family background on the process of attaining education. One could also assign (pseudo-)years of education to each outcome. These are the number of years a 'standard student' would need in order to attain a level of education. For example, if one studied the US educational system one could assign 12 years to someone who only finished high school and 16 for college graduates. Then one can use the results from the sequential logit model to also predict the final outcome, the expected number of pseudo-years of education, and look at how family background influences that outcome. This effect of family background on the highest attained level of education can be decomposed into a weighted sum of the effects of family background on passing transitions. The aim of this article is to show that this decomposition has a substantive interpretation. It makes intuitive sense that the effect on the final outcome is a weighted sum of the effects on each transition. What has been overlooked is that the weights also make substantive sense. In this decomposition a transition receives more weight when:

- more people are at risk of passing that transition,

- passing that transition is more differentiating, that is, the weight is largest when the probability of passing is 50%, which means that neither passing nor failing that transi-

2

tion is virtually universal , and

- people gain more from passing that transition, that is, the difference in expected outcome (in the example above pseudo-years of education) between those that pass and those that fail is larger.

Rather than having separate models for effects on the final outcome and effects on passing different transitions (for example: Shavit and Blossfeld, 1993), this decomposition allows for an integrated discussion of these two related processes. Moreover, this decomposition will allow one to study how changes in the distribution of the outcome variable influence the effect of an explanatory variable on the final outcome. For example, a major development in the educational systems of almost every country in the last century is that more recent cohorts tend to obtain more education (Hout and DiPrete, 2006), a process called educational expansion. This is a change in the marginal distribution, and the method discussed in this article allows one to study the consequences of this development.

The decomposition in this interpretable form exist for the marginal effect for continuous explanatory variable and for discrete changes for categorical explanatory variables. A limitation of this method is that it provides a description of the patterns of association, but these have no causal interpretation.

This article will start with a discussion of the sequential logit model and then derive the decomposition. It will illustrate its use with two examples: the effect of father's education on the offspring's education in the Netherlands and the effect in West-Germany of women's education on the number of children they get.

2 Decomposing the effect on the final outcome

2.1 The sequential logit model

The decomposition of the effect of an explanatory variable on the final outcome into a weighted sum of effects on the odds of passing each transitions starts with a sequential logit model. Consider, a hypothetical process consisting of four states: A, B, C, D, as represented in Figure 1. These states could be educational levels, such that A is no education, B is primary education, C is secondary education, and D is tertiary education. Figure 1 shows how respondents face three transitions: they can attend primary education or opt for no education at all; if they opt for primary education they can choose to leave the system once they have completed primary education, or go on to secondary education; and if they opt for secondary education, they can then either choose to leave once they have completed this level or go on to tertiary education. Alternatively, these states could be the number of children such that A is no children, B is 1 child, C is 2 children, and D is 3 or more children. An implication of this model is that if someone's highest-achieved level is B, then that person was 'at risk' of passing the first two transitions, but not the third. Furthermore, it implies that the person passed the first transition, but failed the second.

The model assumes that one has to be 'at risk' of passing a transition — that is, to have passed through all lower transitions — in order to make a decision at that transition about

Figure 1: Hypothetical process
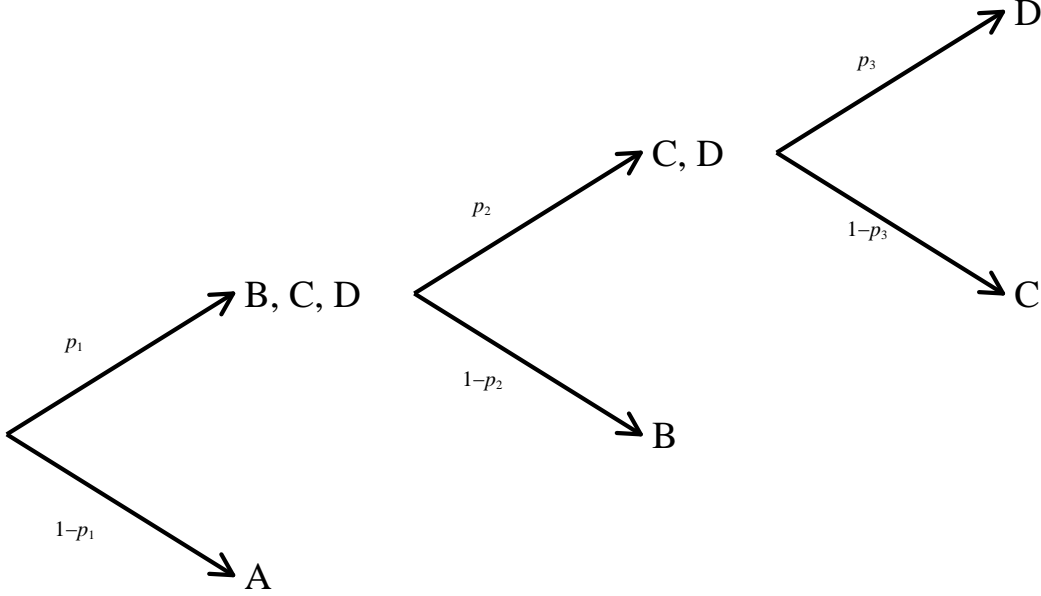
whether to continue or to leave the system. Aside from this, these decisions are assumed to be completely independent. As a result, one can estimate the effects on passing each transition by running separate logistic regressions for each transition on the appropriate sub-sample (Mare, 1980). This model is shown in equation (1).

$$
\begin{aligned}
\Pr(pass_{1,i} = 1 | x_{1i},\, x_{2i}) = \hat{p}_{1i} &= \Lambda(\hat{\lambda}_{01} + \hat{\lambda}_{11}x_{1i} + \hat{\lambda}_{21}x_{2i}) \\
\Pr(pass_{2,i} = 1 | x_{1i},\, x_{2i},\, pass_1 = 1) = \hat{p}_{2i} &= \Lambda(\hat{\lambda}_{02} + \hat{\lambda}_{12}x_{1i} + \hat{\lambda}_{22}x_{2i}) \quad \text{if} \quad pass_{1\,i} = 1 \quad (1) \\
\Pr(pass_{3,i} = 1 | x_{1i},\, x_{2i},\, pass_2 = 1) = \hat{p}_{3i} &= \Lambda(\hat{\lambda}_{03} + \hat{\lambda}_{13}x_{1i} + \hat{\lambda}_{23}x_{2i}) \quad \text{if} \quad pass_{2\,i} = 1
\end{aligned}
$$

The function $\Lambda()$ is the standard logistic function, meaning that $\Lambda(\cdot) = \frac{\exp(\cdot)}{1+\exp(\cdot)}$. The conditional probability that person i passes transition k is $\hat{p}_{ki}$. The association between x_1 and the probability of passing transition k is $\hat{\lambda}_{1k}$, and the constant for transition k is $\hat{\lambda}_{0k}$. Whether or not individual i has passed transition k is indicated by the indicator variable $pass_{k\,i}$. It is assumed that everybody is at risk of passing the first transition.

Cameron and Heckman (1998) have given an important critique of the sequential logit model, in which they question whether the results of a sequential logit model can be given a causal interpretation. They give two reasons for that[2].

First, they argue that the scale of the dependent variable cannot be compared across transitions, and that as a consequence the effects can also not be compared across transitions.

This problem has its origin in the latent propensity representation of the logistic regression model. In this representation of the logistic regression model there is a latent propensity of success which depends linearly on a set of explanatory variables and an error term. A success will occur when the latent propensity is larger than zero, and it is these successes and failures that are actually observed. In order to identify the scale of the latent propensity, and thus the scale of the effects, the variance of the error terms is fixed to a constant ($\pi^2/3$ in case of a logistic regression). However, if the actual variance of the error term differs across groups, this scale for the latent variable will change accordingly. So if we were comparing two groups, and we had reason to believe that the residual variances differ across groups, the scale of the dependent variable would differ and a comparison of effects would be meaningless (Neuhaus and Jewell, 1993; Allison, 1999; Williams, 2009; Mood, 2010). This problem is especially relevant when comparing coefficients in a sequential logit model across transitions, as we have good reason to expect that the error variances will differ across transitions. Preceding transitions will in all likelihood also have selected on variables that are part of the error term, thus reducing the variance of the error term in later transitions (Cameron and Heckman, 1998; Mare, 2006). Notice, however, that this problem only occurs because we assumed that there is a latent variable whose scale we need to identify. There is an alternative representation of the logistic regression model in which it is a non-linear model for the probability of success or a linear model for the log-odds of success. The scale of the dependent variable is now identified: it is respectively the expected proportion of success and the logarithm of the expected number of success per failure. However, both the non-linear probability representation and the latent variable representation lead to exactly the same model, with exactly the same coefficients. How can the same numbers from the same model have simultaneously an identified and unidentified scale? The solution to that paradox is that there is also an unidentified aspect to the probability, odd and log-odds: these measure the degree of (un)certainty that an event will happen, and thus directly depends on the information used and the information not used in estimating the probability; it is the information that was not or could not be used that makes us uncertain. For example, if we estimate the probability of entering university conditional on having completed pre-university education, then that condition is part of the information we used. As a consequence, the probability thus estimated has limited value for predicting what would happen if suddenly many more students became at risk of entering university. In general, one cannot give probabilities, odds, and log-odds a counterfactual interpretation, as they are defined by the model and the sample on which they are estimated. However, this does not preclude the comparison of probabilities, odds, and log-odds in a purely descriptive sense.

The second reason why the results of a sequential logit cannot be given a causal interpretation, is that the selection of students at earlier transitions will result in a correlation between the error term and the observed variables that will influence the estimated coefficients (Mare, 1980; Cameron and Heckman, 1998). For example, one might suspect that mainly high ability children from less privileged backgrounds pass transitions while low ability children from privileged backgrounds also have a fair chance of passing. In that case less privileged children in higher transitions will on average have higher ability than more privileged children in that

same transition. So the selection process created a correlation at higher transitions between family background and ability. If ability was not observed, then the effect of family background would be underestimated at higher transitions. However, if the aim of the study is describing the patterns in the data, then a possible correlation between unobserved variables and observed variables is not relevant. So, one can use sequential logit models in a purely descriptive manner (Mare, 2011; Xie, 2011).

2.2 The decomposition for marginal effects

In order to make a link between the effects[3] of a variable on passing transitions (the $\hat{\lambda}_k$s) and the effect of that variable on the final outcome, it is necessary to first think about how the 'final outcome' can be quantified. For example, if we study educational attainment the outcomes are diplomas. A common way to give each of these diplomas a value is to look at the number of years a 'standard student' would take to achieve that diploma (for example: Korupp et al., 2002). Alternatively, one could use estimated scales of education that are in some way optimal as proposed by Buis (2010, Chapter 3) or Schröder and Ganzeboom (2014). The choice of which values to assign to the outcomes becomes a crucial part of the model, as it defines what the final outcome measures. Once each outcome has a meaningful numerical value (l_k), it becomes possible to use the sequential logit model to calculate the expected value for the final outcome ($E(L_i)$). This is set out in equation (2). The expected value is a weighted sum of the l_ks, whereby the weights are the probabilities of attaining transition k. These probabilities can be derived from the transition probabilities estimated in the sequential logit model using Figure 1.

$$E(L_i) = (1 - \hat{p}_{1i})l_A + \hat{p}_{1i}(1 - \hat{p}_{2i})l_B + \hat{p}_{1i}\hat{p}_{2i}(1 - \hat{p}_{3i})l_C + \hat{p}_{1i}\hat{p}_{2i}\hat{p}_{3i}l_D \qquad (2)$$

For some applications it is hard to find meaningful values for each outcome. For example, what would the values for each outcome be if one studied a process where the first transition consists of whether or not someone ever served in the military and the second transition if they were ever exposed to combat or not (MacLean, 2011)? Similarly, what would the values for each outcome be for a process where potential respondents are first successfully contacted or not, and given that contact was established they decide whether to participate in a study or not (Hoogendoorn and Daalmans, 2009)? Since in both examples the main interest is in one outcome, combat exposure and participation in the survey, a solution is to assign the value 1 to 'combat exposure' and 'study participation' and the value 0 to the outcomes 'no military service' and 'military service without combat exposure' for the first example, and 'unsuccessful contact' and 'contact but refuse to participate' for the second example. That way the expected final outcome is the probability of having experienced combat or participating in a survey and the effect of explanatory variables on the final outcome are the marginal effect on those probabilities. If one wants to apply this decomposition to a research project in which all outcomes are of interest, one can repeat the decomposition such that the value 1 is assigned to a different outcome in each repetition.

