VOLUME 6 | ISSUE 4 | **OCTOBER 2015**

ISSN 1757-9597

Longitudinal and Life Course Studies: International Journal

Published by

SLLS Society for Longitudinal and Life Course Studies

Inside this issue

- Parental criminality and children's life trajectories
- Rose's paradox and delinquency prevention
- Covariate effects in latent class analysis
- Debate on population sampling in longitudinal studies
- Response to debate on social class differences in early cognitive development

Co-sponsors:

VU university medical center

iSER INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

SRC SURVEY RESEARCH CENTER UNIVERSITY OF MICHIGAN

LLCS EDITORIAL BOARD

EDITORIAL COMMITTEE

SUBSCRIPTIONS

For immediate access to current issue and preceding 2015 issues (Volume 6):

***Annual Individual rate:* £20 (discounted rate £55 for 3 years)**.

***Annual Library* rate: £200** to register any number of library readers.

***Annual Society for Longitudinal and Life Course Studies (SLLS) Membership: Individual* £65, *Student* £26, *Corporate* £300**.

SLLS members have automatic free access to the journal among many other benefits. For further information about SLLS and details of how to become a member, go to http://www.slls.org.uk/#!services/ch6q

All issues are accessible free of charge 12 months after publication.

Print on demand

An attractive printed version of each Issue of the journal, including all back numbers, is available at a price of £10 (plus postage and packaging). If you would like to purchase a full set or individual copies, visit the SLLS Bookshop at http://www.slls.org.uk/#!journal-bookshop/czkc Depending on distance, you will receive the order within two weeks of submitting your order.

Please note

The reselling of personal registration subscriptions and hard copies of the journal is prohibited.

SLLS disclaimer

The journal's editorial committee makes every effort to ensure the accuracy of all information (journal content). However the Society makes no representations or warranties whatsoever as to the accuracy, completeness or suitability for any purpose of the content and disclaims all such representations and warranties, whether express or implied to the maximum extent permitted by law. The SLLS grants authorisation for individuals to photocopy copyright material for private research use only.

INTRODUCTION

PAPERS

STUDY PROFILE

COMMENT AND DEBATE

LLCS Journal can be accessed online at: www.llcsjournal.org

Published by
The Society for Longitudinal and Life Course Studies
Institute of Education, 20 Bedford Way, London, WC1H 0AL
info@slls.org.uk http://www.slls.org.uk

SLLS Society for Longitudinal and Life Course Studies

SLLS Society for Longitudinal and Life Course Studies

Join our mailing lists…

Cohort Network Group

SLLS is proud to host a forum for people working in and on longitudinal studies. It aims to build on links made under the EUCCONET (*European Child Cohort Network*) whose funding for co-ordination and communication between child cohorts ended in 2013. That venture brought together researchers across the behavioural, developmental, and health and statistical sciences, and the professional data, survey and communications managers who are also an important part of the interdisciplinary teams who create and run these studies.

Key objectives of the network are the maintenance and continuation of existing studies and the facilitation of the development of new ones at local or national level, even if the aspiration for a pan-European cohort seems unrealistic.

For full details and to join the CN mailing list visit http://www.slls.org.uk/#!cohort-network/c21hq

Interdisciplinary Health Research Group

Large-scale social surveys increasingly collect biomedical data, but at present an inter-disciplinary forum concerned with making best use of these combined social and biological data, is lacking.

A preparatory meeting was held at the SLLS Annual Conference 2014, to assess whether SLLS could fill this gap. Twenty conference delegates from the social and biological sciences attended the preparatory meeting and agreed to propose to the SLLS Executive Committee that a SLLS sub-group on *Interdisciplinary Health Research* be established. The Executive Committee agreed the group with the following remit:

- To enable informed use of biomarkers by social scientists
- To enable informed use of social data by biologists
- To bring together SLLS researchers from a variety of disciplines who work on or have an interest in health and health-related issues

For full details and to join the IHR mailing list visit www.slls.org.uk/#!health-research/c1njv

Policy Group

Life course study and longitudinal research are potentially of central importance to the policy process. The burgeoning of major longitudinal studies throughout the world and the allocation of large-scale government funding to building longitudinal resources reflect this growing interest. In this respect, SLLS is well placed to identify the expertise and research resources needed to underpin the relevant evidence base in different policy domains. For this reason the SLLS Executive Committee decided to create a database registering members' expertise, relevant experience and policy interest areas. It acts as a source of partners for collaboration on international longitudinal research projects directed at policy issues; helps the Executive Committee respond to policy debates; and broaden the scope of our international journal, LLCS, in policy research directions.

For full details and to join the PG mailing list visit www.slls.org.uk/#!policy-group/c99m

www.slls.org.uk

Editorial

John Bynner

The publication of the October issue of the journal marks a number of significant events in the journal's life history. The early volumes comprising three issues now give way to our first four-issue volume – Volume 6. The Society for Longitudinal and Life Course Studies conference just held in Dublin was similarly sixth in its series, beginning with the society's foundation in Cambridge in 2010.

Annual conference

Held in the magnificent setting of Dublin Castle, the latest conference had the highest attendance yet with 340 participants from more than 20 countries. There were more papers presented than ever before in the five programme strands, including a 'workshop' strand comprising five symposia devoted to the longitudinal research/policy interface.

Every paper and symposium is a potential publication for the journal, so the symbiosis between journal and society yields dividends all round. The larger the number of participants, the larger the potential number of authors of individual papers and special sections.

Editorial challenges

Yet enthusiasm needs to be tempered with one major concern from the Editorial Committee meeting – the increasing difficulty in persuading subject experts to accept invitations to review papers. With the expansion of journal content, the number of peer reviews conducted continues to increase, rising from 146 in 2013 to 185 in 2015 – an ever pressing demand on an ever-shortening supply.

We brain-stormed in Dublin various ways of heading off refusals, and these will be tested in the coming months, but the need is always for more experts to approach. And that is partly a matter of reputation. We rely on you, the 2000+ writers and readers of the journal, to sing its praises whenever you can.

Current issue

The content of this issue is also breaking new ground in a number of ways. It starts with two papers in the relatively new area for the journal of *life course criminology.* The first focuses on family life courses and child outcomes in high crime risk socioeconomic backgrounds, covering life events recorded from a range of administrative and survey data collected for a Dutch cohort of 522 across the age period 18-50.

The second paper similarly investigates criminal propensity risk. This time attention is focused on the paradox that low risk population populations tend to produce relatively more negative (criminal) outcomes than the expected high risk minority ones, arguing against highly targeted interventions.

Using simulated and large scale birth cohort study data, the third paper moves to optimising bias removing strategy for progressing the results of mixture modelling directed at patterns of change – at different levels of classification quality ('entropy') – to the identification of latent classes.

The paper following is again a first of its kind, profiling the long-standing Zurich longitudinal study spanning a period of 40 years starting with the transition from school to work. The journal welcomes the opportunity, unique to longitudinal research, of such life histories giving unparalleled insights into the ways important research design and operation decisions were taken and their consequences for the later development of the study and its outputs.

Next comes the second of our new LCCS ventures, 'Comment and Debate' on major topical issues in longitudinal and life course research. This issue's debate is about the role of national population sampling in longitudinal research and comprises a discussion paper from Harvey Goldstein challenging the need for such sampling. He argues that scientific advance is best gained from multiple replication in different settings rather than parameter estimation for a population that, from first contact, is getting progressively out of date. The paper is followed by responses from leading experts in the field to whom he exercises his right of reply.

Finally, we complete the debate which began in the July issue on the socioeconomic gradient in cognitive development with the response from lead author Leon Feinstein.

Debates in subsequent issues will address ‘Allostatic Load‘ and ‘Positive Health’. The whole series may lend itself to reproduction for wider readership in monograph form and will be kept under review.

Parental criminality and children's family-life trajectories: Findings for a mid-20th century cohort

Doreen Huschek University of Stockholm, Sweden
The Netherlands Institute for the Study of Crime and Law Enforcement
DoreenHuschek@criminology.su.se
Catrien Bijleveld VU University Amsterdam, Netherlands

(Received December 2013 Revised August 2014) http://dx.doi.org/10.14301/llcs.v6i4.282

Abstract

The paper analyses the family life courses of sons and daughters from families with low socioeconomic status and at high risk to offend. For this Dutch cohort (N=522), born on average in 1932, register and archive data on offending and family-life events from age 18 to 50 years are investigated. We discuss different mechanisms of how parental criminality may affect demographic behaviours, such as marriage and parenthood. As these demographic behaviours are interlinked, and as their ordering is meaningful, we apply a holistic approach by using sequence and cluster analysis to construct family-life courses. Findings indicate four family-life trajectories that are almost similar for the sons and daughters, although criminal fathers appear to affect sons' and daughters' trajectories differently. Daughters' family-life trajectories seem directly affected by father's offending whereas sons' trajectories are only affected by their own juvenile offending.

Keywords

Intergenerational transmission, demographic behavior, life course, sequence analysis, incarcerated parents

Introduction

It is well known that the behaviour of children is linked to that of their parents. This intergenerational behavioural continuity also relates to deviant behaviours or non-normative relationships. For example, children of criminal parents are more likely to become criminals themselves (Farrington, Coid, & Murray, 2009; Thornberry, 2005) and parental divorce increases children's own risk of divorce (Dronkers & Härkönen, 2008). Apart from these associations within domains, there are also associations across behavioural domains, such as the link between offending and family transitions. For instance, parental divorce (Burt, Barnes, McGue, & Iacono, 2008; Fergusson, Horwood, & Lynskey, 1992) and early parenthood (Pogarsky, Lizotte, & Thornberry, 2003) both increase the likelihood of offspring offending. Similarly, parental incarceration has been linked to a wide array of adverse outcomes for children such as anti-social behavior, internalizing behaviours, poor wellbeing and educational outcomes (Comfort, 2007; Foster & Hagan, 2007; Murray & Farrington, 2008) but also off-time demographic transitions (Osgood, Foster, Flanagan, & Ruth, 2005) such as early marriage or parenthood or children out of wedlock. Due to the sharp increase of prisoners in the United States in the past few decades, although less in Europe, there is a growing interest in studying the possible collateral effects of parental prison terms on prisoners' families and children.

In this paper, we investigate the long-term outcomes of parental criminality and associated family risk factors on children's demographic life courses. We aim to add to existing research in several ways. First, most studies that investigate offending and demographic transitions have taken offending as the outcome variable, with demographic and other transitions as predictors. Few studies have looked at data from the other direction: how (parental) offending may influence demographic transitions.

Second, this paper takes a life-course approach. Studies have shown that vulnerable populations such as children of incarcerated parents are more likely to experience off-time transitions, for example early parenthood and early marriage, and are at a higher risk of divorce (Elder, 1994; Osgood, Foster, Flanagan, & Ruth, 2005; Settersten, 2003). However, most studies have focused on examining a single life-course transition. This makes it difficult to know whether the different off-time demographic transitions cluster within a small group of vulnerable individuals who experience all the off-time transitions while the majority of vulnerable individuals experience standard life courses, or whether the likelihood of different off-time transitions is equally distributed among vulnerable groups with, for example, some individuals experiencing early parenthood and others experiencing early marriage. Furthermore, the study of isolated life-course transitions may lead to seemingly inconsistent findings, for instance, out-of-wedlock parenthood is differently associated with offending than parenthood within marriage (Zoutewelle-Terovan, Van der Geest, Liefbroer, & Bijleveld, 2014). Moreover, such single-event analyses disregard the ordering of demographic transitions. This study will combine various demographic behaviours such as marriage, divorce and parenthood into sequences or, in other words, family life courses, thereby studying not only the occurrence but also the ordering, co-occurrence and timing of demographic behaviours. With this approach, we can much better understand how family life courses and parental offending are intertwined.

Third, previous studies have investigated predominantly men so that much less is known about gendered effects. There is reason to expect these as some demographic transitions are much more age-constrained for women than for men, such as the transition to parenthood. Also, there is evidence that sons and daughters are differently affected by paternal offending: sons appear at increased risk to offend if their father offended (Farrington et al., 2009; Van de Rakt, Nieuwbeerta, & De Graaf, 2008) but daughters seem to be at increased risk of leaving home early in case of the replacement of an incarcerated father by an abusive non-biological father (Foster & Hagan, 2007).

In this paper, we investigate the impact of family risk factors, such as juvenile offending, parental demographic behaviour and criminality, on children's family-life courses from age 18-50 by using register and archival data on a sample of sons and daughters born on average in 1932 in the Netherlands. They were born into families with low socioeconomic status and at high-risk to offend. This sample matured into adulthood in a period where life courses became increasingly standardized with large regularity in the timing and occurrence of family-life transitions as well as low levels of crime. As our sample originates from marginalised and poor segments of the society, we may expect a relatively large number of non-standard or off-time life courses (McLanahan, 2004; McLeod & Kessler, 1990; Settersten, 2003). Given their high risk of offending, any associations between offending and life course trajectories are bound to be easily detectable.

Our analyses will firstly depict family-formation patterns in our sample in comparison to the general Dutch population. Next, we will describe the most common sequences of family life courses of our sample. Finally, we will investigate to what extent family-risk factors are associated with the previously described family life courses.

The sequencing of demographic transitions in the life course

The life course is connected to age and sequencing norms and internalised orders of events (Elder, 1994; Settersten, 2003). Most men and women follow relatively standard sequences of demographic transitions. A standard sequence nowadays in many Western societies is cohabitation, marriage, followed by parenthood. What is a "standard" life course depends, however, on cultural norms as well as on the era in which transitions take place. A nonstandard life course can consist of transitions occurring early (for instance, early parenthood), late, or not at all (for instance, remaining single). Nonstandard life courses may include repeated events such as marrying several times or negative events such as divorce. They can also consist of a nonstandard order of transitions.