The explanatory variables are part of equation (2) through the $\hat{p}_{ki}$s as described in equa-

tion (1). Equation (2) can be understood as a regression equation showing a non-linear relationship between these explanatory variables and the final outcome. In fact, if one plots the 'regression line' from equation (2) together with a regression line from a normal linear regression, one will generally find that the two are very close. The difference between the two lines comes from the fact that the functional form of implied by equation (2) ensures the predicted outcome cannot be less than the lowest level or more than the highest levels, while linear regression does not impose this constraint. Section 3.1 contains an example of a situation where equation (2) would deviate from a normal linear regression. In the earliest cohort studied, around 1910, a large portion of the Dutch population had only primary education. The fact that equation (2) respects that lower bound leads to noticeable non-linearity.

One way to quantify the effect of an explanatory variable on the final outcome is the first derivative of equation (2) with respect to this explanatory variable. This derivative is shown in equation (3). A step-by-step derivation is set out in Appendix A.

$$
\begin{aligned}
\frac{\partial E(L_i)}{\partial x_j} \quad = \\
\{1 \quad &\times \quad \hat{p}_{1i}(1-\hat{p}_{1i}) \quad \times \quad [(1-\hat{p}_{2i})l_B + \hat{p}_{2i}(1-\hat{p}_{3i})l_C + \\
&\qquad\qquad\qquad\qquad\quad \hat{p}_{2i}\hat{p}_{3i}l_D - l_A] \qquad\qquad\qquad \} \quad \hat{\lambda}_{j1} \quad + \\
\{\hat{p}_{1i} \quad &\times \quad \hat{p}_{2i}(1-\hat{p}_{2i}) \quad \times \quad [(1-\hat{p}_{3i})l_C + \hat{p}_{3i}l_D - l_B] \qquad \} \quad \hat{\lambda}_{j2} \quad + \\
\{\hat{p}_{1i}\hat{p}_{2i} \quad &\times \quad \hat{p}_{3i}(1-\hat{p}_{3i}) \quad \times \quad [(l_D - l_C)] \qquad\qquad\qquad\quad \} \quad \hat{\lambda}_{j3}
\end{aligned}
\tag{3}
$$

Equation (3) shows that the effect on the final outcome $\left(\frac{\partial E(L_i)}{\partial x_j}\right)$ is a weighted sum of the effects on each transition (the $\hat{\lambda}_{jk}$s). The weights (the sections between curly brackets) consist of three parts:

1. The predicted proportion of people at risk of passing a transition. For the first transition, this proportion is 1; for the second it is the proportion of respondents who complete the first transition, $\hat{p}_{1i}$; and for the third transition, it is the proportion who completed the second transition, $\hat{p}_{1i}\hat{p}_{2i}$. Substantively, this means that a transition is more important when more people are at risk of passing it.

2. The differentiating capacity of the kth transition, as captured by the function $\hat{p}_{ki}(1-\hat{p}_{ki})$. This function is lowest if virtually everybody passes or fails, and is highest when the probability of passing is .5. This makes sense at a substantive level, because if only a few people pass or fail a transition, any effect of the explanatory variable at this stage will only affect a few people.

3. The differences between the expected outcome of those who pass the transitions and those who do not. These are the parts in the square brackets. For instance, the expected outcome of those who pass the first transition is $(1 - \hat{p}_{2i})l_B + \hat{p}_{2i}(1 - \hat{p}_{3i})l_C + \hat{p}_{2i}\hat{p}_{3i}l_D$ and the expected outcome for those that fail the first transition is l_A. The difference between the two is the expected gain from passing the first transition. The substantive interpretation of this is that a transition becomes more important if passing leads to a greater expected increase in the final outcome.

As discussed in Appendix A, equation (3) can also be written as equation (4), which can be a useful variant. The marginal effect of x_j on the final outcome is a weighted sum

of the marginal effects of x_j on passing the individual transitions, and this weight consists of two of the three elements in equation (3): the proportion at risk and the expected gain from passing that transition. One advantage of this version of the decomposition is that it is true regardless of the link function used, so it also applies to seqential probits, sequential cloglogs, etc. Moreover, many people find the marginal effect on passing a transition used in equation (4) easier to interpret than the log odds ratio of passing a transition used in equation (3). However, this marginal effect is usually[4] also a function of the proportion passing that transition, so this decomposition no longer clearly separates the distribution from the effect. If this separation is important, then that would be an argument for using the decomposition in equation (3) instead. Section 3.1 discusses an example where this separation is important.

$$
\begin{aligned}
\frac{\partial E(L_i)}{\partial x_j} \quad = \quad & \\
\{1 \quad & \times \quad [(1-\hat{p}_{2i})l_B + \hat{p}_{2i}(1-\hat{p}_{3i})l_C + \hat{p}_{2i}\hat{p}_{3i}l_D - l_A] \quad \} \quad \frac{\partial \hat{p}_1}{\partial x_j} \quad + \\
\{\hat{p}_{1i} \quad & \times \quad [(1-\hat{p}_{3i})l_C + \hat{p}_{3i}l_D - l_B] \quad \} \quad \frac{\partial \hat{p}_2}{\partial x_j} \quad + \\
\{\hat{p}_{1i}\hat{p}_{2i} \quad & \times \quad [(l_D - l_C)] \quad \} \quad \frac{\partial \hat{p}_3}{\partial x_j}
\end{aligned}
\tag{4}
$$

What is being decomposed in both decompositions is the marginal effect of an explanatory variable on the expected value of the final outcome at typical values of all explanatory variables. One limitation of this marginal effect is that if the relationship between the explanatory variable and the expected outcome is non-linear, then each observation has its own marginal effect depending on where on the curve each observation is located. This limitations becomes more pressing the more pronounced the degree of non-linearity is. The example in section 3.1 illustrates this point. Moreover, what is typical at the first transition may not be typical at later transitions. For example, a father with a skilled manual job may be typical during the first transition in an educational system, but is probably less typical among the students at risk of entering university. For both these reasons, the marginal effect at typical values of the explanatory variables is not necessarily the best summary measure of the effect. To deal with this problem some have proposed to look at average marginal effects — that is, compute the marginal effect for each observation and average those — as a way of summarizing the effect. This averaged effect of an explanatory variable on the expected final outcome can also be decomposed into a weighted sum of effects on each transitions, where the weights are now the average of the weights predicted for each individual. However, the point of the decomposition is that the weights also have a substantive interpretation as the product of three substantively interesting components. This point is lost by averaging the weights, as the average of a product of three components is not the same as the product of the averages of these three components.

2.3 The decomposition for discrete changes[5]

Another potential problem with marginal effects is that in non-linear relationships the marginal effect is somewhat artificial: it is the effect of a unit change if the rate of change remains unchanged, but the whole point of a non-linear model is that the rate of change is not constant. This is especially problematic when the explanatory variable is categorical, as the instanta-

neous rate of change (the marginal effect) is not defined for such a variable. To deal with this problem some (for example: Long, 1997) have proposed looking at discrete changes — that is, looking at an actual unit change instead of marginal effects. One can derive a decomposition of the effects of a unit change in a similar way as the marginal effect. Lets call the discrete change in the final outcome $\Delta \hat{L}_i$. Continuing the example above, but now assuming that the explanatory variable of interest x is binary with values 0 and 1, and using equation (2), $\Delta \hat{L}_i$ can be written as:

$$
\begin{aligned}
\Delta \hat{L}_i &= \mathrm{E}(L_i|x=1) - \mathrm{E}(L_i|x=0) \\
&= \left[(1-\hat{p}_{1i}^1)l_A + \hat{p}_{1i}^1(1-\hat{p}_{2i}^1)l_B + \hat{p}_{1i}^1\hat{p}_{2i}^1(1-\hat{p}_{3i}^1)l_C + \hat{p}_{1i}^1\hat{p}_{2i}^1\hat{p}_{3i}^1 l_D \right] - \\
&\quad \left[(1-\hat{p}_{1i}^0)l_A + \hat{p}_{1i}^0(1-\hat{p}_{2i}^0)l_B + \hat{p}_{1i}^0\hat{p}_{2i}^0(1-\hat{p}_{3i}^0)l_C + \hat{p}_{1i}^0\hat{p}_{2i}^0\hat{p}_{3i}^0 l_D \right]
\end{aligned} \tag{5}
$$

Where $\hat{p}_{mi}^n$ is short for the predicted probability of person i of passing the mth transition when the explanatory variable of interest is n. Equation (5) can be rewritten as equation (6). A step-by-step derivation of this formula can be found in Appendix B.

$$
\begin{aligned}
\Delta \hat{L}_i = \;& \{ \quad 1 \quad \times \quad [(1-\hat{p}_{2i}^*)l_B + \hat{p}_{2i}^*(1-\hat{p}_{3i}^*)l_C + \hat{p}_{2i}^*\hat{p}_{3i}^* l_D - l_A] \quad \} \quad \Delta \hat{p}_{1i} \quad + \\
& \{ \quad \hat{p}_{1i}^* \quad \times \quad [(1-\hat{p}_{3i}^*)l_C + \hat{p}_{3i}^* l_D - l_B] \quad \} \quad \Delta \hat{p}_{2i} \quad + \\
& \{ \quad \hat{p}_{1,2i}^* \quad \times \quad [(l_D - l_C)] \quad \} \quad \Delta \hat{p}_{3i}
\end{aligned} \tag{6}
$$

Where $\hat{p}_{mi}^* = \frac{\hat{p}_{mi}^1 + \hat{p}_{mi}^0}{2}$, that is, the average probability of passing transition m if categories 1 and 0 are equally common. Similarly, $\hat{p}_{m,ni}^* = \frac{\hat{p}_{mi}^1\hat{p}_{ni}^1 + \hat{p}_{mi}^0\hat{p}_{ni}^0}{2}$, that is, the average probability of passing both transitions m and n if categories 1 and 0 are equally common. Analogous to the notation used for the discrete change in the final outcome, the discrete change in the probability of passing transition m is written as $\Delta \hat{p}_{mi}$, which is $\hat{p}_{mi}^1 - \hat{p}_{mi}^0$. A similar decomposition was also derived by Mare (1977, chapter 5).

The structure of equation (6) is very similar to equation (4). As a consequence, the interpretation is also similar: The discrete change in final outcome is a weighted sum of discrete changes in each transition, such that a transition receives more weight when more people are at risk of passing that transition and the difference in expected outcome between those that pass and those that fail that transition is larger. Notice however that the proportions used in this interpretation are rather artificial: they assume that during each transition the number of observations in each category of the explanatory variable is equal. Even if, by accident, this happens to be (almost) true for the first transition, selection will make sure that this is no longer the case in later transitions.