For example, a person may have a child outside marriage or before marriage, nowadays a common sequence, but considered deviant in a large part of the previous century.

Individuals who go through demographic transitions too early or too fast according to prevailing norms may be subjected to informal normative control and sanctions (Neugarten & Hagestad, 1976). The resulting consequences of non-standard transitions are for example a higher likelihood of depression, lower self-esteem, a higher chance of divorce, lower well-being, and lower achievement in education and work. Such non-standard transitions therefore may also affect children's outcomes in various domains (e.g., Gilman, Kawachi, Fitzmaurice, & Buka, 2003; Koropeckyi-Cox, Pienta, & Brown, 2007; Pogarsky, Thornberry, & Lizotte, 2006; Sigle-Rushton, 2005).

Parental criminality and children's demographic behaviors

There is evidence that vulnerable groups, such as youth from low socioeconomic backgrounds or from single, divorced, criminal or imprisoned parents, are less likely to follow standard life courses (Elder, 1994; Settersten, 2003) and experience earlier transitions to parenthood and marriage (Osgood, Foster, Flanagan, & Ruth, 2005).

How such a high-risk background is associated with embarking on nonstandard life courses has hardly been addressed. A number of causal mechanisms can be envisaged through which children of criminal parents would be at increased risk to follow nonstandard life courses, that is, to experience transitions to parenthood or marriage either early, not at all, late, or in nonstandard order, and if married to be at increased risk of divorce.

In a first mechanism, children may be *socialised* to not internalise or even reject conventional norms by experiencing and observing their parents breaking them. Parents serve as role models and transmit their preferences, attitudes and behaviours to their children by rewarding or punishing certain behaviors. If a parent is criminal, events such as divorce or out-of-wedlock childbirth may be judged less as a break of norms by their children, as children may transpose 'being deviant' to other domains. Thus, children may divorce or have children outside marriage even if their criminal parents did not divorce or have children out of marriage themselves.

Secondly, *stigma* is pivotal. For children of arrested, convicted, or incarcerated parents, such stigma has been extensively reported (e.g., Foster & Hagan, 2007; Phillips & Gates, 2011; Murray & Farrington, 2008). Other individuals can attribute the negative characteristics of incarcerated parents to their children. Out of fear for such stigmatisation, fear of harassment, and bullying, these children may hide the fact that their parent is in prison (Hissel, Bijleveld, & Kruttschnitt, 2011; Nesmith & Ruhland, 2008). Phillips and Gates (2011) reported how children may also internalise such societal reactions or beliefs about their parents and themselves. In this case the stigma may fuel a self-fulfilling prophecy, concurring with labeling explanations. Stigma and associated experiences such as social isolation are likely more severe in times and contexts where offending is of low prevalence and consequently is a rare type of societal norm breaking - like the context studied in this paper. Such stigma may extend into adulthood and reduce marriage chances: ethnographic research (Anderson, 1999; Edin, 2000) has shown that women from poorer segments of society heavily weigh the bread-earning capacity of prospective spouses.

In addition, the *interaction with the criminal justice system*, which relatives and children of offenders experience, can lead to negative outcomes. Witnessing arrest is known to cause trauma in children (Comfort, 2007) or can be experienced as highly emotional and disturbing (Braman, 2004; Hissel et al., 2011). These experiences and the parent-child separation may affect parental attachment and trust, with subsequent effects for later relationship formation, for instance a higher likelihood of divorce.

In all these mechanisms, it is the parental criminality itself that *directly* generates nonstandard life-course outcomes.

From the literature, it is possible to derive a number of *indirect* paths through which parental criminality predisposes children to experience nonstandard demographic life courses. It has been well-established that parental criminality increases the risk for offspring delinquency (e.g., Farrington et al., 2009; Thornberry, 2005). Offspring offending could subsequently affect these children's life course outcomes, for similar reasons as for parental criminality in the "stigma" explanation above. As Svarer (2011) showed, convicted men are regarded by women, particularly women from better-off families, as a less good "investment", because the men's earning potential is considered lowered by their criminal records. The association between parental criminality and children's demographic outcomes is channeled here through offspring criminality.

A last *indirect* mechanism could be parental offending affecting parental life-course outcomes, which in turn are then transmitted to the children: the literature has provided ample evidence of such intergenerational transmission of demographic outcomes (e.g., Amato, 2000; Dronkers & Härkönen, 2008). The association between parental criminality and children's demographic outcomes is here channeled through parental demographic behavior.

Thus, in the first mechanisms (socialisation into deviant behaviours, stigma and contact with criminal justice system), parental criminality is directly affecting children's life-course outcomes. In the last two mechanisms, there is no "cross-over" between criminality of the parents and children's demographic outcome. Rather it is criminality or demographic behaviours that are transmitted across generations and that are generating an association between criminality and children's demographic outcomes.

Gender-specific intergenerational transmission

The life course of sons and daughters is likely to be differently affected by father's criminality. There are general behavioural difference between men and women, for example they differ in the timing of their life course, particularly concerning family roles: marriage and parenthood. Men and women also have a different sequencing and combination of these roles (Elder, 1998; Oesterle, Hawkins, Hill, & Bailey, 2010). Women are also much less likely to offend than men (Block, Blokland, Van der Werff, Van Os, & Nieuwbeerta, 2010).

Intergenerational transmission of fertility generally suggests a stronger transmission from mothers to children and from mothers to daughters in particular (Murphy, 1999), however many studies only include mothers and their daughters (Furstenberg, Levine, & Brooks-Gunn, 1990; Horwitz, Klerman, Kuo, & Jekel, 1991; Murphy & Knudsen, 2002). It is further indicated that daughters are more susceptible to home and family influences than sons due to gender-specific socialisation processes (Murphy & Knudsen, 2002).

Due to data restrictions and the low prevalence of offending among women, research on intergenerational transmission of offending focuses more often on sons than on daughters and there are only few studies that include both genders. Farrington et al. (2009) found significant intergenerational transmission of offending from fathers to sons, but much less strong transmission from mother to sons or fathers to daughters, although, in a Dutch study, Van de Rakt et al. (2008) found intergenerational transmission of criminal careers of the father to both sons and daughters. Foster and Hagan (2007) indicated furthermore that daughters but not sons are at increased risk of leaving home early due to the replacement of an incarcerated father by an abusive non-biological father.

Historical background: Life course and crime levels in the Netherlands

The 20th century was marked by many changes in the life course of young adults in Europe and North America. During the first half of the century, markers of the transition to adulthood such as leaving home, marriage, and parenthood tended to occur earlier and followed increasingly standard trajectories. This trend had started already in the later second half of the 19th century in Europe (Bras, Liefbroer, & Elzinga, 2010). Standardised life courses were most pronounced among individuals marrying in the 1950s and first half of the 1960s. Economic growth after World War II allowed earlier and more plannable life choices in work and family spheres (Fussell & Furstenberg, 2005). Among Dutch individuals born between 1921 and 1940, few remained childless or unmarried. Unmarried cohabitation was below 5% and marriages were generally long and stable (Liefbroer & Dykstra, 2000; see also table 2).

Thus, the sample studied here, who were on average starting on the path to adulthood a few years after the Nazi occupation (1940-1945) of the Netherlands, belong to birth cohorts with the most standardised life courses in the 20th century. By contrast, for cohorts born in the second half of the 20th century family formation patterns became increasingly de-standardised (Elzinga & Liefbroer, 2007).

In this period, crime levels in the Netherlands were moving towards an all-time low for the 20th century. From the 1950s to the early 1970s, the number of prisoners per capita was at an historic low (Tonry & Bijleveld, 2007), even though the number of police registered offences started rising from the 1960s onwards. This means that the sample under study entered adulthood in a period when deviation from the norm, either in terms of demographic behaviour or criminal behaviour, was a rare event.

Method

Sample

We analyse data from the *TransFive* study (Bijleveld & Wijkman, 2009) that has register information on family formation and offending for five generations of men and women born within 198 families in the Netherlands. The starting point of the study were 198 men who had been placed in a reform school between 1911 and 1914 either because of concerns about their character and behaviour (including some petty delinquency) or because their parents had been unable to take proper care of them according to guardian organisations. Previous studies showed that the sample was from a poorly educated, disadvantaged background (Bijleveld, Wijkman, & Stuifbergen, 2007; Ramakers, Bijleveld, & Ruiter, 2011).

All descendants of these men were traced in Dutch genealogical and municipal records, entailing a 100% retrieval rate. The main data collection took place between 2004 and 2007 and register data were updated in 2012. 141 of the 198 original men fathered a total of 621 children who constitute our sample of interest. Being born on average in 1932, these sons and daughters are now mainly between the ages 60 to 85 years old. However, as we are interested in their family life course, we limit our focus on the age range 18 to 50 years. Age 50 is a common cutoff point for family life studies as basically all relationship formations and dissolution as well as childbirths have occurred by this age. We excluded those who died before their 19th birthday (8% of the sample) and those who migrated between the age of 0 and 50 years (7% of the sample), leaving 522 sample members: 259 sons and 263 daughters, nested within 141 families.

Family-life trajectories

Individual family-life trajectories were constructed from register data (see Bijleveld & Wijkman, 2009 for information on the data retrieval procedure). For each sample member, archival records on the date of birth, date of marriage(s), date of divorce(s), date of migration and date of death, as well as date of birth and death of any children are available. These variables were used to construct the life-course sequences explained in detail in the analytical strategy section. From the literature we would expect that our sample is likely to experience early transitions to marriage and parenthood as well as divorce or non-transitions. Furthermore, they are likely to experience a non-standard order of family-life transitions. We therefore chose to take into account the nuptial states: single, married, widowed, divorced and remarried. In the fertility domain we decided to not control for the number of children but rather whether our sample members ever became parents and whether their children were born outside marriage, conceived before marriage or after marriage. We include whether our sample members had conceived a child before marriage, as in the Dutch population during our sample's youth, children conceived both before or outside marriage were a rare event that generally signaled deviance.

Analytical strategy

The analysis contains three steps. The first step consists of visualizing and describing the family-life courses between the age years 18 to 50 separately for sons and daughters using sequence analysis. This approach looks at the life course in its entirety and allows for the study of the timing, duration, and order in which transitions take place as well as the building of typologies. For that, we first define our states of interest. We distinguish 11 different states: ten combined states in the fertility and union formation domains as well as the state 'death'.

In the union formation domain, the possible states are: single, married, divorced, widowed, and remarried. Third and higher order marriages are combined into one remarriage category. In the fertility domain, states indicate whether or not an individual had at least one child. Children are further distinguished by whether or not a child born in a specific age year was born out-of-wedlock, within the first seven months of a marriage or born within eight+ months of marriage. Combined these domains result in ten different states: (S) single, (M) married without children, (Cw) having a child out of wedlock, (MC) married with children, (MC7) married with a child born within seven months of marriage or married within a year of having a child out of wedlock, (D(C)) divorced (with or without child), (DCw) having a child out of wedlock while divorced, (W(C)) widowed (with or without children), (M2+(C)) remarried (with or without children), (M2+C7) remarried with seven-month child. As additional final state (11), a person can have died (DT).

Respondents can experience any number of these 11 states. They can also move back and forth between some states, for example they can marry, divorce, remarry, and divorce again. The lowest number of states a respondent can be in is one: this is an individual who stays unmarried and has no children during the entire 33 years of observation (age range 18-50).

Each respondent follows an individual sequence of states. As *n* individual sequences cannot be meaningfully interpreted, a (dis-)similarity matrix is calculated that compares how individual sequences resemble each other. The most commonly used method is optimal matching (OM) that calculates distances between sequences based on the costs of insertions, deletions, and substitutions needed to turn one sequence into another (Abbott, 1995). Early applications of OM have been criticised in social science literature (Aisenbrey & Fasang, 2010; Barban & Billari, 2012) as for instance the process of insertions, deletions and substitutions lacked a linkage with theory and the transformation costs were arbitrary, the validation of distinct groups of similar sequences was weak, unequal sequence length due to missing or incomplete data contributed to distance measurement, and the timing and order in sequences was not accounted for. In response to these critiques, new technological implementations of OM and alternative measures have been developed and have increasingly been used in the social science (see Aisenbrey & Fasang (2010) for an overview). We use an alternative method proposed by Elzinga (2007) which is based on the longest common subsequence (LCS) to calculate the distance matrix[1]. This method estimates the similarity of pairs of sequences by finding the longest common subsequence for each pair of sequences and taking the length of the sequences into account when transforming the similarity into a dissimilarity (for the calculation see Elzinga, 2007). A common subsequence is a sequence that contains similar states in the same order in a pair of sequences. Thereby, states can be deleted to derive a common subsequence. For example, the following sequences a and b share two common subsequences:

a: S – M – MC – D
b: S – M – D

The sequence pair shares the subsequence S-M, but also the longer subsequence S-M-D that represents the longest common subsequence of the pair and this subsequence would be used to calculate the dis(similarity) measure.

The advantage of this approach is that it is not required to attribute costs for the different states, that it takes into account the order of events occurring and that it is intuitive: The bigger the longest common subsequence of a pair of sequences (corrected for the length of the pair of sequences), the more similar this pair is (Barban & Billari, 2012). In a way, LCS disregards small dissimilarities and emphasizes the most common order and timing of states. For our context, where we want to explore which are the most common life trajectories in the fertility and marriage domains - with a special emphasis on when a child was born around the marriage as it could signify "deviance" or "off-time" in the demographic sense that can have repercussions for other life domains - this method is therefore appropriate.