3 Examples

The purpose of the following two examples is twofold: On the one hand, the examples will be used to illustrate how this decomposition can be used. On the other hand, they will be used to show some extensions to this decomposition. The first example will be used to show how one can apply this decomposition to more complex decision trees and emphasize the way this

method can be used to study the effect of changes in the distribution of the final outcome. The second example illustrates a special case where the decomposition of discrete differences is non-parametric.

3.1 Educational expansion and the effect of father's occupation on the son's education

The second example considers the influence of family background on educational attainment in the Netherlands and how this process changed over cohorts. This is an application with a more complex decisions tree, as the Dutch education system is a tracked system. So rather than a hierarchical set of 'continue or stop' decision, this decision tree will contain a branching point, as is illustrated in Figure 2. It is also an application where changes in weights are of substantive interest. Changes in weights are driven by changes in the probabilities of passing the different transitions. As in most countries (Hout and DiPrete, 2006), these probabilities have changed considerably over cohorts in the sense that more recent cohorts are much more likely to attain higher levels of education than older cohorts; this is illustrated in Figure 3. This process is called educational expansion. Much of the effort in studying these educational systems is aimed at getting estimates of effects of parental background on educational attainment while controlling for educational expansion (Mare, 1981; Breen and Jonsson, 2005). However, since educational expansion is such a common and far reaching change in most educational systems, one might want to know how educational expansion has changed the influence of family background on educational attainment. The decomposition proposed in this article provides a useful way to answer such a question (for example: Ballarino et al., 2014).[6] This example will only focus on the sons. Substantively, comparing the decomposition of male and female respondents would be interesting, as the timing of educational expansion was different for men and women. However, the mechanics of the decomposition, which is what this example is supposed to illustrate, would be the same for both men and women.

The simplified representation of the Dutch educational system in Figure 2 assumes that all respondents complete primary education. After this, they face a choice between leaving the schooling system and continuing[7]. If they opt for the latter choice, they have to choose between the higher secondary and lower secondary school. Once they have finished their second diploma in either track, they can choose whether or not to get a third diploma, continuing with: vocational education if they are in the low track, or tertiary education if they are in the high track. The values assigned to each level of education are estimated in such a way that the effect of education on income is maximized while controlling for the father's occupational status, as described in (Buis, 2010, Chapter 3)[8]. The values are standardized to have a mean of 0 and standard deviaiton of of 1 for the cohort that was 12 in 1950. These values together with the Dutch names for the different educational levels are presented in Table 1.

Sequential logit models can also be used for tracked systems like the one in Figure 2 (Lucas, 2001; Breen and Jonsson, 2000). Appendix C shows that the decomposition discussed in this article also applies to such tracked systems and it has exactly the same interpretation. Notice that this simplified representation of the Dutch education system reduces it to a system of

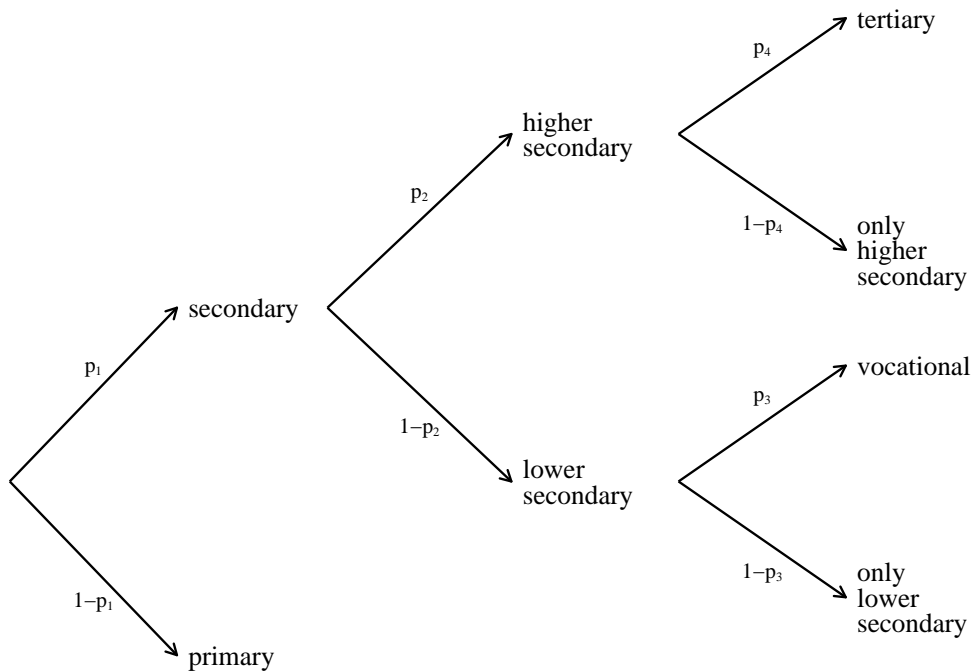Figure 2: Simplified model of the Dutch education system

Table 1: Description of the Dutch education system

Dutch names	English description	value
LO	primary	$l_A = -1.20$
LBO or MAVO	lower secondary	$l_B = -0.468$
HAVO or VWO	higher secondary	$l_C = 0.479$
MBO	vocational	$l_D = -0.031$
HBO or WO	Tertiary	$l_E = 1.72$

binary choices. This is convenient, and proved to be enough to reveal the main interesting trends, but it is not necessary. As discussed in Appendix A.1, the decomposition can also be computed for systems with multinomial choices. Assuming one has data on actual transitions made by respondents instead of only the highest achieved level education used in this example, one can extend this model and the decomposition further to account for the possibility that some respondents changed track and that there is thus more than one way in which one can obtain a given diploma.

In this example it makes sense to represent the effects on transitions as log odds ratios so one can more clearly separate the influence of educational expansion from other changes. Educational expansion is captured in a sequential logit model by changes in the baseline odds of passing a transition. An odds ratio controls for changing baseline odds by being a relative effect. So the argument is that a doubling of the odds is considered to be the same effect regardless how large the baseline odds is, just as two percentages are sometimes compared using the argument that a hundredth of some total is comparable to a hundredth of some other total.[9] This is how changes in odds ratios capture changes net of educational expansion, while the weights represent the influence of educational expansion.

The data were obtained from the International Stratification and Mobility File (ISMF) (Ganzeboom and Treiman, 2009). The ISMF contains 54 surveys[10] on the Netherland, carried out between 1958 and 2006. The surveys were post-harmonized and merged to increase the time period covered and the number of respondents, and to lessen the effect of individual surveys' idiosyncrasies. Time was measured by the year in which the respondent was 12, scaled in decades since 1950. Information is available for the cohorts born between 1905 and 1991. Cohort is allowed to have a non-linear effect by representing it as a cubic B-spline parameterized such that the coefficients of each variable represents the difference (in this case ratio) between its year and a reference year (Newson, 2012). So, these parameters can be interpreted in a similar way as a set of indicator or dummy variables, except that these points are linked by a smooth curve rather than a step-function. The knots are 1910, 1930, 1950, 1970, and 1990 and the reference year is 1950. A family's SES was measured according to the father's score on the International Socio-Economic Index (ISEI) of occupational status (Ganzeboom and Treiman, 2003), as information on the father's occupation was available for the largest number of cohorts. The original ISEI score is a continuous variable ranging from 10 to 90, but it was standardized to have the value 0 for those fathers who are skilled manual workers (close to the mean in 1950) and a standard deviation of 1 for the cohort that were 12 in 1950. The effect of father's occupational status was allowed to change linearly with cohort, by adding an interaction term. Survey weights were used where available. After removing respondents with missing observations on any of the variables, 43,768 male respondents remained.

The distribution of the highest achieved level of education for different cohorts is shown in Figure 3. The proportion of pupils who only achieved primary education dropped dramatically, while the proportion attaining tertiary education and vocational education strongly increased. Figure 3 also shows that vocational education is a relatively recent level of education for the Netherlands. Whereas no one from the earlier cohorts completed this level of education, vocational education completion has rapidly grown to about 40%.

Figure 3: Distribution of highest achieved level of education for men over cohorts

The sequential logit model is presented in table 2. The main effects of the father's occupational status (fisei) represent the effect for the cohort that was 12 in 1950. Thus, for men from that cohort a standard deviation increase in father's occupational status is associated with an 88% increase in the odds of passing the first and second transition, a 29% increase in the odds of continuing within the lower track, and a 36% increase in the odds of continuing in the higher track. This effect of father's occupational status has significantly declined by 6.4% per decade for the first transition, 1.7% per decade for the second transition and 3.3% per decade for the last transition within the upper track. The trend in the effect of father's occupational status was non-significant in the final transition within the lower track. Educational expansion is captured by the baseline odds plus the four cohort variables. So for men who were 12 in 1950 and whose father was a skilled manual worker one would expect to find 5.04 individuals that passed the first transition for each individual that failed the first transition. For similar men from the 1910 cohort this odds would only be a tenth of the odds of the 1950 cohort, that is, one would expect to find only half a person that passed this transition for every person that failed it. These estimates show a picture similar to Figure 3.

With these estimates and the values that were attached to each level of education one can compute the expected level of education for people with different father's occupational status and born in different years. These are shown in Figure 4. The effect of father's occupational status on the highest achieved level of education are the slopes of these curves. As was discussed in section 2.2, the slope will differ depending on at which value of father's

Table 2: Sequential logit model for men

	primary v. secondary	lower v. higher secondary	lower secondary v. vocational	higher secondary v. tertiary
fisei	1.88	1.88	1.29	1.36
	(31.2)	(38.5)	(9.87)	(12.4)
fisei×cohort[a]	.936	.983	.987	.967
	(-4.94)	(-1.86)	(-.886)	(-2.36)
cohort 1910[b]	.104	.247	.000242	.677
	(-10.9)	(-3.78)	(-5.88)	(-.582)
cohort 1930[b]	.326	.588	.3	.395
	(-21.2)	(-9.06)	(-10.8)	(-9.45)
cohort 1970[b]	3.26	1.05	1.98	1.22
	(24.1)	(1.57)	(15.2)	(3.73)
cohort 1990[b]	7.75	1.23	6.39	2.25
	(10.2)	(2.32)	(14.1)	(4.88)
constant	5.04	0.674	0.361	2.42
	(59.0)	(-17.8)	(-32.6)	(23.9)
N	43762			
log likelihood	-49986.0			

Exponentiated coefficients; z statistics in parentheses

[a] Cohort is entered linearly in the interaction

[b] Reference splines, parameters can be interpreted as dummy variables with 1950 as a reference

occupational status the slope is computed because these curves are non-linear. For example, I chose the value 0 to be 'typical' (the occupational status of a skilled manual worker, which is close to the mean for the cohort that was 12 in 1950), so it makes sense to look at the slopes at that value. At that value the slopes initially become steeper from 1910 to 1930 and then become flatter after that. If we look instead at the value 2 standard deviations above 0 (the occupational status of a teacher at a secondary school), we don't see that initial increase in the effect of father's occupational status. Instead the slope is steadily declining over time. Mechanically, this difference is due to the fact that for the oldest cohort the lower bound forces the slope to flatter at a father's occupational status of 0 compared with a father's occupational status of 2. This also makes substantive sense. Around 1910 a large portion of the population were basically equal because they all had little education. As a consequence, if the father got a bit more occupational status, then that did not do much when he started as a skilled manual worker. However, a small increase in father's occupational status was more consequential when the father started as a teacher, as all educational outcomes were for those children more or less realistic options. So it is important to carefully consider at which point you evaluate the effects.

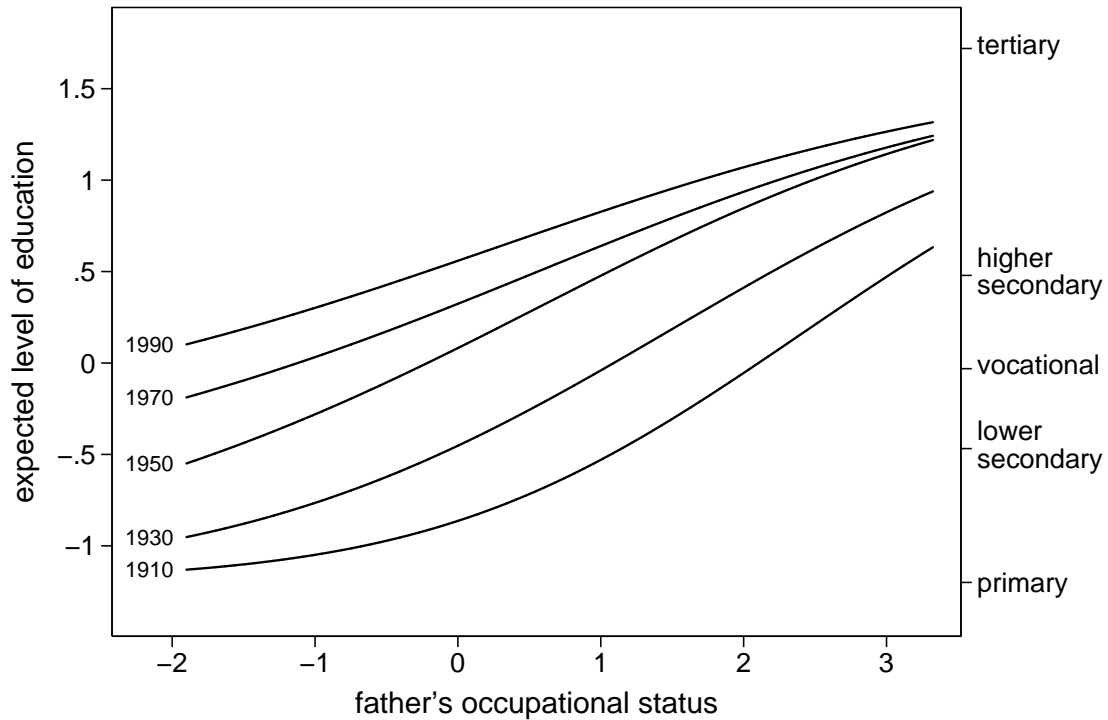Figure 4: Expected highest achieved level of education (years refer to the year in which the respondent was 12)

Figure 5 shows the effects of father's occupational status for sons of skilled manual workers. Both highest achieved level of education and the father's occupational status are scaled in such a way that the standard deviation is 1. So this measure of the total effect is similar to

15

a standardized regression coefficient. Figure 5 confirms the pattern found in Figure 4: The effect of father's occupational status initially increases, till around 1945 and then decreased[11].

The first step in explaining this pattern is to investigate the contributions of each transition to the total effect. A striking feature is that the final two transitions (continuing in the lower track and continuing in the higher track) contribute negligible amounts to the total effect. Moreover, there has been a shift between the first and the second transitions as to the dominant source of the total effect. The contribution of the first transition initially increased and then decreased, while the contribution of the second transition increased throughout the period being studied. The initial increase in the contributions of the first and second transition more or less coincide, thus reinforcing one another. The later increase in the contribution of the second transition was not enough to compensate for the decrease in the contribution of the first transition, thus leading to the decrease in the effect of father's occupational status on the highest achieved level of education.

Figure 5: Contribution of different transitions to the effect of father's occupation
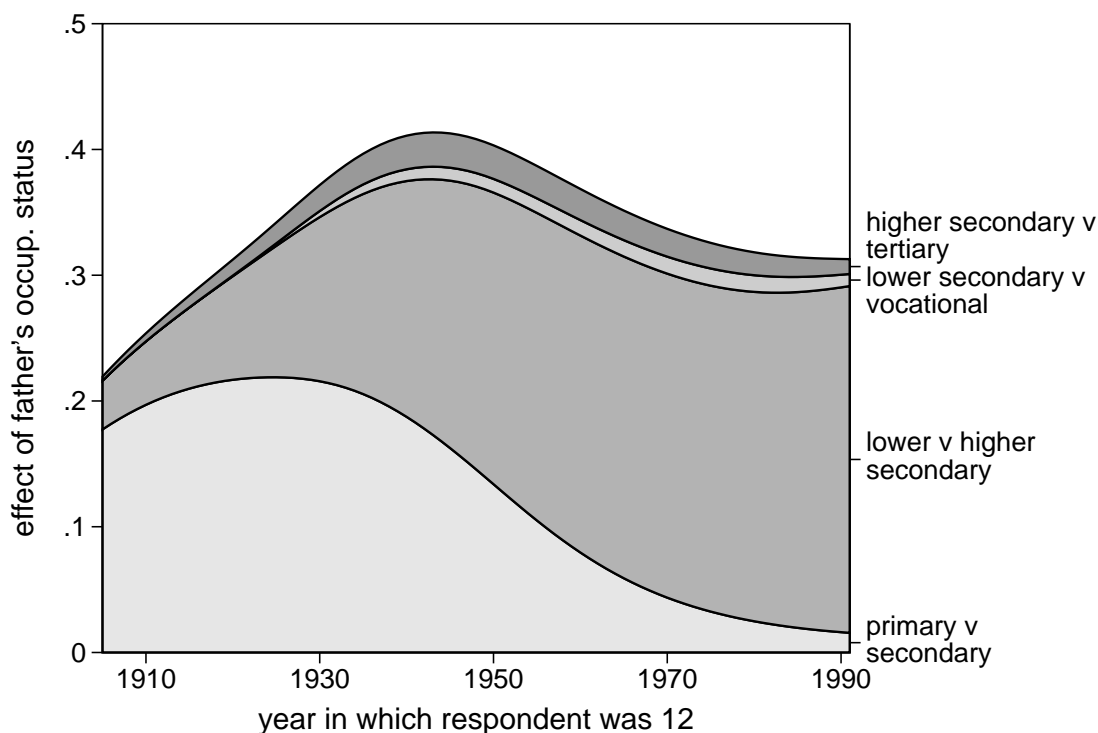
The second step of this explanation consists of breaking up each transition's contribution into its two parts: the weight and the effect on each transition. Since the contribution is the product of these two terms, it can be visualized as the area of a rectangle, with a height equal to the effect and a width equal to the weight. This is shown in Figure 6. The horizontal axis shows the weights and the vertical axis the effects, while the columns represent the cohorts and the rows represent the transitions. These figures show that the initial increase in the contribution of the first transition is due to an increase in its weight, while the later decrease

of this transition is due to both a decrease in the weight and a decrease in the effects on the transitions. The increase in importance of the second transition is entirely due to the increase in the weight of this transition. This increase in weight actually more than offsets a decrease in the effects. The low contributions of both higher transitions are due to both low effects and low weight.

Figure 6: Decomposition of the effect of father's occupational status for men into effects on passing transitions and weights
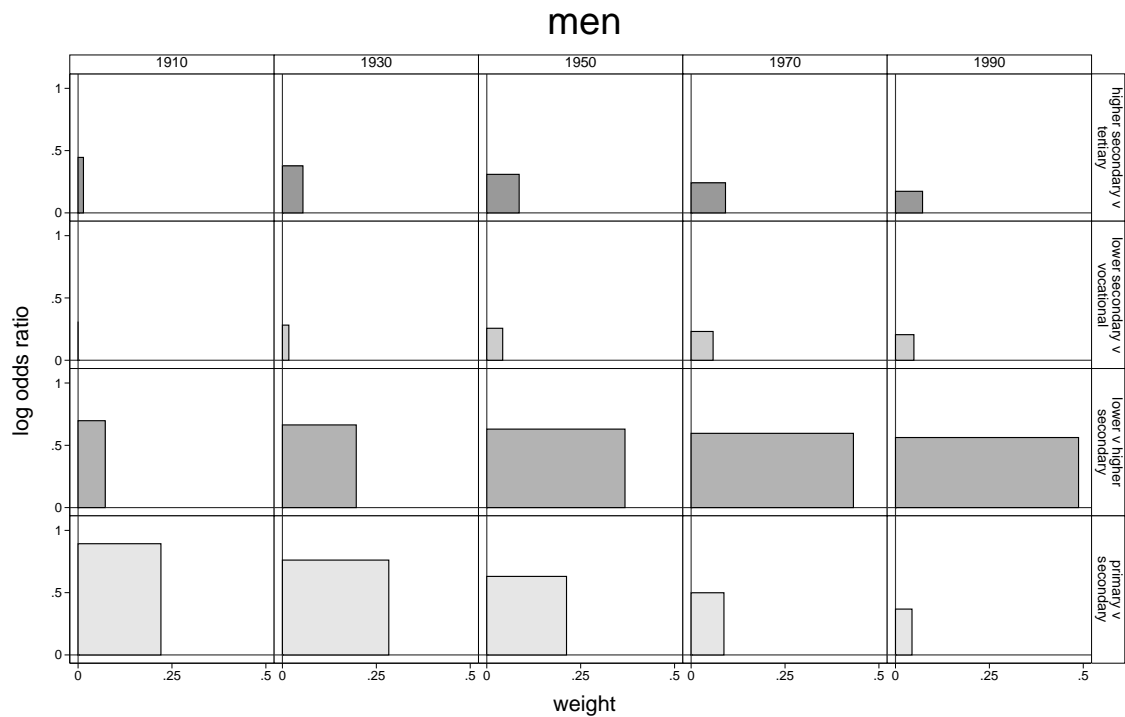
The third step breaks the weights down into three components, which is done in Table 3. This table shows that the initial increase and the later decline in the first transition's influence is primarily due to the differentiating capacity of that transition. Initially, any inequality at the first transition affected few people, because only a low proportion of students passed. As the proportion of students passing increased, the transition received more weight, until half of the students passed, after which inequality at this transition affected less people again because only a few students failed. The increase in importance of the second transition is partly due to the differentiating capacity, but also to a strong increase in the number of students that are at risk of making this transition. The last two transitions receive relatively small weights because relatively few people are at risk of passing these transitions, and those who pass gain relatively little. Those who pass the first two transitions gain both the immediate increase in level of education and the possibility of gaining an extra level of education, while in the third and fourth transition, people gain only the immediate increase in level of education. These developments provide a mechanism through which educational expansion influences the effect

of father's occupational status on highest achieved level of education.