In the second part of the analysis, we use the (dis-)similarity matrix to combine sequences into groups of family-life trajectories or clusters with the hierarchical clustering method 'Ward' (for further information on cluster analysis see for example Kaufman & Rousseeuw, 2005). Sequences are thus fused into successively larger clusters by calculating the total sum of squared deviations from the mean of a cluster. In the case of sequence analysis, the squared deviation is expressed in terms of pairwise distances and is obtained by using an analogy of the general formula (Studer et al., 2011). Generally, each clustering step aims to minimize the increase in the error sums of squares. As we are interested in gender differences, we ran analyses separately for men and women. A four-cluster solution for both men and women appeared to be the optimal number of clusters according to the visual analysis of the dendrograms which give a graphical representation of the data's hierarchical clustering structure. These clusters are then described by state frequency plots that give the percentage of the different states per year over the observed period. For the analysis, the TraMineR package (version 1.8) of the statistical software R was used (Gabadinho, Ritschard, Müller, & Studer, 2011).

In the third step of the analysis, we test whether early risk factors such as familial criminality are linked with a particular family-life cluster. The general effects are likely to be small as these clusters comprise different demographic behaviors, i.e. childbearing and union formation, and clusters may be heterogeneous, i.e. combining different types of family-life trajectories. We therefore also include significant effects at the p<0.1 level. We run a multinomial logistic regression and control for clustering at the family level, because our individuals are nested within 141 families and our outcome variable, family-life cluster, is categorical.

Family risk factors

Table 1 gives an overview of the family factors that may influence the family-life cluster that we analyse in the third step of the analysis. We take into account both demographic and offending variables in our multivariate analysis. Three bivariate variables capture the demographic behavior of the parents: (a) *parents divorced,* (b) *mother had a child out of wedlock* and (c) *mother had a child within seven months of marriage* (meaning that she became pregnant outside of marriage). Parental divorce was coded as '1' when the parents divorced before a respondent turned 18 years, and otherwise '0'. Similar demographic indicators were constructed for the offspring in the sequence analysis. Furthermore, we constructed the categorical variable *birth cohort*. Generally, the distribution of births followed a slightly left-skewed bell shape with most births concentrated around the average year of birth 1932. The majority of the sons and daughters were born between 1921 and 1940[2]. Due to this uneven distribution, we chose not to employ five-year or decennial cutoff points (used by Statistics Netherlands or also in the study by Liefbroer & Dijkstra 2000) to capture period effects, but rather chose to include the following largely even sized four birth cohorts: *<= 1925, 1926-1930, 1931-1940, >= 1941*.

Table 1: Overview of family factors

	Sons (n=259)			Daughters (n=263)		
	M	SD	Range	M	SD	Range
Father never convicted	0.54	0.50	0 – 1	0.61	0.49	0 – 1
Father convicted, no prison	0.12	0.33	0 – 1	0.11	0.31	0 – 1
Father served prison term	0.34	0.48	0 – 1	0.28	0.45	0 – 1
Parents divorced	0.14	0.35	0 – 1	0.20	0.40	0 – 1
Mother had a child out of wedlock	0.09	0.29	0 – 1	0.11	0.31	0 – 1
Mother had a child within seven months of marriage	0.40	0.49	0 – 1	0.32	0.47	0 – 1
Birth year <= 1925	0.19	0.39	0 – 1	0.23	0.42	0 – 1
Birth year 1926-1930	0.27	0.45	0 – 1	0.26	0.44	0 – 1
Birth year 1931-1940	0.37	0.48	0 – 1	0.35	0.48	0 – 1
Birth year >= 1941	0.17	0.38	0 – 1	0.16	0.37	0 – 1
(Serious) Juvenile delinquency	0.15	0.33	0 – 1	0.03	0.17	0 – 1

Criminal offending variables were retrieved from archives and judicial records. Only eight of the mothers were convicted for any offence and because of this low number we did not include their offending in the analysis. For our multivariate analysis, we constructed the following categorical variable: (a) *father never convicted*, (b) *father convicted, no prison*, and (c) *father served a prison term*, as imprisonment is more 'visible' to outsiders than a conviction and likely carried greater stigma. We also take into account sons' and daughters' own *juvenile delinquency* (age 12-17). For daughters, we include all juvenile offences as prevalence is very low, and for sons, we include only serious juvenile offences. This is done as minor juvenile offences are not uncommon for male adolescents, thus only the more severe will likely signal deviance. Given that we study offending through criminal justice contacts and not self-reported delinquency, our offending measures constitute the lower limit of actual criminal behaviour.

Results

Setting the context: Family-life transitions in comparison to the general Dutch population born between 1921 and 1940

A first inspection of the data shows that our sample of low socio-economic status differed as expected from the average Dutch population in their family formation patterns (table 2). The age of marriage and parenthood was lower than in the general Dutch population for both men and women. As they were on average more poorly educated, this could be expected, but the differences are quite large: two to three years for men and three to four years for women. For both men and women, marriage and parenthood were, as in the general Dutch population, closely linked. For 50% of the Dutch population, a first child was born within 17 months of marriage. Our sample had a much higher percentage of weddings where the woman was already pregnant when she married. This suggests that risk-taking behaviour such as early sexual activity was more common in our sample and that in the case of a pregnancy, it was absolved by getting married. In both the general Dutch population and in our sample, a birth outside marriage was a rare event. In the Netherlands until the mid-1980s fewer than 5% of the children were born outside marriage (Statistics Netherlands 2013).

Table 2: Comparison of family-life indicators between the high-risk sample and general Dutch birth cohorts born between 1921 and 1940 (in percent)

	Men			Women		
		birth cohort			birth cohort	
	Offspring sample	1921-30	1931-40	**Offspring sample**	1921-30	1931-40
Age when 50% experienced first marriage	**24.5**	27.6	26.4	**21.8**	25.0	24.3
Age when 50% experienced birth first child	**25.7**	29.1	28.3	**23.3**	27.0	25.9
Not married by age 35	**20**	15	12	**9**	10	10
Childless by age 40	**27**	17	16	**17**	16	11
Average number of children	**2.8**	2.7	2.5	**2.9**	2.9	2.7
Ended first marriages after 20ys (widowhood and divorce)	**24**	8	10	**24**	12	12
Ended first marriages after 20ys (divorce only)	**20**	-	-	**16**	-	-
No differentiation by gender:						
Time between marriage & childbirth for 25% of cohort				**4 months**	10 months	10 months
Time between marriage & childbirth for 50% of cohort				**11 months**	17 months	16 months

Sources: TransFive and Liefbroer and Dykstra 2000

Men more often remained unmarried (20%) or childless (27%) compared to the general male Dutch population of similar birth cohorts. By contrast, the high-risk daughters did not differ from similar birth cohorts in the general Dutch population: 9% remained unmarried and 17% childless. Thus, the high-risk sons were more often excluded from certain transitions than the daughters. Also, in our sample, the divorce rates for men were more than twice as high as among the general population, for women they were twice as high. Both men and women had approximately 2.8 children, similar to the average Dutch population.

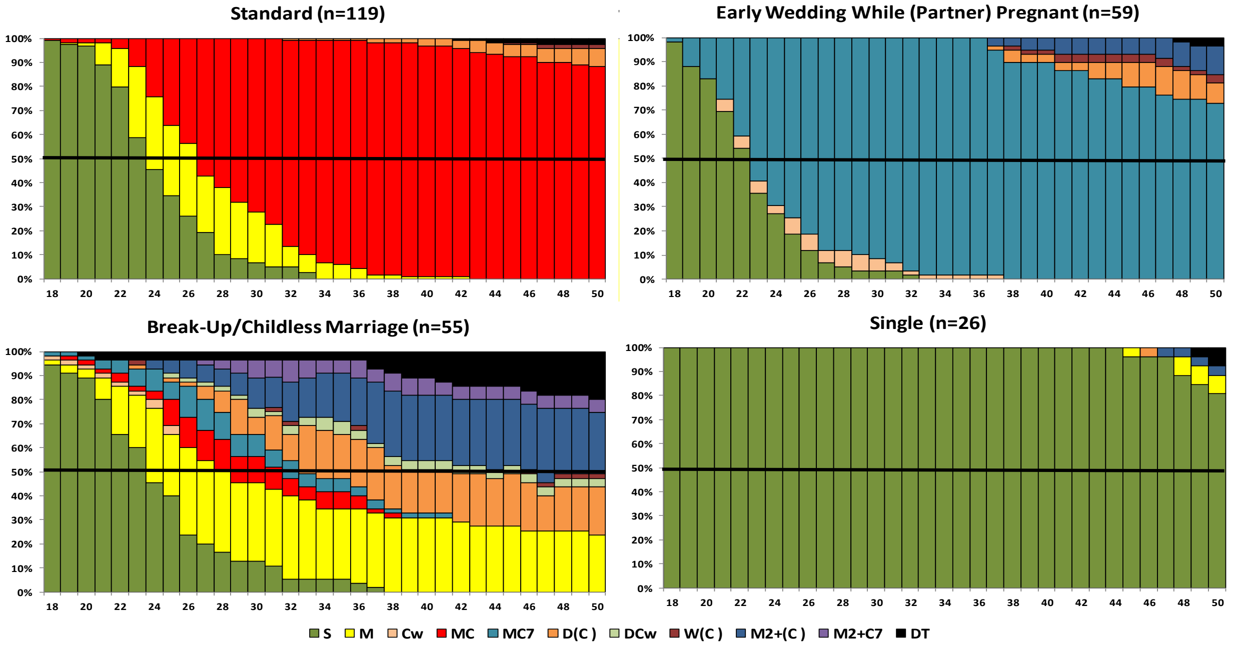
Description of the family life courses

We constructed family life courses with the help of sequence analysis to describe not only whether demographic events occurred but also in which order and timing these events occurred in the lives of our sample members. In order to describe whether some demographic behaviour concentrated in some part of the sample, the different individual life course sequences were grouped into clusters. In the following section, we summarise the most common points of these clusters in terms of timing or demographic transitions occurring. Figures 1 and 2 give a graphical presentation of the clusters by showing the percentage of each of the possible 11 behavioural states by age year.

The four clusters for men and women are fairly similar in their grouping of life courses. Although they are not identical, the ordering of events is comparable for men and women and the labels are therefore similar as well. The clusters were named *standard, early wedding while (partner) pregnant, break-up/childless marriage* and *single* for men as well as *standard*, *early wedding while pregnant, break-up* and *single/late childless marriage* for women.

Almost half, i.e. 46% of the men and 48% of the women followed a fairly standard family-life trajectory. They married, had one or more children born at least eight months after their wedding and stayed married for a substantial part of their life. Among the sons who fall into the standard cluster, 66% were married and more than half of them already had children by age 25. In the daughters' standard cluster, 85% were married and two thirds of them had their first child by age 25. Divorce occurred only for a few in these standard family-life clusters. Marriage was thus characterised by "till death do us part". These two clusters resemble the highly standardised life courses that prevailed until 1965. Thus it is not surprising that almost half of our sample belong to these clusters. Although these transitions were in the "standard" order, the transitions into marriage and parenthood did, as stated, occur earlier than among the general Dutch population.

Figure 1: State frequency plots by cluster showing the percentage of all 11 states for ages 18 to 50 years for sons

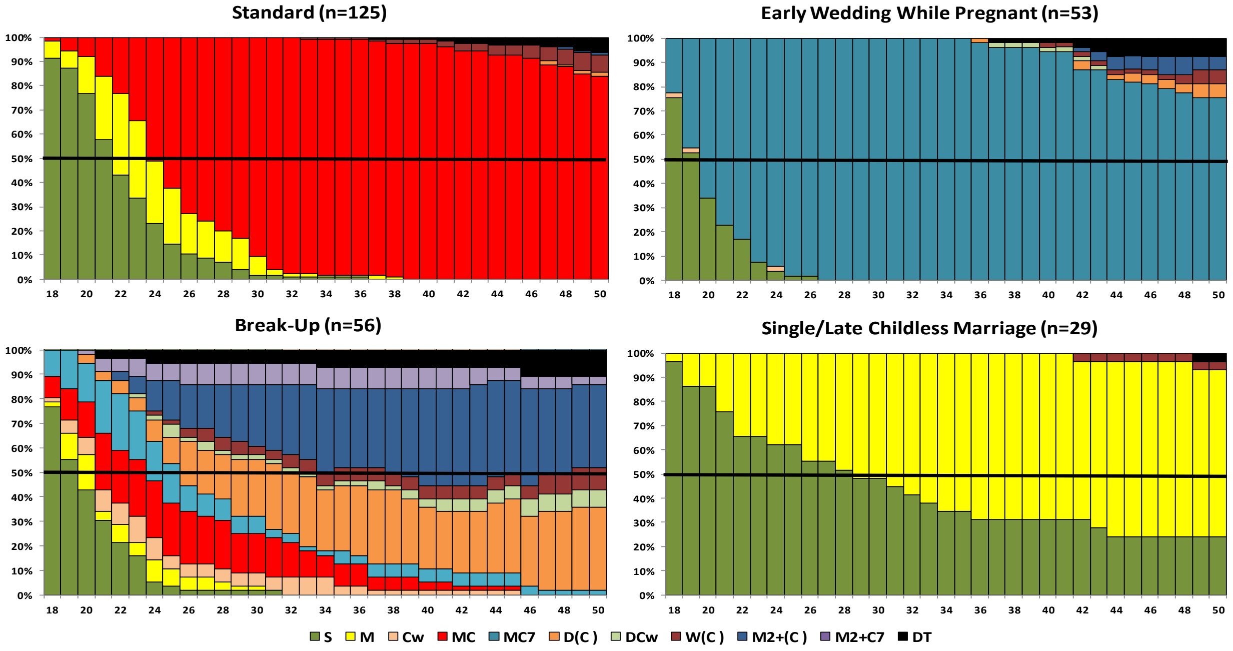

Standard (n=119)

Early Wedding While (Partner) Pregnant (n=59)

Break-Up/Childless Marriage (n=55)

Single (n=26)

Figure 2: State frequency plots by cluster showing the percentage of all 11 states for ages 18 to 50 years for daugthers



Legend: (S) single, (M) married without children, (Cw) having a child out of wedlock, (MC) married with children, (MC7) married with a child born within seven months of marriage, (D(C)) divorced (with or without child), (DCw) having a child out of wedlock while divorced, (W(C)) widowed (with or without children), (M2+(C)) remarried (with or without children), (M2+C7) remarried with seven-month child, (DT) died

Slightly more than half of the men and women in our high-risk group followed nonstandard family-formation patterns. Firstly, 23% of men and 20% of women fall into what we labeled the 'early wedding while (partner) pregnant' and 'early wedding while pregnant' clusters. By age 25, 75% of these men were married with a seven-month child and 7% had acknowledged a child born out of wedlock. Women married especially young in this cluster. By age 25, 96% of the women were married with a seven-month child. Starting in their late 30s and early 40s, the marriages in these two clusters began to break up due to divorce, widowhood and death (30% men, 25% women). Some men and women remarried. Those few who had a child out of wedlock married quickly afterwards and then remained in a stable relationship. Dutch cohorts born between 1921 and 1940 also had a substantial share of individuals with a short duration between marriage and childbirth, but this behavior was much more prevalent among our sample.