It is informative to go through all the elements of the first cohort in Table 3 in order to illustrate how the decomposition works. One can see that the effect on the final outcome is the weighted sum of the effects on passing transitions: $0.893 \times 0.221 + 0.698 \times 0.072 + 0.308 \times 0.000 + 0.446 \times 0.014 = 0.254$. The effects on passing transitions can be computed from Table 2: $\ln(1.88 \times 0.936^{-4}) \approx 0.893$, $\ln(1.88 \times 0.983^{-4}) \approx 0.698$, $\ln(1.29 \times 0.987^{-4}) \approx 0.308$, and $\ln(1.36 \times 0.967^{-4}) = 0.446$ (the differences are due to rounding error that happened because only three digits are reported in the tables). The weights are the products of the proportion at risk, the degree of differentiation, and the gain of passing: $1 \times .0.226 \times 0.977 = 0.221$, $0.344 \times 0.122 \times 1.718 = 0.072$, $0.295 \times 0 \times 0.437 = 0$ and $0.049 \times 0.235 \times 1.241 = 0.014$. The proportions at risk follow directly from the probabilities of passing transitions (the $\hat{p}$s): The proportion of respondents at risk in the first transition is 1, the proportion at risk of passing the second transition is $\hat{p}_1 = 0.344$, the proportion at risk of passing the third transitions is $\hat{p}_1 \times (1 - \hat{p}_2) = 0.344 \times (1 - 0.143) = 0.295$, and the proportion at risk of passing the fourth transition is $\hat{p}_1 \times \hat{p}_2 = 0.344 \times 0.143 = 0.049$. The degrees of differentiation are $0.344 \times (1 - 0.344) = 0.226$, $0.143 \times (1 - 0.143) = 0.122$, $0 \times (1 - 0) = 0$, and $0.621 \times (1 - 0.621) = 0.235$. The gain from passing the different transitions is $(1 - 0.143) \times (1 - 0) \times -0.468 + 0.143 \times (1 - 0.621) \times 0.479 + 0.143 \times 0 \times -0.031 + 0.143 \times 0.621 \times 1.72 - -1.20 \approx 0.977$, $0.621 \times 1.72 + (1 - 0.621) \times 0.479 - 0 \times -0.031 - (1 - 0) \times -0.468 = 1.718$, $-0.031 - -0.468 = 0.437$, and $1.72 - 0.479 = 1.241$

3.2 Women's education and the probability of having a large family
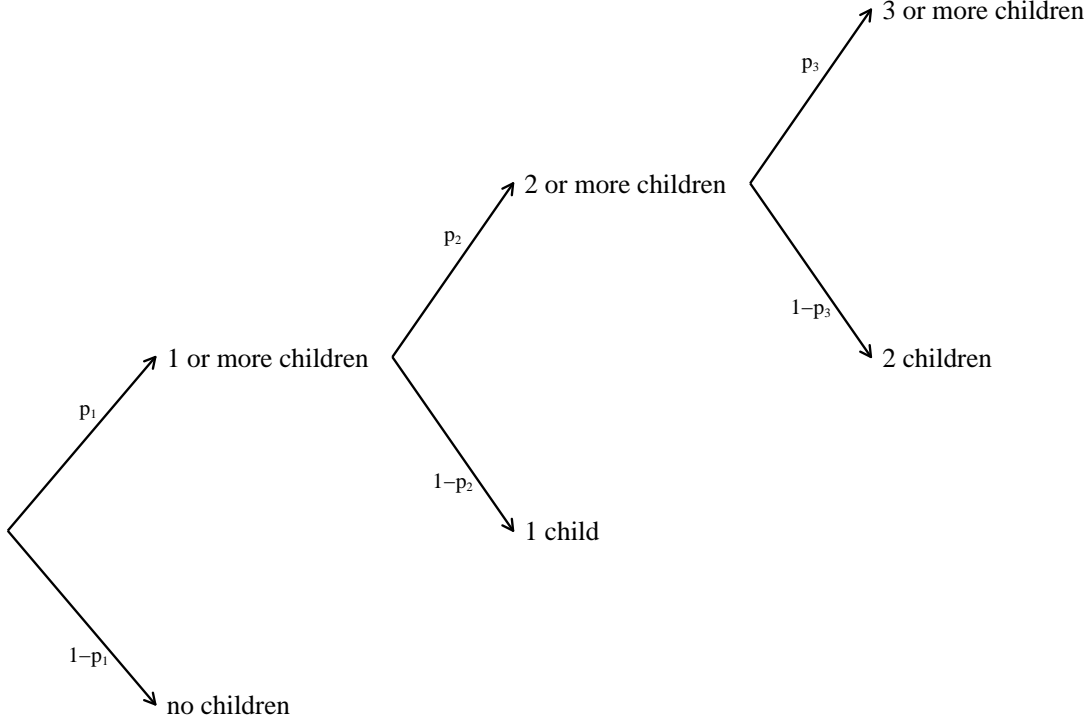
In this example I will study the influence of women's education on the process leading to 'large' families for women born between 1934 and 1936 in West-Germany. The main purpose of this example is to illustrate the decomposition in terms of discrete differences. For illustrative purposes it is convenient to have a fairly small decision tree. So in this example the decision tree stops at '3 or more children' rather than continue for all the children a women could have. One can imagine a research project in which the interest is on 'large' families rather than the number of children *per se*. For this cohort in West-Germany the modal number of children is clearly 2, so I will consider a family with 3 or more children as large. The cohorts are interesting because they are to a large extend responsible for the 'baby-boom' in Germany (Buis et al., 2012). The process is described in Figure 7.

The next step is to assign values to each outcome. One solution would be to not censor the number of children at 3 or more, but estimate a model for the complete distribution of completed fertility. In that case it would make sense to assign the value 0 to no children, 1 to 1 child, etc. The final outcome would in that case be the expected number of children. However, since the interest is in large families rather than the number of children, I assign the value 1 to '3 or more children' and the value 0 to all other outcomes. That way the expected final outcome is the probability of having a large family and the effect of mother's education on the final outcome is the marginal effect on the probability of having a large family. By choosing these values the decomposition in equation (6) simplifies to equation (7).

Table 3: Detailed decomposition of the effect of father' occupational status on men's educational attainment

year		primary v. secondary	lower v. higher secondary	lower secondary v. vocational	higher secondary v. tertiary
1910	effects on transitions	0.893	0.698	0.308	0.446
	weight	0.221	0.072	0.000	0.014
	at risk	1.000	0.344	0.295	0.049
	differentiate	0.226	0.122	0.000	0.235
	gain	0.977	1.718	0.437	1.241
	$\hat{p}$	0.344	0.143	0.000	0.621
	total effect	0.254			
1930	effects on transitions	0.762	0.664	0.282	0.378
	weight	0.283	0.197	0.017	0.055
	at risk	1.000	0.621	0.445	0.176
	differentiate	0.235	0.203	0.088	0.250
	gain	1.204	1.557	0.437	1.241
	$\hat{p}$	0.621	0.284	0.098	0.489
	total effect	0.368			
1950	effects on transitions	0.630	0.630	0.257	0.310
	weight	0.212	0.368	0.042	0.086
	at risk	1.000	0.834	0.498	0.336
	differentiate	0.138	0.241	0.195	0.207
	gain	1.536	1.834	0.437	1.241
	$\hat{p}$	0.834	0.403	0.265	0.708
	total effect	0.388			
1970	effects on transitions	0.499	0.596	0.231	0.241
	weight	0.087	0.432	0.059	0.091
	at risk	1.000	0.943	0.551	0.391
	differentiate	0.054	0.243	0.243	0.188
	gain	1.617	1.888	0.437	1.241
	$\hat{p}$	0.943	0.415	0.416	0.748
	total effect	0.310			
1990	effects on transitions	0.368	0.562	0.205	0.173
	weight	0.044	0.487	0.049	0.072
	at risk	1.000	0.975	0.532	0.443
	differentiate	0.024	0.248	0.211	0.131
	gain	1.804	2.017	0.437	1.241
	$\hat{p}$	0.975	0.454	0.697	0.845
	total effect	0.268			

Figure 7: Process leading to 'big' families (3 or more children)

$$\Delta \hat{L}_i = \begin{array}{llll} \{ & 1 & \times & [\hat{p}_{2i}^* \hat{p}_{3i}^*] & \} & \Delta \hat{p}_{1i} & + \\ \{ & \hat{p}_{1i}^* & \times & [\hat{p}_{3i}^*] & \} & \Delta \hat{p}_{2i} & + \\ \{ & \hat{p}_{1,2i}^* & \times & [1] & \} & \Delta \hat{p}_{3i} & \end{array} \tag{7}$$

The data are from the scientific use file of the Mikrozensus 2008. It contains 6,261 women born between 1934 and 1936 with valid information on the number of children they received and their education. Education is measured by 3 educational categories: basic education, which I define as only general lower secondary education (*Hauptschule* or *Realschule*) or less, lower secondary education plus vocational education, and higher general secondary or more (*Abitur* or *Abitur* followed by vocational education or tertiary education.). The distribution of the number of children born to women conditional on their education is described in Table 4. In this table, one can see that women in this cohort were not well educated; 46% of the women end up with only basic education. One can also see that having a large family is not uncommon as 32% of the women in this cohort got three or more children. Finally, one can see that higher educated women tended to get fewer children. It is this pattern that this example wants to describe in more detail.

One could estimate a sequential logit model using this data, and then compute the decomposition. However, that is in this case not necessary. Notice, that the decomposition only depends on the predicted probabilities of passing transitions and that the model would be fully saturated. A characteristic of fully saturated models is that predictions from that model

Table 4: Distribution of number of children conditional on mother's education

	no children	1 child	2 children	3 or more children	Total
	(row %)				(N)
basic	8.7	19.2	33.9	38.2	2,891
vocational	14.4	24.3	35.5	25.9	2,986
Abitur or more	21.9	21.9	30.7	25.5	384
Total (row %)	12.2	21.8	34.4	31.6	100.0
(N)	764	1,365	2,156	1,976	6,261

do not differ from predictions one could directly compute from Table 4. For example the probability of passing the first transition, that is getting one or more children, for a women with basic education is $19.2 + 33.9 + 38.2 = 91.3\%$. If we first estimated a sequential logit (or sequential probit, or a sequential cloglog, or any other sequential model), followed by predicting the transition probabilities, then we would have gotten the exact same number. In this sense the decomposition of discrete differences in a fully saturated model is non-parameteric. If the model is no longer saturated, for example by adding a control variable either linearly or without adding all the interactions, then the decomposition would differ depending on which link function was used, and the decomposition would thus no longer be non-parametric.

The decomposition of discrete changes is shown in Table 5. Women with vocational education tend to have a 12.3% lower chance of getting a large family compared to women with basic education. All transitions contribute to this negative effect but the largest contribution comes from third transition, which has both the largest effect and the largest weight. The weight for the final transition is highest as the gain is high — everybody who passes the third transition by definition has a large family — and there is still a substantial proportion of women at risk of passing that transition. For the first transition everybody is at risk, but only a small proportion who pass the first transition actually get a large family.

Women with Abitur tend to have only 0.4% lower chance of getting a large family compared to women with vocational education. This small effect is the result of two opposing effects: women with an Abitur are less likely to pass the first transition, that is, they are more likely to remain childless compared to women with vocational education. However, once a woman with Abitur has 2 children, she is more likely to get a third child. The negative effect on passing the first transition is larger than the positive effect at the third transition, but the weight of the first transition is lower than the weight of the third transition.

It is informative to show how the elements in Table 5 are related. The total effect is a weighted sum of the effects on passing the transitions. For example, the total effect for the comparison of vocational education versus basic education is $-0.057 \times 0.358 - 0.073 \times 0.421 - 0.108 \times 0.667 = -0.123$. The weights are the product of the proportion at risk and the gain: For example, the weight of the first transition is $1 \times 0.358 = 0.358$.

It is also informative to show how Table 5 can be derived from Table 4. The total effects can be computed directly from Table 4: For example, the probability of getting 3 or more children is 0.259 when a woman has vocational education and 0.382 when a woman has basic

Table 5: Decomposition of the effect of education on the probability of having a big family

		first child	second child	third child
vocational versus basic	$\Delta \hat{p}_i$	-0.057	-0.073	-0.108
	weight	0.358	0.421	0.667
	at risk	1.000	0.885	0.667
	gain	0.358	0.476	1.000
	Total effect	-0.123		
Abitur or more versus vocational	$\Delta \hat{p}_i$	-0.075	0.004	0.032
	weight	0.314	0.358	0.588
	at risk	1.000	0.819	0.588
	gain	0.314	0.438	1.000
	Total effect	-0.004		

Table 6: transition probabilities conditional on mother's education

	first child	second child	third child
basic	0.913	0.789	0.530
vocational	0.856	0.716	0.422
Abitur or more	0.781	0.720	0.454

education, so the difference is -0.123, which is the total effect reported in Table 5. In order to compute the differences in the transition probabilities one needs to compute those transition probabilities from Table 4. To repeat the example above: the probability of passing the first transition for a women with basic education is $19.2 + 33.9 + 38.2 = 91.3\%$. The other transition probabilities can be computed in the same way, and these are shown in Table 6. The effect on passing the first transition for the comparison between vocational education and basic education is $0.856 - 0.913 = -0.057$.

To compute the the weights we need the "average" transition probabilites ($\hat{p}_i^*$). For example, the $\hat{p}_{1i}^*$ for the comparison of vocational versus basic education can be computed as $(0.913 + 0.856)/2 = 0.885$. This and the other "average" transition probabilities are shown in Table 7. The comparison of Abitur and vocational education is a good illustration that these averages can be rather artificial: the transition probabilities of each group are given the same weight even though there are 2,986 women with vocational education in the sample and only 384 women with Abitur. These "average" probabilities can now be used to compute the two components of the weights as reported in Table 5. For example, for the comparison between vocational education and basic education, the gain from passing the first transition is $\hat{p}_{2i}^* \times \hat{p}_{3i}^* = 0.753 \times 0.476 = 0.358$. Similarly, for this same comparison the proportion at risk for the second transition is $\hat{p}_{1i}^* = 0.885$.

4 Conclusion

The purpose of this article was to discuss a method for studying processes consisting of a sequence of steps and especially relating effects of a variable on passing each transition to the

Table 7: "Average" transition probabilities ($\hat{p}_i^*$) for the different contrasts

	$\hat{p}_{1i}^*$	$\hat{p}_{2i}^*$	$\hat{p}_{3i}^*$	$\hat{p}_{1,2i}^*$
vocational versus basic	0.885	0.753	0.476	0.667
Abitur or more versus vocational	0.819	0.718	0.438	0.588

effect of that variable on the final outcome. The idea behind this is that effects on passing each transition and effects on the final outcome are not competing descriptions of the process being studied but natural complements. Treating these effects as complementary creates the challenge to move beyond a separate discussion of these two estimates to an integrated discussion.

This challenge was met by showing that the sequential logit model and other sequential models also imply an estimate for the effect on the final outcome. In case of a sequential logit model, this estimate is a weighted sum of effects (log odds ratios) on passing transitions such that the effect receives more weight if more people are at risk of passing that transition; if the transition is more differentiating (that is, if the proportion of respondents who pass is closer to 50%); and if there is a larger difference in the expected outcome between people who pass and fail that transition. This method decomposes the marginal effect of an explanatory variable on the final outcome, and thus applies to continuous explanatory variables. A variation of this decomposition was developed that decomposes the discrete change in final outcome in a weighted sum of discrete changes in the probabilities of passing transitions. In this decomposition a transition receives more weight when more people are at risk and the expected gain from passing a transition is larger. A limitation of this decomposition is that it is only descriptive and does not have a causal interpretation.

Apart from enabling an integrated discussion of effects on each transition and effects on the outcome, this decomposition also allows for studying the impact of changes in the distribution of the outcome on the effects of explanatory variables on that outcome. For example, if one is studying the effect of parental background on educational attainment over cohorts, it is interesting to also study the impact of the general increase in highest achieved level of education over cohorts on the effect of parental background. This can be especially interesting since the increase in average educational attainment across cohorts is one of the most universal and far-reaching changes in educational systems across countries during the 20th century (Hout and DiPrete, 2006).

This method was illustrated using two examples that are also used to introduce several extensions. The first example considered the effect of father's occupational status on the process of educational attainment in the Netherland between 1905 and 1991. The Dutch educational system is a tracked system in that children at about the age of 12 are selected in different tracks. This example was used to illustrate that the decomposition can also be applied to decision trees that are more complex than a set of hierarchical pass or fail decisions. It showed that the composition of the effect of father's occupation on the offspring's highest achieved level of education shifted from being primarily determined by the first transition

(whether or not to continue after primary education) to being primarily determined by the second transition (the choice between the vocational and the academic track). The last two transitions (whether to finish vocational or tertiary education) contributed relatively little throughout the period being studied. The differences in the distribution of education across cohorts (educational expansion) were shown to explain this shift in importance between the first and second transitions and a striking feature of the trend in effect on the final educational outcome. This feature is that the trend over cohorts showed an initial increase followed by a decrease. The initial increase can be explained by the increase in the proportion of students that pass the first two transitions from less than 50% to around 50%, thus initially increasing the weights for both transitions. The weight for the second transition also increased as more students became at risk of passing that transition. The subsequent decrease happened because the weight of the first transition sharply decreased since passing that transition became nearly universal. These changes also explain the shift in importance between the first and second transitions.

The second example considered the effect of women's education on the process through which they get a 'large' (three or more) number of children for West-German women born between 1934 and 1936. This example was used to illustrate the decomposition in terms of discrete differences instead of marginal effects. Women with vocational education have a 12% larger chance of having a large family compared to women with only lower secondary education. Most of this effect is due to the last transition (whether to get a third child after having two children). Women with higher secondary education or more have about the same chance of getting a large family compared to women with vocational education. However this is due to two effects canceling each other out: Women with higher secondary education are less likely get a first child, but are more likely to get a third child if they already have two children compared to women with vocational education.

This example was used to show what can be done when it is not possible to assign a value to each outcome state. This is problematic as the final outcome of the process is defined as the average outcome, and the average can only be computed if each outcome state has a value. However, in those cases, there is often one state that is of particular interest, and by assigning the value one to that state and the value zero to all other states, the outcome becomes the probability of attaining the state of interest. This way the effect on the final outcome becomes a marginal effect on the probability of attaining the state of interest.

In conclusion, this article has shown how the study of a sequential process, like attaining education or getting children, can be enriched by studying both effects on passing each transition within that process and effects on the final outcome of that process as complementary pieces of information.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological methodology 13*(1), 61–98.

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

Ballarino, G., F. Bernardi, M. Requena, and H. Schadee (2009). Persistent inequalities? expansion of education and class inequality in Italy and Spain. *European Sociological Review 25*(1), 123–138.

Ballarino, G., N. Panichella, and M. Triventi (2014). School expansion and uneven modernization. comparing educational inequality in northern and southern Italy. *Research in Social Stratification and Mobility 36*, 69 – 86.

Ballarino, G. and H. Schadee (2010). Allocation and distribution: A discussion of the educational transition model, with reference to the Italian case. *Research in Social Stratification and Mobility 28*(1), 45–58.

Breen, R. and J. O. Jonsson (2000). Analyzing educational careers: A multinomial transition model. *American Sociological Review 65*(5), 754–772.

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

Buis, M. L. (2010). *Inequality of Educational Outcome and Inequality of Educational Opportunityin the Netherlands during the 20th Century*. Ph. D. thesis, Faculty of Social Sciences, VU-University Amsterdam.

Buis, M. L. (2011). The consequences of unobserved heterogeneity in a sequential logit model. *Research in Social Stratification and Mobility 29*(3), 247–262.

Buis, M. L. (forthcomming). Lange-termijn trends in onderwijsuitkomsten in nederland. In H. van der Werfhorst and H. B. G. Ganzeboom (Eds.), *Een kloof van alle tijden. Verschillen tussen lager en hoger opgeleiden in werk, cultuur en politiek*, Chapter 9. Amsterdam: Amsterdam University Press.

Buis, M. L., B. Mönkediek, and S. Hillmert (2012). Educational expansion and the role of demographic factors: The case of West Germany. *Population Review 51*(2), 1–15.

Cameron, S. V. and J. J. Heckman (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.

Cole, S. R. and C. V. Ananth (2001). Regression models for unconstrained, partially or fully constrained continuation odds ratios. *International Journal of Epidemiology 30*(6), 1379–1382.

Fienberg, S. E. (1977). *The Analyses of Cross-Classified Categorical Data*. MIT Press.

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

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research 38*(2), 306–347.

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

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

Gill, J. (2006). *Essential Mathematics for Political and Social Research*. New York: Cambridge University Press.

Hauser, R. M. and M. Andrew (2006). Another look at the stratification of educational transitions: The logistic response model with partial proportionality constraints. *Sociological Methodology 36*(1), 1–26.

Hoogendoorn, A. and J. Daalmans (2009). Nonresponse in the recruitment of an internet panel based on probability sampling. *Survey Research Methods 3*(2), 59–72.

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

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford bulletin of economics and statistics 57*(1), 129–136.

Korupp, S. E., H. B. Ganzeboom, and T. Van Der Lippe (2002). Do mothers matter? a comparison of models of the influence of mothers' and fathers' educational and occupational status on children's educational attainment. *Quality and Quantity 36*(1), 17–42.

Läärä, E. and J. N. S. Matthews (1985). The equivalence of two models for ordinal data. *Biometrika 72*(1), 206–207.

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

Lucas, S. R. (2001). Effectively maintainted inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology 106*(6), 1642–1690.

MacLean, A. (2011). The stratification of military service and combat exposure, 1934–1994. *Social Science Research 40*(1), 336–348.

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

Mare, R. D. (1977). *Growth and Distribution of Schooling in White Male American Cohorts: 1907–1952.* Ph. D. thesis, University of Michigan.

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

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

Mare, R. D. (2006). Response: Statistical models of educational stratification — Hauser and Andrew's models for school transitions. *Sociological Methodology 36*, 27–37.

Mare, R. D. (2011). Introduction to symposium on unmeasured heterogeneity in school transition models. *Research in Social Stratification and Mobility 29*(3), 239–245.

Marks, G. N. (2004). The measurement of socio-economic inequalities in education: A further comment. *Acta Sociologica 47*(1), 91–93.

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

Newson, R. B. (2012). Sensible parameters for univariate and multivariate splines. *Stata Journal 12*(3), 479–504.

Nieuwbeerta, P. and S. Rijken (1996). Educational expansion and educational reproduction in eastern Europe, 1940–1979. *Czech Sociological Review 4*(2), 187–210.

Schröder, H. and H. B. G. Ganzeboom (2014). Measuring and modelling level of education in european societies. *European Sociological Review 30*(1), 119–136.

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

Smith, H. L. and P. P. Cheung (1986). Trends in the effects of family background on educational attainment in the Philippines. *The American Journal of Sociology 91*(6), 1387–1408.

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

Xie, Y. (2011). Values and limitations of statistical models. *Research in Social Stratification and Mobility 29*(3), 343–349.

Notes

[1]This model is also known under a variety of other names: sequential response model (Maddala, 1983), (unconstrained) continuation ratio logit (Agresti, 2002; Cole and Ananth, 2001), model for nested dichotomies (Fox, 1997), and Mare model (Cameron and Heckman, 1998).

[2]While both arguments by Cameron and Heckman (1998) are valid, there is some evidence that in at least one application — the effect of parental background on educational attainment in the Netherlands —- the effect of these problems is not big enough to change the qualitative conclusions (Buis, 2011).

[3]The word effect is here used in a descriptive rather than a causal sense.

[4]The exception would be when an identity link function was used.

[5]I thank an anonymous reviewer for pointing this generalization of the decomposition out to me.

[6]An alternative approach to capturing the influence of educational expansion based on the ordered logit model instead of the sequential logit model was proposed by Ballarino et al. (2009) and Ballarino and Schadee (2010).

[7]Since I measure education as the highest finished level of education, continuing education actually means continuing *and* finishing a subsequent level of eduction. Even though continuing education after primary education was compulsory during almost the entire historical period that is being studied, finishing a subsequent level of education was not compulsory.