Among the men, 21% were grouped into the 'break-up/childless marriage' cluster and among the women, 21% into the 'break-up' cluster. Individuals in these clusters generally experienced a break-up of a first relationship. For men, the most common states were divorce, remarriage, out-of-wedlock parenthood, widowhood and early death but also childless marriage. By age 25, 40% of the men were still single, 25% were married without child; the remainder was comprised of all kinds of other states. By age 50, the most common categories were married without children (22%), divorced (24%), remarried (31%), and 20% had died. For women, the break-up trajectory is equally heterogeneous. By age 25, the most common states were remarried (23%), married with child (21%), married with seven-month child (16%), divorced (11%) and child out of wedlock (12%, not married or after a divorce). By age 50, all women were either divorced, remarried, widowed or had died. The common denominator of the individuals in this cluster is the absence of the long stable marriage periods observed for the two previous clusters.

The final family-life clusters differed slightly for men and women. For men, the final cluster was comprised of the 10% of the sample that remained single. The cluster also included a few late marriages (age 40 years and older). Women in the final cluster also either remained single or they married late and had no children, but among the 11% of women who followed this trajectory more married: by age 25, 62% of the women were single compared to 100% of the men. By age 50, 24% were single compared to 81% of the men.

The influence of family risk factors on the type of family-life trajectory

As a next step, the influence of various family risk factors on belonging to a specific family-life cluster is examined. Table 3 and 4 give the results of the multinomial logistic regression analysis. We compare belonging to the standard family-life cluster versus being assigned to the wedding while (partner) pregnant, break-up/late childless marriage, single cluster for men and early wedding while pregnant, break-up, and single/late childless marriage cluster for women.

Table 3: Results of the multinomial logistic regression of family-life cluster for sons

	-- cluster 'standard' is reference outcome --					
	Wedding while (partner) pregnant		Break-up/childless marriage		Single	
	B	SE	B	SE	B	SE
Convicted father, never prison	-0.15	0.52	0.05	0.71	0.90	0.71
Convicted father, prison term	0.02	0.40	0.05	0.42	0.48	0.45
Serious juvenile delinquency	1.06*	0.47	0.92#	0.50	1.20*	0.53
Parents ever divorced	0.97#	0.53	1.50**	0.52	0.76	0.76
Mother had a child out of wedlock	0.89	0.60	-0.18	0.53	0.61	0.78
Mother had a child born within 7 months of marriage	0.34	0.37	0.78*	0.39	0.40	0.46
Birth year <1925	Ref.		Ref.		Ref.	
Birth year 1926-1930	0.59	0.47	0.11	0.54	-1.64	1.16
Birth year 1931-1940	0.16	0.48	0.34	0.49	0.72	0.68
Birth year >1941	0.41	0.56	0.29	0.55	1.11	0.77
Constant	-1.49***	0.45	-1.65***	0.47	-2.68***	0.74
Number of persons	259					
x2(def)	62.97(27)					
adjusted for 116 family clusters						

**p<0.01, * p<0.05, # p<0.1

Table 4: Results of the multinomial logistic regression of family-life cluster for daughters

	-- cluster 'standard' is reference outcome --					
	Early wedding while pregnant		Break-up		Single/ late childless marriage	
	B	SE	B	SE	B	SE
Convicted father, never prison	-0.62	0.85	1.54**	0.56	2.19**	0.82
Convicted father, prison term	0.16	0.41	0.32	0.51	0.37	0.64
Serious juvenile delinquency	0.71	0.96	0.28	0.84	0.41	1.37
Parents ever divorced	1.29*	0.53	1.42**	0.53	0.36	0.68
Mother had a child out of wedlock	-2.11#	1.14	-0.92	0.91	0.45	1.02
Mother had a child born within 7 months of marriage	-0.06	0.43	0.59	0.46	1.08#	0.57
Birth year <1925	Ref.		Ref.		Ref.	
Birth year 1926-1930	1.74***	0.53	0.48	0.46	0.41	0.60
Birth year 1931-1940	1.04*	0.49	0.20	0.45	-0.36	0.59
Birth year >1941	1.12#	0.66	1.08*	0.52	-0.02	0.79
Constant	-2.00***	0.52	-1.86***	0.44	-2.42***	0.53
Number of persons	263					
x2(def)	60.75(27)					
adjusted for 113 family clusters						

**p<0.01, * p<0.05, # p<0.1

For men, offending of the father (independent of whether or not he was incarcerated) does not influence the likelihood of belonging to a certain family-life cluster, but men's own early criminal career is linked to their later family-life trajectory (table 3). Having a juvenile conviction increases the likelihood of belonging to one of the "nonstandard" family-life trajectories compared to belonging to the standard one. The effect was strongest for the single men cluster. In a step-wise modeling approach, we saw that the direct effect of father's offending on son's likelihood to belong to the single cluster disappeared once we took the son's own delinquency into account[3]. Among the parental demographic variables, we find that parental divorce is the strongest predictor for belonging to the early wedding while pregnant and break-up clusters - but not for staying single. If the mother had been pregnant when she married, the son had a higher likelihood of belonging to the break-up/childless marriage cluster compared to belonging to the standard cluster.

A somewhat different picture emerges for women (table 4). Having a criminal father, who did not serve a prison term, significantly increases the likelihood of belonging to the break-up or the single/late childless marriage cluster compared to belonging to the standard cluster. By contrast, a father who was incarcerated does not increase the likelihood of belonging to a particular cluster. In addition, for women, juvenile delinquency has no influence on cluster membership. Observing the demographic behaviour of the parents, we find - similar to the sons - that parental divorce increases the likelihood of belonging to the early wedding while pregnant and break-up clusters. Daughters whose mother had a child out of wedlock are less likely to belong to the early wedding while pregnant cluster compared to the standard family-life cluster, whereas daughters whose mother had a child within seven months of marriage are more likely to belong to the single/late childless marriage cluster. However, these findings were only significant at the p<0.1 level.

Generally, we see that parental divorce is the best predictor for both men and women to have a "nonstandard" family-life trajectory, in which also divorce occurs often. Other parental demographic behaviour is not that clearly transmitted in our study, i.e. sons and daughters are not more likely to belong to the early wedding while pregnant cluster when their mother had a child within seven months of marriage and similarly out-of-wedlock childbirth of the mother is not associated with belonging to the early wedding or the break-up cluster.

Conclusion and discussion

The aim of this study was to examine the influence of family criminality and other family risk factors on children's family life courses.

In a first step, we found that the daughters in our Dutch offspring sample of low socioeconomic status born on average in 1932 followed the demographic behaviour of the average Dutch population more than the sons. The sample differed from the Dutch population born in similar birth cohorts in that women and men married younger, had more often married when the woman was already pregnant, and had elevated divorce rates. Men in addition remained comparatively often single and childless.

As demographic behaviours are often interlinked or may cluster within certain individuals, we visualized and explored sequences of family-life transitions from ages 18 to 50 years. We identified four behavioural clusters that were fairly similar for men and women. A standard pattern of marriage followed by childbirth and long stable marriages was present in slightly less than half of the sample. The other half of the sample had non-normative family-life trajectories. A particularity of the sample was that a substantial part had their children very early and within less than seven months of their marriage, and while their relationships were long-lasting they were at increased risk to end in divorce. Another share experienced an early break-up of a marriage. The final clusters were single for men and single/late childless marriage for women.

In the next part of the analysis, we examined the link between parental offending and the identified groups of family-life sequences. How can the findings be interpreted in light of the proposed mechanisms: a direct mechanism and two indirect ones where parental offending influences children's demographic behaviour via juvenile offending or via parent's own demographic behaviour?

We found no evidence of a direct effect of father's offending on son's family-life trajectories. The findings seem to support more an indirect mechanism via juvenile offending. A son's own juvenile deviance, as reflected by his serious offending, is a consistent predictor of a "nonstandard" family-life course. Juvenile delinquency was previously shown to be linked to other deviant behaviours, for example early parenthood and marriage (Eggelston-Doherty, Green, & Ensminger, 2012). As such, we find many juvenile offenders in the 'early wedding while (partner) pregnant' family-life trajectory. Also, serious juvenile delinquency may make men less attractive marriage partners: men who had been seriously delinquent as an adolescent were more likely to remain single.

For daughters, the findings differed. Father's offending was found to have a direct influence on the family-life trajectories of their daughters. Daughters whose fathers offended but were not removed from the parental home through imprisonment were more often present in the break-up cluster and in the single/late childless marriage cluster. This indicates that exposure to a criminal father generated that risk. The finding allows different interpretations: socialisation into deviant behaviours, escaping a difficult household situation by marrying an unsuitable partner and stigma preventing family formation.

The daughters were probably – like the sons – socialised through their father's offending behaviour to reject certain conventional norms and behaviours, and they may have been more likely to accept divorce as a solution to marital conflict. Another explanation for the gender difference may be that girls were more restricted to the home environment than boys and were more exposed to and affected by a present criminal father. Some studies have found negative outcomes for children when a violent or drug addicted parent is present compared to a parent who was absent and incarcerated (Finlay & Neumark 2010; Jaffee, Moffitt, Caspi, & Taylor, 2003; Wildeman, 2010). Girls may also have seen marriage as the fastest route to escape a difficult house situation (see also Foster & Hagan, 2007) and may have ended up with too hastily chosen partners whom they divorced afterwards. Similarly, another share may have avoided marriage or foregone childbearing because of their difficult childhood home as described in Reading and Amatea's (1986) review of the psychological literature.

Another reason for the effect of fathers' offending on daughters' demographic behaviour may be that women in the period under study were less able to escape their family background by building up their lives through for example employment. This would explain why having a delinquent father increased the odds most for remaining single or marrying late or having short relationships: Women from such a stigmatised background may – like men - have been less sought-after partners. What argues against this explanation is that having a father who went to prison – likely generating a greater stigma – did not increase the likelihood for such behaviour.

Although these explanations are tentative and need additional study, our findings indicate that a father's offending affects the family-life of his children, but that the mechanisms differ for sons and daughters. While fathers' offending does not affect sons' demographic outcomes directly but appears to do so through transmission of criminality, for daughters there does appear to be a direct effect of a father's criminality on her family-life course.

Our study has strong and weak points. A strong point is that the sample allows the modeling of entire family-life courses for a group with low socioeconomic status and at high risk of offending, in which relations are possible to emerge. A possible critique is the choice we made concerning the distinction between having a child before seven months of marriage or after eight months of marriage in the definition of states in the sequences analysis. This adds to the distinctive clusters for early wedding while pregnant and standard clusters, placing individuals in the states (MC and MC7) for long periods of time, although the event that decided this state is "long past". However, we believe it is important to account for the timing of childbirth, because it can drive the timing and duration of marriage. Without including the timing of childbirth in combination with the marriage trajectory, we could not holistically construct the family life trajectories. The timing difference early in life seems to be the main difference between these cluster types (and in some cases the earlier break-up of the early wedding marriages). Still our findings also highlight that the birth timing is important. We find these two distinct groups, because the two groups are apparently common among our sample. In addition, the results of the multinomial regression suggest that there are underlying differences between the groups. For sons, we see the higher level of juvenile delinquency and for daughters a parental divorce effect for the early wedding clusters versus the standard clusters.

A weak point of our study is that we did not have any information beyond parental criminality, juvenile offending, and adult social roles. Thus, some of the associations we found may be spurious. Further studies should attempt to incorporate information on personality traits, early life events and contexts, such as aggression, impulsivity, and neighborhood. Another issue is the historical nature of the sample. One may question what relevance our findings have for current generations. This is, however, inherent to research where one wants to study individuals over a long time. Related to this is also the fact that our sample may be considered special in that the sons and daughters entered adulthood in a period with exceptionally standard life courses. However, as we compared individuals within the sample, the internal validity of our conclusions should not be affected. Our choice to study family-life courses from ages 18 to 50 – the age range where most family-life transitions occur – has its drawbacks too. One consequence is that our effect sizes are small – with numerous life events combined into long periods, effects are bound not to have emerged as sharply as when we would have investigated single events close in time. Nevertheless, this approach allowed us to model sequences and combinations of behaviour that otherwise would not have emerged and that was associated with different risk factors. If we had analysed single demographic transitions, we would have shown that the sample under study is at an increased risk of divorce, early marriage etc. Our analysis, however, was able to show that for a sizeable part of our sample, none of these off-time or negative transitions occurred and that almost half of our sample followed standard life courses, thus presenting a more nuanced view of the demographic careers of these high-risk men and women.

Further research is needed to study the mechanisms in more detail. Also, it would be worthwhile to study whether the found effects are also present in current cohorts that grew up in times where large parts of the population experienced a de-standardisation of the family-life course, although the length of the life course studied would be limited in current cohorts.

To sum up, our study illustrated the far-reaching consequences of parental offending. While it has been well-documented that parental criminality is transmitted intergenerationally, our study showed how parental crime – in tandem with their demographic behaviour – affects the demographic trajectories of their children – and possibly eventually also the lives of their grandchildren.

References

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology, 21,* 93-113. http://dx.doi.org/10.1146/annurev.so.21.080195.000521

Aisenbrey, S., & Fasang, A. E. (2010). New Life for Old Ideas. The 'Second Wave' of Sequence Analysis - Bringing the Course Back into the Life Course. *Sociological Methods & Research*, *38*(3), 420-462. http://dx.doi.org/10.1177/0049124109357532

Amato, P. R. (2000). The consequences of divorce for adults and children. *Journal of Marriage and the Family, 62,* 1269-1287. http://dx.doi.org/10.1111/j.1741-3737.2000.01269.x

Anderson, E. (1999). Code of the street: Decency, violence, and the moral life of the inner city. New York: W. W. Norton & Company.

Barban, N., & Billari, F. C. (2012). Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61*(5), 765-784. http://dx.doi.org/10.1111/j.1467-9876.2012.01047.x

Bijleveld, C. C., & Wijkman, M.D. (2009). Intergenerational continuity in convictions: A five-generation study. *Criminal Behaviour and Mental Health, 19,* 142-55. http://dx.doi.org/10.1002/cbm.714

Bijleveld, C. C., Wijkman, M. D., & Stuifbergen, J. A. (2007). 198 Boefjes? De maatschappelijke integratie van moeilijk opvoedbare jongens 100 jaar geleden. [198 Rascals? The societal reintegration of difficult boys 100 years ago]. NSCR report.

Block, C. R., Blokland, A. A., Van der Werff, C., Van Os, R., & Nieuwbeerta, P. (2010). Long-term patterns of offending in women. *Feminist Criminology, 5,* 73-107. http://dx.doi.org/10.1177/1557085109356520

Braman, Donald. 2004. *Doing Time on the Outside: Incarceration and Family Life in Urban America.* Ann Arbor: University of Michigan Press.

Bras, H., Liefbroer, A. C., & Elzinga, C. H. (2010). Standardization of pathways to adulthood? An analysis of Dutch cohorts born between 1850 and 1900. *Demography, 47,* 1013-1034. http://dx.doi.org/10.1007/BF03213737

Burt, S. A., Barnes, A. R., McGue, M., Iacono, W. G. (2008). Parental divorce and adolescent delinquency: Ruling out the impact of common genes. *Developmental Psychology, 44,* 1668-1677. http://dx.doi.org/10.1037/a0013477

Comfort, M. (2007). Punishment beyond the legal offender. *Annual Review of Law and Social Science, 3,* 271-96. http://dx.doi.org/10.1146/annurev.lawsocsci.3.081806.112829

Dronkers, J., & Härkönen, J. (2008). The intergenerational transmission of divorce in cross-national perspective: Results from the Fertility and Family Surveys. *Population Studies, 62,* 273-88. http://dx.doi.org/10.1080/00324720802320475

Edin, K. (2000). Few good men: Why poor mothers don't marry or remarry. *American Prospect, 11,* 26–31.

Eggelston-Doherty, E., Green, K. M., & Ensminger, M. E. (2012). The impact of adolescent deviance on marital trajectories. *Deviant Behavior, 33*, 185-2006. http://dx.doi.org/10.1080/01639625.2010.548303

Elder, G. H. Jr. (1998). The life course as developmental theory. *Child Development, 69,* 1-12. http://dx.doi.org/10.1111/j.1467-8624.1998.tb06128.x

Elder, G. H. Jr. (1994). Time, human agency, and social change: Perspectives on the life course. *Social Psychology Quarterly, 57,* 4-15. http://dx.doi.org/10.2307/2786971

Elzinga, C. H. (2007). Sequence analysis: Metric representations of categorical time series. Manuscript. Department of Social Science Research Methods, VU University Amsterdam.

Elzinga, C. H., & Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population, 23,* 225-50. http://dx.doi.org/10.1007/s10680-007-9133-7

Farrington, D. P., Coid, J. W., & Murray, J. (2009). Family factors in the intergenerational transmission of offending. *Criminal Behaviour and Mental Health, 19,* 109–124. http://dx.doi.org/10.1002/cbm.717

Fergusson, D. M, Horwood, L. J., & Lynskey, M. T. (1992). Family change, parental discord and early offending. *Journal of Child Psychology and Psychiatry, 33,* 1059-1075. http://dx.doi.org/10.1111/j.1469-7610.1992.tb00925.x

Finlay, K., & Neumark, D. (2010). Is Marriage always good for children? Evidence from families affected by incarceration. *Journal of Human Resources, 45*, 1046-1088. http://dx.doi.org/10.1353/jhr.2010.0026

Foster, H., & Hagan, J. (2007). Incarceration and intergenerational social exclusion. *Social Problems, 54,* 399-433. http://dx.doi.org/10.1525/sp.2007.54.4.399

Furstenberg, F. F., Levine, J. A., & Brooks-Gunn, J. (1990). The children of teenage mothers: Patterns of early childbearing in two generations. *Family Planning Perspectives, 22*, 54–61. http://dx.doi.org/10.2307/2135509

Fussell, E., & Furstenberg, F. F. (2005). The transition to adulthood during the twentieth century: Race, nativity, and gender. In R. A. Settersten, F. F. Furstenberg, & R. G. Rumbaut (Eds.), *On the frontier of adulthood: Theory, research, and public policy* (pp. 29-75). Chicago: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226748924.003.0002

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software, 40,* 1-37. http://dx.doi.org/10.18637/jss.v040.i04

Gilman, S. E., Kawachi, I., Fitzmaurice, G., & Buka, S. L. (2003). Socioeconomic status, family disruption and residential stability in childhood: Relation to onset, recurrence and remission of major depression. *Psychological Medicine, 33,* 1341-355. http://dx.doi.org/10.1017/S0033291703008377

Hissel, S., Bijleveld, C. C., & Kruttschnitt, C. (2011). The well-being of children of incarcerated mothers: An exploratory study for the Netherlands. *European Journal of Criminology, 8,* 346-360. http://dx.doi.org/10.1177/1477370811415755

Horwitz, S. M., Klerman, L. V., Kuo, H. S., & Jekel, J. F. (1991). Intergenerational transmission of school-age parenthood. *Family Planning Perspectives, 23,* 168–177. http://dx.doi.org/10.2307/2135740

Jaffee, S. R., Moffitt, T. E., Caspi, A., & Taylor, A. (2003). Life with (or without) father: The benefits of living with two biological parents depend on the father's antisocial behavior. *Child Development, 74*, 109-126. http://dx.doi.org/10.1111/1467-8624.t01-1-00524

Kaufman, L. & Rousseeuw, P. J. (2005). Finding groups in data: An introduction to cluster analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.

Koropeckyj-Cox, T., Pienta, A. M., & Brown, T. H. (2007). Women of the 1950s and the normative life course: The implications of childlessness, fertility timing, and marital status for psychological well-being in late midlife. *International Journal of Aging and Human Development, 64*, 299-330. http://dx.doi.org/10.2190/8PTL-P745-58U1-3330

Liefbroer, A. C., & Dykstra, P. A. (2000). *Levenslopen in verandering: een studie naar ontwikkelingen in de levenslopen van Nederlanders geboren tussen 1900 en 1970, WRR serie*. [Changing life courses: A study over the development in the life courses of Dutch persons born between 1900 and 1970] Den Haag: Sdu Uitgevers. http://www.wrr.nl/fileadmin/nl/publicaties/DVD_WRR_publicaties_1972-2004/V107_Levenslopen_in_verandering.pdf

McLanahan, S. (2004). Diverging destinies: How children are faring under the second demographic transition. *Demography, 41,* 607-627. http://dx.doi.org/10.1353/dem.2004.0033

McLeod, J. D., & Kessler, R. C. (1990). Socioeconomic status differences in vulnerability to undesirable life events. *Journal of Health and Social Behavior, 31*, 162-172. http://dx.doi.org/10.2307/2137170

Murphy, M. (1999). Is the relationship between fertility of parents and children really weak? *Social Biology, 46,* 122-145. http://dx.doi.org/10.1080/19485565.1999.9988991

Murphy, M., & Knudsen, L. B. (2002). The intergenerational transmission of fertility in contemporary Denmark: The effects of number of siblings (full and half), birth order, and whether male or female. *Population Studies, 56,* 235-248. http://dx.doi.org/10.1080/00324720215937

Murray, J., & Farrington, D. P. (2008). The Effects of Parental Imprisonment on Children. *Crime and Justice, 37,* 133-206. http://dx.doi.org/10.1086/520070

Nesmith, A., & Ruhland, E. (2008). Children of incarcerated parents: Challenges and resiliency, in their own words. *Children and Youth Services Review*, *30*, 1119 – 1130. http://dx.doi.org/10.1016/j.childyouth.2008.02.006

Neugarten, B. L., & Hägestad, G. O. (1976). Age and the life course. In R. H. Binstock & E. Shanas (Eds.), *Handbook of Aging and the Social Science* (pp. 35-55). New York: Van Nostrand-Reinhold.

Oesterle, S., Hawkins, J. D., Hill, K. G., & Bailey, J. A. (2010). Men's and women's pathways to adulthood and their adolescent precursors. *Journal of Marriage and the Family, 72,* 1436-1453. http://dx.doi.org/10.1111/j.1741-3737.2010.00775.x

Osgood, D. W., Foster, E. M., Flanagan, C. A., & Ruth, G. R. (2005). *On your own without a net: The transition to adulthood for vulnerable populations*. Chicago: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226637853.001.0001

Phillips, S., & Gates, T. (2011). A conceptual framework for understanding the stigmatization of children of incarcerated parents. *Journal of Child and Family Studies, 20,* 286-294. http://dx.doi.org/10.1007/s10826-010-9391-6

Pogarsky, G., Lizotte, A. J., & Thornberry, T. P. (2003). The delinquency of children born to young mothers: Results from the Rochester Youth Development Study. *Criminology, 41,* 101-138. http://dx.doi.org/10.1111/j.1745-9125.2003.tb01019.x

Pogarsky, G., Thornberry, T. P., & Lizotte, A. J. (2006). Developmental outcomes for children of young mothers. *Journal of Marriage and the Family, 68*, 332-34. http://dx.doi.org/10.1111/j.1741-3737.2006.00256.x

Ramakers, A. A., Bijleveld, C. C., & Ruiter, S. (2011). Escaping the family tradition: A multi-generation study of occupational status and criminal behaviour. *British Journal of Criminology, 51,* 856-874. http://dx.doi.org/10.1093/bjc/azr039

Reading, J., & Amatea, E.S. (1986). Role deviance or role diversification: Reassessing the psychosocial factors affecting the parenthood choices of career-oriented women. *Journal of Marriage and the Family, 48*, 255-260. http://dx.doi.org/10.2307/352392

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Divergent pathways and turning points through life*. Cambridge, MA: Harvard University Press.

Settersten, R. A. (2003). Age structuring and the rhythm of the life course. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the Life Course* (pp. 81-98). New York: Kluwer Academic/Plenum Publishers. http://dx.doi.org/10.1007/978-0-306-48247-2_4

Sigle-Rushton, W. (2005). Young fatherhood and subsequent disadvantage in the United Kingdom. *Journal of Marriage and the Family, 67,* 735-753. http://dx.doi.org/10.1111/j.1741-3737.2005.00166.x

Skardhamar, T., & Lyngstad, T. H. (2009). Family formation, fatherhood and crime. An invitation to a broader perspective on crime and family transitions. Statistics Norway, Research Department: Discussion Papers No. 579. http://www.ssb.no/a/publikasjoner/pdf/DP/dp579.pdf

Statistics Netherlands (2013). Birth; age mother (on 31 December), birth order and fertility rates. Table selection gives the total births of all marital and non-marital live-born children between 1950 and 2012 for the Netherlands. Accessed 11th December 2013 from http://statline.cbs.nl/StatWeb/publication/?DM=SLEN&PA=37744ENG&D1=0,5,10&D2=0&D3=a&LA=EN&VW=T

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research, 40*, 471-510. http://dx.doi.org/10.1177/0049124111415372

Svarer, M. (2011). Crime and partnerships. *Review of Economics of the Household, 9*, 307–325. http://dx.doi.org/10.1007/s11150-010-9104-3

Thornberry, T. P. (2005). Explaining multiple patterns of offending across the life course and across generations. *The ANNALS of the American Academy of Political and Social Science, 602,* 156-195. http://dx.doi.org/10.1177/0002716205280641

Tonry, M., & Bijleveld, C. C. (2007). Crime, criminal justice, and criminology in the Netherlands. In M. Tonry & C. C. Bijleveld (Eds.), *Crime and justice in the Netherlands. Crime and Justice, A review of research, Volume 35* (pp. 1-30). Chicago: University of Chicago Press. http://dx.doi.org/10.1086/650192

Van de Rakt, M., Nieuwbeerta, P., & De Graaf, N. D. (2008). Like father, like son? The relationships between conviction trajectories of fathers and their sons and daughters. *British Journal of Criminology, 48,* 538–556. http://dx.doi.org/10.1093/bjc/azn014

Wildeman, C. (2010). Paternal incarceration and children's physically aggressive behaviors: Evidence from the Fragile Families and Child Wellbeing Study. *Social Forces, 89*, 285-310. http://dx.doi.org/10.1353/sof.2010.0055

Zoutewelle-Terovan, M., Van der Geest, V. R., Liefbroer, A. C., & Bijleveld, C. C. (2014). Criminality and family formation: Effects of marriage and parenthood on criminal behavior for men and women. *Crime and Delinquency, 60*(8), 1209-1234. http://dx.doi.org/10.1177/0011128712441745

Endnotes

[1] Barban & Billari (2012) showed that LCS and OM return fairly similar results.