[8]A similar scale for education was developed by Schröder and Ganzeboom (2014).

[9]For limitations of this argument see for example (Marks, 2004).

[10]These surveys are, using the coding from (Ganzeboom and Treiman, 2009): net58, net67, net67t, net70, net71c, net71, net74p, net76j, net77, net77e, net79p, net81e, net82e, net82n, net82u, net85o, net86e, net86l, net87i, net87j, net87s, net88o, net90, net90o, net91j, net92f, net92o, net92t, net94e, net94h, net94o, net95h, net95s, net95y, net96, net96c, net96o, net96y, net98, net98e, net98f, net98o, net99, net99a, net99i, net00f, net00s, net02e, net03f, net03n, net04e, net04i, net06e, net06i.

[11]A similar pattern was found using a different method in (Buis forthcoming), making it less likely that this pattern is an artefact of the method.

A Derivation and generalization of equations (3) and (4)

Equation (3) is the first derivative of equation (2). Equation (2) is repeated below:

$$E(L_i) = (1 - \hat{p}_{1i})l_A + \hat{p}_{1i}(1 - \hat{p}_{2i})l_B + \hat{p}_{1i}\hat{p}_{2i}(1 - \hat{p}_{3i})l_C + \hat{p}_{1i}\hat{p}_{2i}\hat{p}_{3i}l_D$$

To generalize the result, the $\hat{p}_{ki}$s are assumed to be dependent on the explanatory variables using a link function $g(\cdot)$ and the inverse link function $g^{-1}(\cdot)$, which could be the logit link function used in equation (1) but could also be any other link function.

$$
\begin{aligned}
g(\hat{p}_{ki}) &= \hat{\lambda}_{0k} + \hat{\lambda}_{1k}x_{1i} + \hat{\lambda}_{2k}x_{2i} = \eta_k \quad \text{if} \quad pass_{k-1\,i} = 1 \\
\hat{p}_{ki} &= g^{-1}(\eta_k) \quad \text{if} \quad pass_{k-1\,i} = 1
\end{aligned}
$$

A list of common link function is given in Table 8. An interesting link function in this list is the cloglog link function, as that will result in a discrete time hazard model (Allison, 1982; Läärä and Matthews, 1985; Jenkins, 1995). The probit link function was included in this list because it is such a common alternative to the logit link function. The log and identity link functions need to be handled with care, as they can lead to predicted probabilities outside

Table 8: Common inverse link functions and their derivatives

link name	inverse link function	$\frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}$
logit	$\frac{\exp(\eta_k)}{1+\exp(\eta_k)}$	$\hat{p}_{ki}(1-\hat{p}_{ki})$
probit	$\Phi(\eta_k)$	$\phi(\eta_k)$
cloglog	$1-\exp(-\exp(\eta_k))$	$(\hat{p}_{ki}-1)\ln(1-\hat{p}_{ki})$
log	$\exp(\eta_k)$	$\hat{p}_{ki}$
identity	η_k	1

$\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of the standard normal distribution

the [0, 1] interval. However, they can occasionally be useful when one wants effects in terms of risk ratios (log) or risk differences (identity).

The derivative of the expect outcome with respect to x_j can be computed using the sum rule, the product rule, (for example: Gill, 2006, pp. 190–191), and the derivative of the inverse link function as given in Table 8.

Using the sum rule, the first derivative can be written as:

$$\frac{\partial E(L_i)}{\partial x_j} = \frac{\partial (1-\hat{p}_{1i})l_A}{\partial x_j} + \frac{\partial \hat{p}_{1i}(1-\hat{p}_{2i})l_B}{\partial x_j} + \frac{\partial \hat{p}_{1i}\hat{p}_{2i}(1-\hat{p}_{3i})l_C}{\partial x_j} + \frac{\partial \hat{p}_{1i}\hat{p}_{2i}\hat{p}_{3i}l_D}{\partial x_j}$$

Using the product rule, this can be rewritten as:

$$
\begin{aligned}
\frac{\partial E(L_i)}{\partial x_j} =\ & l_A \frac{\partial(1-\hat{p}_{1i})}{\partial x_j} + \\
& l_B \left(\frac{\partial \hat{p}_{1i}}{\partial x_j}(1-\hat{p}_{2i}) + \frac{\partial(1-\hat{p}_{2i})}{\partial x_j}\hat{p}_{1i} \right) + \\
& l_C \left(\frac{\partial \hat{p}_{1i}}{\partial x_j}\hat{p}_{2i}(1-\hat{p}_{3i}) + \frac{\partial \hat{p}_{2i}}{\partial x_j}\hat{p}_{1i}(1-\hat{p}_{3i}) + \frac{\partial(1-\hat{p}_{3i})}{\partial x_j}\hat{p}_{1i}\hat{p}_{2i} \right) + \\
& l_D \left(\frac{\partial \hat{p}_{1i}}{\partial x_j}\hat{p}_{2i}\hat{p}_{3i} + \frac{\partial \hat{p}_{2i}}{\partial x_j}\hat{p}_{1i}\hat{p}_{3i} + \frac{\partial \hat{p}_{3i}}{\partial x_j}\hat{p}_{1i}\hat{p}_{2i} \right)
\end{aligned}
$$

The terms in this equation can be rearranged in such a way that all elements that have the same $\frac{\partial \hat{p}_{ki}}{\partial x_j}$ in common are grouped together, resulting in equation (4):

$$
\begin{aligned}
\frac{\partial E(L_i)}{\partial x_j} =\ & \\
\{1 \quad & [(1-\hat{p}_{2i})l_B + \hat{p}_{2i}(1-\hat{p}_{3i})l_C + \hat{p}_{2i}\hat{p}_{3i}l_D - l_A]\} & \frac{\partial \hat{p}_{1i}}{\partial x_j} + \\
\{\hat{p}_{1i} \quad & [(1-\hat{p}_{3i})l_C + \hat{p}_{3i}l_D - l_B]\} & \frac{\partial \hat{p}_{2i}}{\partial x_j} + \\
\{\hat{p}_{1i}\hat{p}_{2i} \quad & [l_D - l_C]\} & \frac{\partial \hat{p}_{3i}}{\partial x_j}
\end{aligned}
$$

This equation can be turned into equation (3), by choosing the logit link function, use the chain rule to rewrite $\frac{\partial \hat{p}_k}{\partial x_j}$ as $\frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}\frac{\partial \eta_k}{\partial x_j} = \frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}\lambda_{jk}$, and look up in Table 8 that $\frac{\partial g^{-1}(\eta_k)}{\partial \eta_k}$ is in that case $\hat{p}_{ki}(1-\hat{p}_{ki})$.

The function $\hat{p}_{ki}(1-\hat{p}_{ki})$ is interpreted as the differentiating capacity of the kth transition,

29

because this is a function that is largest when 50% pass that transition and decreases as either passing or failing becomes more universal. This interpretation also carries over when one uses the probit link, as $\phi(\eta_k)$ has a very similar shape. However, this is not generally true. For example, if the cloglog link was chosen the maximum effect would not happen when 50% of the people pass that transition but at approximately 63% (to be exact: $100 \times (1 - e^{-1})$). With the log link the the function steadily increases with the proportion passing that transition, so the maximum occurs when everybody passes that transition. So with those link functions it more meaningful to stick with the decomposition in the form of equation (4).

Some variations on the sequential model apply a constraint on η_k. All one needs to do to apply this decomposition to those models is adjust $\frac{\partial \eta_k}{\partial x_j}$ to incorporate that constraint. For example, a fully constrained continuation odds ratio model (Cole and Ananth, 2001) imposes the constraint that the odds ratio of passing a transition associated with variable x_j is the same for each transition. So in that model $\frac{\partial \eta_k}{\partial x_j} = \lambda_j$ instead of λ_{jk}. Similarly, in the model proposed by Hauser and Andrew (2006) $\eta_k = \lambda_{k0} + \phi_k \sum_j \lambda_j x_j$ with the identifying constraint $\phi_1 = 1$. In that case $\frac{\partial \eta_k}{\partial x_j}$ is $\phi_k \lambda_j$.

A.1 The decomposition with a multinomial link function

This decomposition will also work when the decision tree consists of points with more than two choice options. In that case, one would use the multinomial logit link function when at a transitions k there are more than two options. If there are M choices in transition k, then the link function for choice l is $\frac{\exp(\eta_{kl})}{1 + \sum_{m=1}^{M} \exp(\eta_{km})}$. Consider the example in Figure 8.

We can write the expected outcome as:

$$E(L_i) = (1 - \hat{p}_{1i}\hat{p}_{2i})l_A + \hat{p}_{1i}l_B + \hat{p}_{2i}(1 - \hat{p}_{3i})l_C + \hat{p}_{1i}\hat{p}_{3i}l_D$$

Computing the first derivative, and collecting terms with the same $\frac{\partial \hat{p}_{ki}}{x_j}$ leads to:

$$\frac{\partial E(L_i)}{\partial x_j} = [l_B - l_A]\frac{\partial \hat{p}_{1i}}{x_j} + [(1 - \hat{p}_{3i})l_C + \hat{p}_{3i}l_D - l_A]\frac{\partial \hat{p}_{2i}}{\partial x_j} + \hat{p}_{2i}[l_D - l_C]\frac{\partial \hat{p}_{3i}}{\partial x_j}$$

With a multinomial link function the derivative of $\hat{p}_{kli}$ with respect to x_j is $\hat{p}_{kli}(\lambda_{klj} - \sum_{m=1}^{M} \lambda_{kmj}\hat{p}_{kmi})$. Filling this into the equation and collecting terms with the same λ_{klj} leads to:

$$
\begin{array}{llllll}
\frac{\partial E(L_i)}{\partial x_j} & = & & & & \\
\{1 & \times & \hat{p}_{1i}(1-\hat{p}_{1i}) & \times & [l_B - \frac{1-\hat{p}_{1i}-\hat{p}_{2i}}{1-\hat{p}_{1i}}l_A - \frac{\hat{p}_{2i}(1-\hat{p}_{3i})}{1-\hat{p}_{1i}}l_C - \frac{\hat{p}_{2i}\hat{p}_{3i}}{1-\hat{p}_{1i}}l_D] & \} \hat{\lambda}_{j1} + \\
\{1 & \times & \hat{p}_{2i}(1-\hat{p}_{2i}) & \times & [(1-\hat{p}_{3i})l_C + \hat{p}_{3i}l_D - \frac{1-\hat{p}_{1i}-\hat{p}_{2i}}{1-\hat{p}_{2i}}l_A - \frac{\hat{p}_{1i}}{1-\hat{p}_{2i}}l_B] & \} \hat{\lambda}_{j2} + \\
\{\hat{p}_{2i} & \times & \hat{p}_{3i}(1-\hat{p}_{3i}) & \times & [(l_D - l_C)] & \} \hat{\lambda}_{j3}
\end{array}
$$

The structure of this decomposition is exactly the same as in equation (3), except that when computing the nett gain (the part in square brackets) from choosing l in transition k when there are more than two possible choices, one needs to be explicit that the loss is conditional on not choosing l.