[2] Less than 5% of the offspring sample was born before 1921, 13% was born 1941-1950, and 3.5% was born after 1951.

[3] As this was the only effect that changed, we decided to only show the full models. In another analysis we ran a path analysis contrasting those falling into the "standard" clusters versus all other clusters. There we found a similar effect: father's criminal involvement increased chances of juvenile delinquency for sons but did not significantly influence their family-life course. This additional analysis is available upon request from the corresponding author.

Can Rose's paradox be useful in delinquency prevention?

Mogens Nygaard Christoffersen SFI – The Danish National Centre for Social Research, Denmark
mc@sfi.dk
Heather Joshi UCL Institute of Education, UK

(Received November 2013 Revised May 2014) http://dx.doi.org/10.14301/llcs.v6i4.274

Abstract

Geoffrey Rose's prevention paradox obtains when the majority of cases with an adverse outcome come from a population of low or moderate risk, and only a few from a minority 'high risk' group. Preventive treatment is then better targeted widely than on the 'high risk' minority. This study tests whether the prevention paradox applies to the initiation of criminal behaviour, as recorded in longitudinal administrative data from Denmark. Children born in 1984 are followed from birth to early adulthood. A discrete-time Cox model allows for changing covariates over time. The initiation of criminal behaviour is defined as getting a police record between the ages of 15 and 22 as a result of a criminal matter. This outcome was predicted, more accurately than by chance, by a combination of over twenty risk factors, reflecting the major crime reduction paradigms. However, it seems impossible to identify a minor group (<5%) in the population from whom criminals are exclusively recruited. Our example illustrates how the applicability of Rose's prevention strategy, population based, rather than targeted, depends on how narrowly 'high-risk group' is defined, for a given distribution of estimated risk, and allows for the possible complementarity of population and targeted measures.

Keywords

Birth Cohort, Criminal Behaviour, Juvenile Delinquency, Prevention, Life Course, Childhood Risk Factors, Register data.

Introduction

It is often found that a small group of individuals is responsible for a sizeable proportion of all offending activity. Studies in Britain and Denmark find that about 6% of offenders account for more than half of offending activity (Piquero, Farrington, & Blumstein, 2007; Kyvsgaard, 2002). They imply that the targeting of crime prevention on these high-risk individuals at an early stage, before they have started their criminal career, could bring large benefits to the community, as well as the individuals themselves.

This supports an approach to crime prevention which focuses on the few who are most likely to become involved in breaking the law. Such a narrow focus strategy can be contrasted with the 'population strategy', based on Geoffrey Rose's prevention paradox (Rose, 1992). Rose drew attention to some situations where the majority of cases come from the population at low or moderate risk and only a minority from the high risk group. Targeting the minor high risk group in that case may be ineffective.

Rose's prevention paradox does not always apply. In the health field, for example, a certain liver cancer (*haemangiosarcoma*) can only be caused by exposure to Vinyl chloride monomer and asbestosis can only be caused by exposure to asbestos. In these cases it is possible to locate a specific small group accounting for nearly all the cases. The prevention strategy is simply to minimize the number of persons who get exposed to Vinyl chloride monomer, or asbestos, respectively.

Seat belts in cars provide an example where Rose's prevention paradox is appropriate. Seat belts, the preventive measure, reduce the number of casualties effectively. It is impossible to isolate those with a high risk; a seat belt has to be used every time by everybody in a car, although the risk is very low. The preventive measure spread throughout the population brings large benefits to the community but offers little to each participating individual because their risk is low (Rose, 1992). Rose argues that the population strategy is appropriate when a small causal risk involves a large number of people and it seems difficult to identify a minor group in the population from whom most of the cases are drawn.

In the present paper, the applicability of Rose's paradox for crime prevention is explored in a national total sample of 54,000 children born in 1984 in Denmark, and followed into adulthood. In order to locate a high risk group we estimate the hazards of being registered by the police for criminal behaviour. Other administrative registers provide indicators of major risk factors in parenting, location and individual resource deficits. Risk factors from birth to adulthood are used to estimate the hazards of getting a police record of criminal behaviour between the ages of 15 and 22.

Theories of crime prevention

The rationale for identifying juvenile delinquents who may graduate to crime is articulated in the White Paper from the Danish Ministry of Justice supporting early intervention (Justitsministeriet, 2009). Research on delinquency reduction asks if it can be predicted whether a given adolescent is likely to become a delinquent. The question has been debated over the last fifty years. In the 1940s, 50s and 60s efforts were made to identify delinquents or to spot potential delinquents (mainly among boys). The first studies were based on retrospective data looking back to childhood where many of the factors did not become operative until later in the boys' lives. The purpose was to identify delinquents in advance of any manifestation of criminal behaviour. Childhood factors included 'discipline of boy by father', 'supervision of boy by mother', 'affection of parents for boy', and 'family cohesiveness' (Glueck, 1962; Glueck, 1963). David Farrington and colleagues found that having a father who had been arrested, a young mother, or a bad neighbourhood were links in the causal chain leading to boys' delinquency (Farrington, Jolliffe, Loeber, Stouthamer-Loeber, & Kalb, 2001). Psychometric instruments for identifying youth at risk of delinquency were constructed and evaluated (Loeber, Dishion, & Patterson, 1984; Glueck, 1950). A review by Loeber and Dishion (1983) found that the best predictors of criminal behaviour were reports of the child's stealing, lying, or truancy, with the child's problem behaviour close behind. Parents' family management (supervision and discipline), the child's conduct problems, parental criminality, and the child's poor academic performance were other principal predictors of delinquency. Murray and Farrington find in a review of prospective longitudinal studies that the most important risk factors are impulsiveness, low IQ, low school achievement, poor parental supervision, punitive or erratic parental discipline, cold parental attitude, child physical abuse, parental conflict, antisocial parents, large family size, low income, high delinquency rates in schools, and neighbourhoods (Murray & Farrington, 2010). Most of the studies were based on predicting crime in a group of young people who already had manifested criminal behaviour (Olver, Stockdale, & Wormith, 2009; Leschied, Chiodo, Nowicki, & Rodger, 2008; Loeber et al., 1984).

The research questions

In the present study we want to predict who becomes a criminal, defined as those who are recorded by the police for criminal behaviour by age 22. We construct a screening instrument for individual young people derived from measures collected before any criminal behaviour is recorded in police records, to address the following questions:

1 How well can the method predict criminal behaviour for both boys and girls?

2 How many of the predicted children will actually be involved in criminal behaviour within the follow-up period?

3 How many of the young people who do get into trouble with the police will actually come from the ’high risk‘ group?

4 What level of estimated risk should count as ‘high’?

The answers to these questions may help us to choose between a crime reduction programme based on universal measures (Rose’s population strategy) and one focussed on the ’high-risk’.

The next section of the paper reviews theories of crime prediction and prevention. This is followed by sections on data sources and methods and results. The discussion section includes a reminder of the research questions, before a brief conclusion. The appendices give details of the register data sources, supplementary regression results and the statistical model.

Theories of crime prediction and prevention

The strategy of crime prevention focussed on ‘high risk’ individuals involves screening and early treatment for those at the extreme of the risk distribution. An important aim of this crime prevention strategy is a reduction in the risk factors under the assumption that the risk factors represent a causal link between risk factors and outcome. Per-Olaf Wikström emphasizes the importance of addressing the lack of integration of levels of explanation of how environmental and individual conditions interact (Wikström, 2006). Predictors or risk factors are chosen on the basis of four major paradigms each with its own explanation of crime and potential relevance to crime reduction in primary and secondary crime prevention theories (Hope, 2000; Soothill, Christoffersen, Hussain, & Francis, 2010).

1. Primary crime prevention theory focuses on universal measures to reduce delinquency without reference to individual characteristics. Such a ‘population strategy’ aims at reducing crime by interventions directed at the general population. One paradigm (1.a.) emphasizes the importance of the current situation and opportunities as the most essential factors rather than the individual’s background (Clarke, 1980). Prevention should focus on the setting itself that may prompt, provoke, pressure, or permit an individual to offend (Cornish & Clarke, 2003; Clarke, 1997). The other primary paradigm (1.b.) links criminal behaviour to localities or neighbourhoods and only to a lesser extent to individual characteristics (Sampson, Morenoff, & Raudenbush, 2005; Wikström, 1998). Segregation differentially exposes members of disadvantaged groups to violence and looser informal community controls. This theory implies that generic interventions to improve neighbourhood conditions and support families may reduce violence in the locality, and that moving individuals out of a high risk area will in itself reduce their risk of offending.

2. Secondary crime prevention targets specific subgroups of the population believed to be at greater risk than others - a ‘high risk’ approach. Here again we have two main paradigms: Paradigm (2.a) focuses on developmental theories, parental child-rearing methods and disadvantages during adolescence as the background for deviant behaviour (Farrington & Welsh, 2007; Farrington, 1994; Loeber & Blanc, 1990). Paradigm (2.b) focuses the decision-making processes of young people at high risk of delinquency, explaining delinquency by resources such as their own lack of education, poverty, unemployment, or unstable family status. According to this paradigm, criminality can be seen as a rational behaviour, one among several possibilities to increase income (Becker, 1968). Under this fourth paradigm the prevalence of crime depends on: the possibilities of illegal compared to legal income; the risk of being caught; the severity of the punishment in case of conviction; the possibilities of legal employment; individual willingness to expose oneself to risks and preferences for crime; and the amount of their social capital – reputation, employment, marriage (Williams & Sickles, 2002). Within this paradigm, crime prevention effort focuses on the options of high-risk groups and how to influence their rational choices.

3. Tertiary crime prevention aims to truncate the criminal career and deals with the treatment of known offenders (Pease, 1997). This is further down the line than the point on which we focus – the first-time contact with the police before the criminal career has taken any further steps. Hence the present study seeks inspiration from the theories of primary and secondary crime prevention. These four paradigms are not mutually exclusive, but constitute our frame of reference for selection of potential risk factors to predict crime in a prospective longitudinal study.

Causality

If Rose's prevention paradox is appropriate for forming a crime prevention strategy, there has to be a causal link between effective population intervention measures and criminal activities. There have been several comprehensive reviews on such evidence (for example: Farrington & Welsh, 2007). These measures include: changing parenting practices (Olds et al., 1998); changed environments in preschools (Schweinhart et al., 2005); peer tutoring or mentoring in schools (Hahn, 1999; Welsh, 2007); or school training programs (Pfiffner, R.A.Barkley, & G.J. DuPaul, 2006); and anti-bullying programs in schools (Olweus, 2005). These measures have demonstrated causality via effective prevention in randomized control trials. At the present stage, not all paradigms have been supported when they were implemented as delinquency prevention strategies. Measures targeting localities/neighbourhoods have not yet delivered the expected results in U.S.A. (Welsh & Hoshi, 2006).

Prospective longitudinal studies offer the best way to study the predictors of delinquent and criminal behaviour (Murray, Farrington, & Eisner, 2009). A review by Farrington, Ohlin, & Wilson, (1986) found eleven prospective longitudinal surveys with information about crime and delinquency based on samples of at least hundreds of persons. Liberman (2008) found over 60 longitudinal data sets, and more have been published since. Two thirds of the studies were from the U.S., the rest came from ten other developed countries, including Denmark. One in four studies included males only.

Data

We use longitudinal data assembling indicators of risk factors for a complete cohort of all children born in 1984 (N=27,840 boys and 26,618 girls) in Denmark. The children are followed from birth to early adulthood in 2006. A criterion for participation was that the children were resident in Denmark on 1 January 1998 at 14 years of age. Adolescents known to have emigrated or died were censored at the last person year they appeared in the records. The risks are estimated from birth until they first get a police record or until early adulthood. The register includes individual risk factors such as living in a disadvantaged area, parental circumstances and behaviour, and individual resource deficits recorded for the birth cohort from an early age and in early adulthood (table 1 and appendix A).

The risk factors and outcome variables were chosen on four criteria: 1) A theoretically grounded choice based on the crime paradigms set out above, and on prior empirical evidence. 2) Predictions should rest on a non-biased population-wide base. 3) The risk factors should be registered in the administrative archives. 4) With these constraints on data availability, the outcome was chosen as the event of first getting a record in the police register of criminal behaviour under the Penalty Code. Someone appears in the police register if they are either charged or confined under the criminal Penalty Code, see appendix A for details. The criterion used indicates the event of embarking on what may turn into a 'criminal career'. This measure is not a true measure of crime, because some of those with records, say of arrest only, will not be convicted. Equally, some of those who have committed crimes will not have been brought to the notice of the police. However, this indicator is treated as proxy for criminality, subject to these caveats and we sometimes use this term in what follows where we do not explicitly remind readers that the police register is our source.

Table 1. Information selected from the population-based registers used in the Danish cohort study

Register	Variables	Years included
Police archives	Arrest, pre-trail detention, charges of crime under the Penalty Code	1999-2006
Population statistics	Gender, age, marital status, address	1980-2006
Medical register on vital statistics	Cause of death, suicide	1979-2006
Employment statistics	Unemployment, branch of trade, occupation	1980-2006
Housing statistics	Ownership, number of rooms,	1980-2006
Education statistics	School achievements, education, vocational training	1981-2006
Social assistance act statistics	Children in care, preventive care	1977-2006
Crime statistics	Violation, adjudication, imprisonment	1980-2006
Income compensation benefits	Social benefit, duration	1984-2006
Income statistics tax register	Income	1980-2006
Fertility Database	No. of siblings, parity, link to parents	1980-2006
National inpatient register	ICD-8/10 diagnoses (somatic)	1977-2006
National psychiatric register	ICD-8/10 diagnoses (psychiatric)	1979-2006

Note: information in registers includes both children and parents.

Administrative registers, linked together via personal identity numbers, have the advantage, over survey data, of smaller reporting biases. The data they harvest has three positive attributes:

1) They are registered prospectively - that is, information gathered in calendar year 't+1' has no influence whatsoever on data filed in calendar year 't': these register data are not subject to back -filling, even if later information reveals errors, they are not corrected in the files available for research.

2) Data are provided independently from a number of agencies, who have no knowledge of each other's entries.

3) They have complete coverage of all calendar years from their birth in 1984 until 2006, when the cohort reaches age 22.

One drawback to register data is that they only provide information known to the authorities, not for example attitudes or abilities of parents or children, psychiatric disorders not requiring admission to hospital, unreported domestic violence or undetected offences. Another is that registers are not immune from error. Registers known to be particularly unreliable, according to internal reliability tests or a few external reliability assessments, have not been used here.

Method

The analysis proceeds in three stages. First, variables were selected as potential predictors on the basis of relevance to the theories outlined above, from the more reliable sources, as listed in appendix A. Second, the predictive value of these risk factors was tested in model 1, which includes all risk factors selected on a priori grounds at the first stage. It indicates if any of them prove to be redundant. In the third stage, model 2 drops the risk factors whose estimates were not significant in model 1. The improvement of prediction is estimated by the Hosmer & Lemeshow Goodness-of-fit test (table 2). We use a maximum likelihood method to estimate a discrete-time Cox regression model (Allison, 1982). A similar method has been used in other crime risk studies (Christoffersen, Soothill & Francis., 2003; Soothill et al., 2010). See appendix B for details of the statistical model.

The discrete-time Cox model was chosen to allow for changing covariates over time. The risk factors are divided into three types of time co-variation, according to the number of years for which they are introduced. Type I risk factors identify the presence of that factor in the year before the outcome may occur, for example, living in a disadvantaged area when the subject was 18 will act as a risk factor when the subject is 19 – the following calendar year, and being there at age 19 would be a risk factor for age 20. Type I factors are reversible. Thus, a move out of a disadvantaged area at 20 would affect prediction in for the following year, age 21. Type II risk factors are time-varying, in that they are introduced in the year when they first occur, but once 'switched on' are irreversible, applying for all subsequent years. Family separation, for example is one of the factors assumed to have such a lasting effect. Type III risk factors are those that are taken to indicate a permanent condition throughout the risk period from age 15, for example, if the child didn't ever pass lower secondary ('basic') level this is taken to be an indicator of permanent poor performance.

Results

Among the birth cohort of 54,458 individuals 11.2% (or 6,075) had experienced at least one contact with the police (arrest, confinement or charge under the Penalty Code) between the ages of 15 and 22. This represents 17.0% of the males and 5.0% of the females.

Among the risk factors, whose mean person years are listed in the second column of table 2, parental mental illness during childhood was registered during 11% of person years from 15 to 22. Registered violence to or by parents in the childhood home apply to 8% of the years under observation. In about 60% of the person years, one of the parents had experienced at least one year with more than 21 weeks of unemployment up to that point. During the window when the children were 15 to 22 years old, nearly 38% of person-years had been preceded by a family separation at some point. While these risk factors are examples of relatively common incidence, other predictors are rare. Only 2% of the person-years from age 15 had lived in a disadvantaged area. Child abuse and neglect are registered for 3.8% of the person-years. The focus-child having ever been in social care covers 7.1% of the person years. Less than 1% of the person-years had followed a parental conviction according to the criminal code (Type I).

Risk factors

As shown in table 2, most of the 25 potential risk factors, selected for model 1 on theoretical grounds turned out to be highly significant predictors of getting a police record. Although the effect sizes may be modest for the individual risk factors, the total picture may be predictive. Four turned out to be redundant: parental suicidal behaviour, parental substance abuse, poverty during the young person's upbringing and the young person's hospitalisation for psychiatric disorder. If these factors do have an influence, it is absorbed into other risk factors. Odds ratios greater than 1 confirm a positive association with crime. The estimates were not affected by excluding the redundant covariates. Model 2 showed that parental background factors such as domestic violence, parental mental illness, child abuse and neglect, child in (public) care, and family separation all contributed independent information to the prediction of criminality. Structural factors such as parental vocational qualifications and parental unemployment also contributed to the explanation model, as did the young person living in a disadvantaged area or rented housing, and other indictors of a resource deficit in the young person's 'human capital': low education, youth poverty and youth unemployment. Other variables retained in model 2 pertaining to the young person's behaviour, with relatively high odds ratios, were substance abuse (alcohol or drugs) and attempted suicide (ORs 1.95 and 1.60 respectively). Convictions of the mother, though very rare (0.3% of person years), were strong predictors of their offspring's later police contacts with an odds ratio of 1.76. Convicted fathers were not quite as rare (0.7%) and were less strongly associated (OR 1.34). Up to a point, the results conform to the notion that it is the rarer risk factors which have the higher risks, e.g., child abuse and neglect - 3.8 % of person-years and odds ratio of 1.86. However the risk ratios attaching to childhood or adolescent adversity are dwarfed by the relative risk of being male, where the odds of being registered for criminal behaviour is three times higher than for females.

Table 2: Estimated prognoses for the first-time crime events (arrest, confinement or charged according to the penalty code) the following year. Discrete time Cox model

			Model 1				Model 2			
	Type	% of person-years	Estimate	SE Standard Error		Odds ratio	Estimate	SE Standard Error		Odds ratio
Constant term			-4.56	0.07			-4.55	0.07		
20 year			-0.89	0.06	***		-0.88	0.06	***	
19 year			-0.56	0.05	***		-0.56	0.05	***	
18 year			-0.21	0.05	***		-0.21	0.05	***	
17 year			-0.20	0.04	***		-0.20	0.04	***	
16 year			-0.17	0.04	***		-0.17	0.04	***	
Parental background:										
Parental inpatient mental illness	III	11.2	0.09	0.04	*	1.10	0.10	0.04	*	1.10
Parental substance abuse	III	6.9	0.04	0.05	Ns	1.04			Ns	
Parental suicidal behaviour	III	2.9	-0.05	0.06	Ns	0.95			Ns	
Parental violence	III	8.4	0.39	0.04	***	1.48	0.39	0.04	***	1.48
Non-Danish	I	7.2	0.48	0.04	***	1.62	0.49	0.04	***	1.63
Mother has no vocational qualification	I	70.3	0.14	0.04	***	1.15	0.14	0.04	***	1.15
Father has no vocational qualification	I	76.7	0.24	0,04	***	1.28	0.24	0.04	***	1.28
Parental unemployment >21 weeks	II	60.9	0.32	0.03	***	1.37	0.32	0.03	***	1.37
Poverty (<40% of median income)	III	20.3	0.04	0.04	Ns	1.04			Ns	
Child abuse and neglect	II	3.8	0.61	0.05	***	1.84	0.62	0.04	***	1.86
Family separation	II	37.7	0.43	0.03	***	1.55	0.44	0.03	***	1.55
Mother teenager	II	3.8	0.26	0.05	***	1.30	0.26	0.05	***	1.30
Mother convicted	I	0.3	0.56	0.13	***	1.75	0.57	0.13	***	1.76
Father convicted	I	0.7	0.29	0.10	*	1.34	0.29	0.10	*	1.34
Location:										
Rented housing (not self-owner)	I	30.9	0.16	0.03	***	1.17	0.16	0.03	***	1.18
Disadvantaged area	I	2.1	0.24	0.07	***	1.28	0.24	0.07	**	1.27
Individual resources										
Basic secondary only	III	2.0	0.40	0.06	***	1.50	0.41	0.06	***	1.51
Not in process of training or education	I	15.6	0.12	0.04	**	1.12	0.12	0.04	*	1.13
Not graduated from high school	III	66.8	0.57	0.05	***	1.76	0.57	0.05	***	1.76
Current poverty (< 50 % of median level)	I	9.6	0.18	0.05	**	1.20	0.20	0.05	***	1.23
Own Unemployment >21 weeks	I	1.1	0.46	0.10	***	1.58	0.46	0.10	***	1.59
Focus child ever in care	II	7.1	0.34	0.04	***	1.41	0.36	0.04	***	1.43
Substance abuse (alcohol, drugs)	II	1.9	0.66	0.07	***	1.94	0.67	0.07	***	1.95
Own attempted suicide	II	0.9	0.44	0.10	***	1.56	0.47	0.10	***	1.60
Own in-patient mental illness	II	3.5	0.11	0.06	Ns	1.12			Ns	
Gender (1=boy; 0=girl)		48.3	1.34	0.03	***	3.83	1.34	0.03	***	3.28

Notes for Table 2

Number of first-time events n=6,075. Total number of individuals in the study = 54.458, while the total number of person-years = 300,591.

* p < 0.05; ** p < 0.001; *** p < 0.0001. Ns Non-significant.

Type of time-dependency

Type I: exposed to risk factor at time t then the risk factors is also present at t+1.

Type II: exposed to risk factor at time t then risk factor is also present at all the following years.

Type III: risk factor observed for at time t it also covers the years before and after the years under investigation.

The Hosmer & Lemeshow Goodness-of-fit test shows that prediction capability is increased when using model 2 instead of model 1. Model1: Chi-square 9.32; DF 8; Pr<0.32 while Model 2: Chi-square 8.62; DF 8; Pr<0.38).

Detailed definitions of the variable in Appendix A

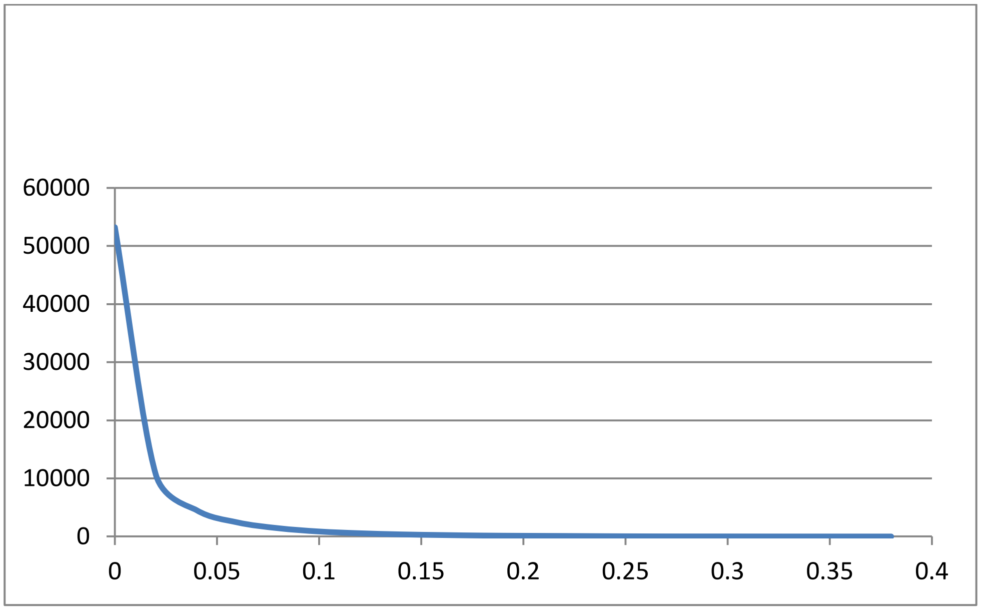
Source: Table 2 is based on: Register data and PolSas, Police Archive data (Christoffersen, Skov Olsen, Vammen, Sander Nielsen, Lausten, & Brauner, 2011).

These estimates are presented for males and females together, with just this substantial intercept to distinguish them. It might be expected that the relationship with risk factors would be sex-specific. To investigate this we estimated a fully interacted version of model 1 allowing the 25 parameters to vary by sex. Most of them (20 out of 25) were identical for men and women, see appendix C. Four risk factors were more predictive for girls than boys: parental mental illness, the mother having been convicted, the young person having attempted suicide or not being in training. On the other hand, other things equal, boys who left secondary school without qualification showed particularly high levels of criminality. Since gender differences are mainly accounted for by the binary intercept, we proceed with a pooled model, rather than a separate one for females, where offences are quite sparse. Although the girls would be particularly outnumbered in any 'high risk' group it might not be possible to exclude them from interventions.