30

Figure 8: A process with a multinomial choice
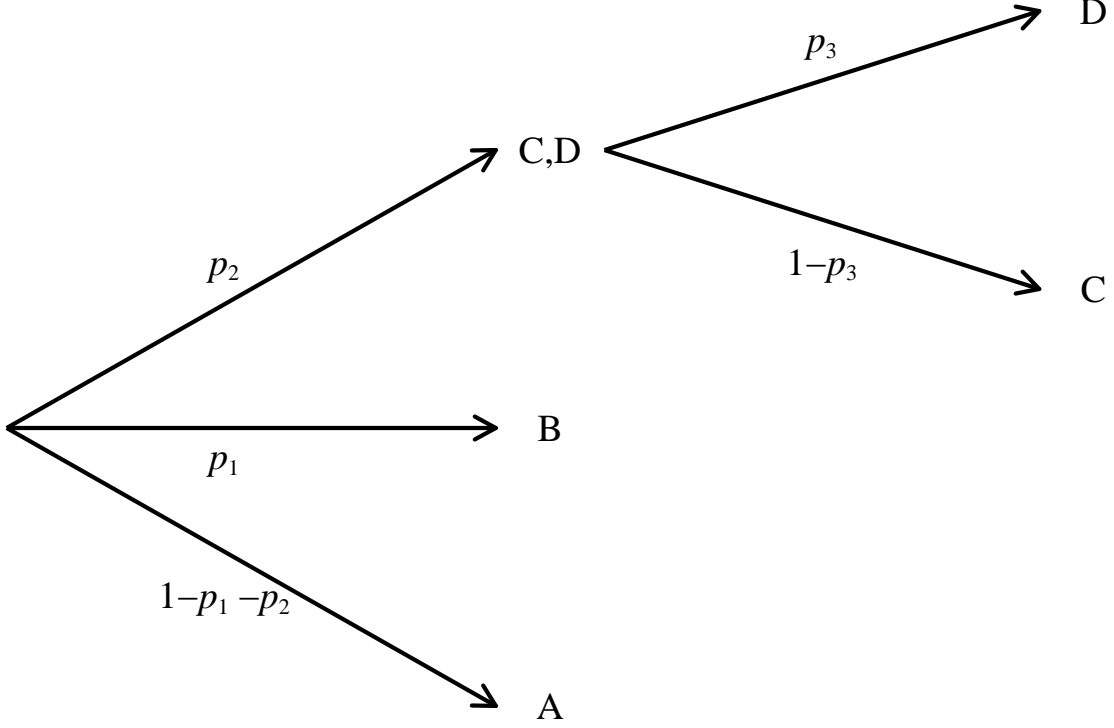
B Derivation of equation (6)

The derivation starts with equation (5), repeated below:

$$\Delta \hat{L}_i = \left[(1-\hat{p}_{1i}^1)l_A + \hat{p}_{1i}^1(1-\hat{p}_{2i}^1)l_B + \hat{p}_{1i}^1\hat{p}_{2i}^1(1-\hat{p}_{3i}^1)l_C + \hat{p}_{1i}^1\hat{p}_{2i}^1\hat{p}_{3i}^1 l_D\right] - \\ \left[(1-\hat{p}_{1i}^0)l_A + \hat{p}_{1i}^0(1-\hat{p}_{2i}^0)l_B + \hat{p}_{1i}^0\hat{p}_{2i}^0(1-\hat{p}_{3i}^0)l_C + \hat{p}_{1i}^0\hat{p}_{2i}^0\hat{p}_{3i}^0 l_D\right]$$

Collecting terms sharing the same l gives:

$$\Delta \hat{L}_i = \left[(1-\hat{p}_{1i}^1) - (1-\hat{p}_{1i}^0)\right]l_A + \\ \left[\hat{p}_{1i}^1(1-\hat{p}_{2i}^1) - \hat{p}_{1i}^0(1-\hat{p}_{2i}^0)\right]l_B + \\ \left[\hat{p}_{1i}^1\hat{p}_{2i}^1(1-\hat{p}_{3i}^1) - \hat{p}_{1i}^0\hat{p}_{2i}^0(1-\hat{p}_{3i}^0)\right]l_C + \\ \left[\hat{p}_{1i}^1\hat{p}_{2i}^1\hat{p}_{3i}^1 - \hat{p}_{1i}^0\hat{p}_{2i}^0\hat{p}_{3i}^0\right]l_D$$

This can be rewritten as:

$$\Delta \hat{L}_i = -\left[\hat{p}_{1i}^1 - \hat{p}_{1i}^0\right]l_A +$$
$$\left[\left\{\frac{(1-\hat{p}_{2i}^1)+(1-\hat{p}_{2i}^0)}{2}\right\}\left\{\hat{p}_{1i}^1 - \hat{p}_{1i}^0\right\} + \left\{\frac{\hat{p}_{1i}^1+\hat{p}_{1i}^0}{2}\right\}\left\{(1-\hat{p}_{2i}^1)-(1-\hat{p}_{2i}^0)\right\}\right]l_B +$$
$$\left[\left\{\frac{\hat{p}_{2i}^1+\hat{p}_{2i}^0}{2}\right\}\left\{\frac{(1-\hat{p}_{3i}^1)+(1-\hat{p}_{3i}^0)}{2}\right\}\left\{\hat{p}_{1i}^1 - \hat{p}_{1i}^0\right\} + \left\{\frac{\hat{p}_{1i}^1+\hat{p}_{1i}^0}{2}\right\}\left\{\frac{(1-\hat{p}_{3i}^1)+(1-\hat{p}_{3i}^0)}{2}\right\}\left\{\hat{p}_{2i}^1 - \hat{p}_{2i}^0\right\} - \right.$$
$$\left.\left\{\frac{\hat{p}_{1i}^1\hat{p}_{2i}^1+\hat{p}_{1i}^0\hat{p}_{2i}^0}{2}\right\}\left\{\hat{p}_{3i}^1 - \hat{p}_{3i}^0\right\}\right]l_C +$$
$$\left[\left\{\frac{\hat{p}_{2i}^1+\hat{p}_{2i}^0}{2}\right\}\left\{\frac{\hat{p}_{3i}^1+\hat{p}_{3i}^0}{2}\right\}\left\{\hat{p}_{1i}^1 - \hat{p}_{1i}^0\right\} + \left\{\frac{\hat{p}_{1i}^1+\hat{p}_{1i}^0}{2}\right\}\left\{\frac{\hat{p}_{3i}^1+\hat{p}_{3i}^0}{2}\right\}\left\{\hat{p}_{2i}^1 - \hat{p}_{2i}^0\right\} + \right.$$
$$\left.\left\{\frac{\hat{p}_{1i}^1\hat{p}_{2i}^1+\hat{p}_{1i}^0\hat{p}_{2i}^0}{2}\right\}\left\{\hat{p}_{3i}^1 - \hat{p}_{3i}^0\right\}\right]l_D$$

This can now be rewritten in terms of $\hat{p}^*$s and $\Delta\hat{p}$s:

$$\Delta\hat{L}_i = -\Delta\hat{p}_{1i}l_A +$$
$$\left[(1-\hat{p}_{2i}^*)\Delta\hat{p}_{1i} - \hat{p}_{1i}^*\Delta\hat{p}_{2i}\right]l_B +$$
$$\left[\hat{p}_{2i}^*(1-\hat{p}_{3i}^*)\Delta\hat{p}_{1i} + \hat{p}_{1i}^*(1-\hat{p}_{3i}^*)\Delta\hat{p}_{2i} - \hat{p}_{1,2i}^*\Delta\hat{p}_{3i}\right]l_C +$$
$$\left[\hat{p}_{2i}^*\hat{p}_{3i}^*\Delta\hat{p}_{1i} + \hat{p}_{1i}^*\hat{p}_{3i}^*\Delta\hat{p}_{2i} + \hat{p}_{1,2i}^*\Delta\hat{p}_{3i}\right]l_D$$

Collecting terms sharing the same $\Delta\hat{p}$ gives equation (6):

$$\Delta\hat{L}_i = \begin{aligned} &\{ \quad 1 \quad \times \quad [(1-\hat{p}_{2i}^*)l_B + \hat{p}_{2i}^*(1-\hat{p}_{3i}^*)l_C + \hat{p}_{2i}^*\hat{p}_{3i}^*l_D - l_A] \quad \} \quad \Delta\hat{p}_{1i} \quad + \\ &\{ \quad \hat{p}_{1i}^* \quad \times \quad [(1-\hat{p}_{3i}^*)l_C + \hat{p}_{3i}^*l_D - l_B] \quad \} \quad \Delta\hat{p}_{2i} \quad + \\ &\{ \quad \hat{p}_{1,2i}^* \quad \times \quad [(l_D - l_C)] \quad \} \quad \Delta\hat{p}_{3i} \end{aligned}$$

Notice that, unlike the decomposition of marginal effects, the link function does not play a role in this decomposition.

C   Applying the decomposition to the Dutch education system

Using Figure 1 one can write the expected value of the final outcome — in this case, the highest achieved level of education — as equation (8).

$$E(L_i) = (1-\hat{p}_{1i})l_A + \hat{p}_{1i}(1-\hat{p}_{2i})(1-\hat{p}_{3i})l_B + \hat{p}_{1i}(1-\hat{p}_{2i})\hat{p}_{3i}l_C + \qquad (8)$$
$$p_{1i}\hat{p}_{2i}(1-\hat{p}_{4i})l_D + p_{1i}\hat{p}_{2i}\hat{p}_{4i}l_E$$

Using similar steps as in Appendix A, the effect of variable x_j on the highest achieved level of education can then be written as equation (9).

$$
\begin{aligned}
\frac{\partial E(L_i)}{\partial x_j} \quad &= \\
\{1 \quad &\times \quad \hat{p}_{1i}(1-\hat{p}_{1i}) \quad \times \quad [(1-\hat{p}_{2i})(1-\hat{p}_{3i})l_B+ \\
& \qquad\qquad\qquad\qquad\qquad\quad (1-\hat{p}_{2i})\hat{p}_{3i}l_C+ \\
& \qquad\qquad\qquad\qquad\qquad\quad \hat{p}_{2i}(1-\hat{p}_{4i})l_D+ \\
& \qquad\qquad\qquad\qquad\qquad\quad \hat{p}_{2i}\hat{p}_{4i}l_E - l_A] \qquad\quad \} \quad \hat{\lambda}_{j1} \quad + \\
\{\hat{p}_{1i} \quad &\times \quad \hat{p}_{2i}(1-\hat{p}_{2i}) \quad \times \quad [(1-\hat{p}_{4i})l_D+\hat{p}_{4i}l_E- \\
& \qquad\qquad\qquad\qquad\qquad\quad (1-\hat{p}_{3i})l_B-\hat{p}_{3i}l_C] \quad \} \quad \hat{\lambda}_{j2} \quad + \\
\{\hat{p}_{1i}(1-\hat{p}_{2i}) \quad &\times \quad \hat{p}_{3i}(1-\hat{p}_{3i}) \quad \times \quad [(l_C-l_B)] \qquad\qquad\qquad \} \quad \hat{\lambda}_{j3} \quad + \\
\{\hat{p}_{1i}\hat{p}_{2i} \quad &\times \quad \hat{p}_{4i}(1-\hat{p}_{4i}) \quad \times \quad [(l_E-l_D)] \qquad\qquad\qquad \} \quad \hat{\lambda}_{j4}
\end{aligned}
\tag{9}
$$

This equation has the same structure as equation (3), the effect on the final outcome is a weighted sum of the effects on passing each transition, the $\hat{\lambda}_{jk}$s. The weights (the parts between curly brackets) consist of the same three parts:

1. The proportion of people at risk (1, $\hat{p}_{1i}$, $\hat{p}_{1i}(1-\hat{p}_{2i})$, and $\hat{p}_{1i}\hat{p}_{2i}$ respectively).

2. A part ($\hat{p}_{ki}(1-\hat{p}_{ki})$) that is small if virtually everybody passes or fails that transition and is largest when the probability of passing is 0.5.

3. The differences between the expected levels of education of those who pass the transitions and those who do not (these are the parts in the square brackets).