Another presumption is that many young people with a police record will have just one episode of offending and none further (Moffitt, 1993). To investigate this we estimated the odds ratio, in the original cohort, for a second contact with the police, and for third, fourth and fifth contacts. We do not estimate transitions to a higher number of contacts conditional on having reached the previous level since our aim is to explore early interventions based on information about risk factors from before the first contact. Contrary to the assumption that one-time-only offenders have a different risk profile to those who repeat, the parameters estimated for a second contact with the police were similar to that of a first. However the risks did rise slightly for having the third or further registered encounters with the police. This suggests that the minority of 'hard core recidivists' were somewhat more strongly associated with some risk factors than those with one or two, as shown in appendix D. Effect sizes (Odds ratios) increased slightly for boys, for young people who had been in care, who had experienced a violent childhood (domestic violence, abuse and neglect), who were non-Danish, and for those with poor school performance.

Figure 1 shows the number of people and their estimated risk of being placed on the police criminal register between ages 15 and 22. No-one is estimated to have a risk over 0.4, in fact very few over 0.2. Many are estimated to have very low risks. In trying to predict which of the young people will eventually get a police record of criminal behaviour we need to specify a level of risk that characterises the target.

Figure 1: Exposure in the population / Number of people with expected risk



Note: the horizontal axis is the estimated risk levels 0 to 0.40

Table 3: Classification over probability cut-point, estimated expected risk of criminality, and observed criminality in all the years under observation

	Numbers				Percentages	
Cut-point	Criminality:					
Probability level	Observed Expected	Not observed Not expected	Not observed Expected	Observed Not expected	False positive[1] %	False-negative[2] %
			False positive:	False negative:		
	(a)	(b)	(c)	(d)	c/(a+c)%	d/(b+d) %
0.00	6075	0	47154	0	88.6	-
0.02	4371	40833	6321	1704	59.1	4.0
0.04	2673	45295	1859	3402	41.0	7.0
0.06	1631	46359	795	4444	32.8	8.7
0.08	1052	46778	376	5023	26.3	9.7
0.10	681	46970	184	5394	21.3	10.3
0.12	467	47060	94	5608	16.8	10.6
0.14	317	47104	50	5758	13.6	10.9
0.16	223	47123	31	5852	12.2	11.0
0.18	146	47132	22	5929	13.1	11.2
0.20	108	47137	17	5967	13.6	11.2
0.22	83	47142	12	5992	12.6	11.3
0.24	51	47148	6	6024	10.5	11.3
0.26	28	47150	4	6047	12.5	11.4
0.28	17	47151	3	6058	15.0	11.4
0.30	16	47152	2	6059	11.1	11.4
0.32	10	47153	1	6065	9.1	11.4
0.34	7	47153	1	6068	12.5	11.4
0.36	5	47153	1	6070	16.7	11.4
0.38	3	47153	1	6072	25.0	11.4

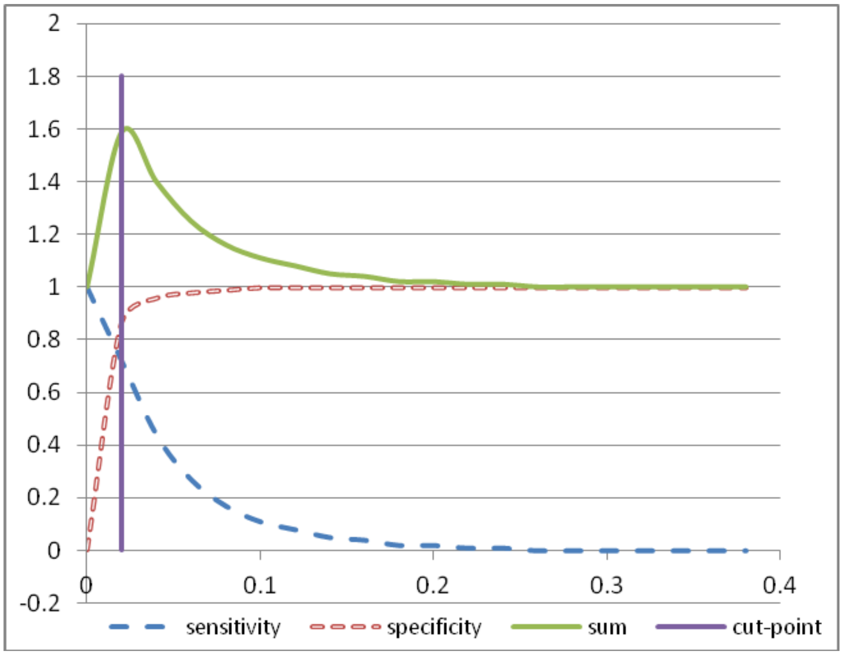
Maximisation of sensitivity and specificity

One aim would be to maximize 'sensitivity' - the proportion of those registered ('observed') with a police record who are correctly predicted by the model to have one ('expected'). We also aim to maximize 'specificity' - the proportion of those with no police record who are correctly predicted to have no record. To operationalize sensitivity and specificity, we need to set a level of estimated probability sufficiently high to constitute a positive prediction. No individual has an estimated certainty of getting a police record with probability of 100%. Assuming a cut-point to distinguish between cases and non-cases in a target group, we can compute the sensitivity and specificity, as plotted in figure 2. The analytical tool (Receiver Operating Characteristic, ROC) maximises both these measures simultaneously (Woodward, 1999). It turns out that the cut-point at 0.02 maximizes the sum of these two measures. Prediction using the cut-point 0.02 (one in fifty) is much better than chance (P<0.0001). Table 3 shows it predicts 10,692 individuals out of the total birth cohort of 54,458 to have entered police criminal records. This will correctly assign 4,371 out of 6,075 of those observed in the register (71.9%), the' true positives'. Unfortunately, it also includes 6,321 'false positives', that is those with no police record who were predicted as having one – this is a majority (59.1 per cent) of all 'expected' cases. Thus this cut-point will also include some moderate and low risk individuals. As the cut-point rises, the number of cases falsely assigned to the criminal category declines, but so does sensitivity (see figure 2). By the time the cut-point reaches around 0.2 (one in five) , there are very few 'positives', either true or false. Almost all of the criminals would be 'false negatives'. Attempting to characterize a very high risk minority would miss a lot of offenders.

Rose's paradox

This illustrates Rose's paradox that efforts to prevent criminal behaviour aimed at screening and treatment of individuals at 'very high risk' are likely to have limited population impact, if the majority of cases do not occur in the minority high-risk group. The likelihood of finding empirical support for the prevention paradox rests upon the relative size of the minority high-risk group (Romelsjö & Danielsson, 2012). Table 4 operationalizes the high-risk group as reducing from 20% to 2% of the population as the cut-point rises from 0.02 to 0.10, an increasingly narrow definition of the high risk group. For example, the 5% of the population with an estimated risk of at least 0.06 will include only 27% of the people actually registered for criminal behaviour during the observation period. However, if we drop the cut-point to 0.04, extending the high-risk group to 9% of the population, nearly half of the young persons with police contact (44%) will be correctly predicted. On the other hand, about 41% of the predictions of criminality will be incorrect (false positive).

Figure 2: Sensitivity, specificity, and their sum against cut-points used to distinguish expected criminals from non-expected criminals for the data in table 3



True positive rate (sensitivity): number of observed and expected criminal persons in relation to number of observed criminals.

Specificity: number of expected and observed non-criminals in relation to number of observed non-criminals.

The vertical line represents a cut-point of 0.02

Source: Table 2, model 2

Table 4: Classification table over probability cut-point: estimated percentage of population, and observed, and expected first-time offenders. (selected cut-points)

	Percentage of population	Expected as percentage of observed first-time offenders
	(a)	(b)
0.02	20	72
0.04	9	44
0.06	5	27
0.08	3	17
0.10	2	11
Total	53,229	6,075

Source: Table 3

The cut-point which maximises the sum of sensitivity and specificity, 0.02, extends the target group to 20% of the population. It implies that around one fifth of the population (10,692), including 72% of actual offenders, might be offered a targeted intervention. Less than half the target (41% or 4,371) would have been correctly identified as criminal while 1,704 offenders (3% of the population) would be missed. Screening one fifth of the total population (on the basis of information in principle available in many registers) could be seen as more cost-effective than reaching the wider population, although targeting 20% of the population could not be called a very small minority. A universal intervention would reach all of the cases who actually ended up on the register (no false negatives), but it would also cover the 89% of the population who do not come into police contact – the' false positives'. They are analogous to the seat-belt users who never have a road accident.

Discussion

The *first research question* was whether accurate predictions of risk of offending could be made. Our analysis of a wide range of information from administrative registers clearly gives a better prediction of future criminal behaviour than chance. It was expected that girls would have a lower risk-level and also a different risk profile than boys, but the risk factors generally had the similar effect sizes in boys and girls - although boys tended to have a higher starting point. We also investigated whether 'hard core criminals' had a different childhood risk profile than the young people whose record includes only one offence. Our results indicate only small differences between the first and the second contact with the police. The *second and third research questions* concerned the sensitivity and specificity of the predictions. The answer to these questions depends on where we draw the line between high and moderate risk. The majority of criminal persons come from a population with low or moderate risk and only a minority of the criminals come from the high-risk population. The optimal cut-point (0.02) in this dataset means that 20% of the population would be targeted as 'at risk' but only 8% will be correctly identified as offenders; 12% would be false positive while another 3% of the population would be offenders who get missed (false negative). In other words, for more than half of the young people whose previous life events predicted a probability of (registered) criminal behaviour above one in fifty, there was no police record during the follow-up years. Thus, this 'low threshold' indicator of a (relatively) high risk profile apparently makes a false 'accusation' to nearly 60% of the subjects. Although it is possible that they might have engaged in delinquent behaviour which escaped police notice, these results demonstrate the problem of labelling or stigmatising young adults with a high risk profile according to administrative data. There is still a small but widespread risk in the other 80% of the population which accounts for 28% of recorded delinquency.

The *fourth research question* asked what proportion of estimated risk should count as 'high'. Other empirical studies have operationalized the minority at high risk at between 5% and 35%, although 10% is commonly reported. This exercise has focussed on a definition based on maximising the sum of sensitivity and specificity, which makes the cut at 20% of this population. We have also shown the implications of varying the high risk criterion.

What do we conclude about the suitability of a population strategy of universal measures (Rose's population strategy) rather than one focussed on the 'high-risk' individuals in a programme of delinquency reduction? We have found that the risk of criminal behaviour displays a continuum in the population. A large number of people have a small risk and give rise to more cases of criminal behaviour than a small number of people with a high risk. Geoffrey Rose found that though it is possible to focus preventive efforts on very high-risk groups these are a relatively small proportion of the population and cases (Rose, 1992). This has led to the impetus to identify the factors that may influence the population distribution of risk factors. Measures that decrease the average level of criminal behaviour will decrease the prevalence of excessive criminal behaviour according to the 'mass population strategy'.

Some p*rimary crime prevention* interventions could be recommended on a universal level, at a very early age, before, the prediction of crime is possible. Studies have shown significant crime reducing effect of family training; and home visiting nurses (Barth, Hacking & Ash, 1988; Gray, Cutler, Dean & Kempe, 1979; Olds, Henderson, Chamberlin & Tatelbaum, 1986; Olds et al. 1998; Olds, Henderson, Tatalbaum & Chamberlin, 1988). Home visiting and pre-schools are provided on a universal basis in Denmark. High Scope and similar pre-school interventions, though they tend to be targeted at vulnerable groups in USA, have been associated with a significant reduction of youth delinquency among low-income families (Berrueta-Clement, Schweinhart, Barnett, Epstein, & Weikart, 1984; Schweinhart, Montie, Xiang, Barnett, Belfield, & Nores, 2005 ; Schweinhart, Barnes, & Weikart, 1993). Likewise, some universal school programs such as peer tutoring or mentoring (Hahn, 1999; Welsh, 2007), school non-bullying programs (Olweus, 1994; Olweus, 1995; Olweus, 2005) and cohesive school programs (Gottfredson, Wilson, & Najaka, 2002) seem to be appropriate as universal programs (Farrington, 2013).

A crucial question is the huge amount of delinquency prevention measures which have no supporting evidence. The definition of a well-established treatment or intervention is that it has been compared in two or more design manualized experiments and shown to have to have significant effects over another treatment or placebo (Chambless & Ollendick, 2001). Well-established experiments should give information about costs and outcomes of treatment side effects as well as intended effects. Without this, policy makers are unaware of the possible damage and costs of the chosen intervention measures.

To consider the relative merits of population based vs high risk strategies when there are unintended side effects, consider two illustrative scenarios, not necessarily exhaustive:

- I. Criminal behaviour has devastating consequences for the individual and for society at large, and the preventive measures have no adverse side effects for the individual.
- II. Criminal behaviour has minor consequences for the individual and society and the side effects of the preventive measures have high costs for the treated persons and society.

In scenario I we would tolerate a large number of false predictions of criminality (false-positives) as side effects are minimal, though one would have to consider the cost of targeting people who did not 'need' the intervention – known as deadweight loss. In scenario II we would be less inclined to accept a high false-positive rate. It would be unethical to force or convince people to participate if the side-effects are devastating and many of those treated would not be actual criminals. In scenario II the population strategy looks less attractive

In our example, a strategy of targeting the riskiest 5% could only include about a quarter (27%) of those later observed to be criminals; we end up with a relatively high false positive rate using the administrative data to predict future criminal behaviour. The focus should be upon evidence-supported preventive measures which have little or no adverse side-effects and also measures regarded as positive by the participants.

This study has a least two important limitations. All the risk factors are correlated with the outcome, and precede it, but the study insufficiently demonstrates a causal link to the outcome and the longitudinal study needs to be combined with experimental prevention programs to test effects of interventions (Murray & Farrington, 2010). Consequently, influencing these risk factors is an uncertain crime prevention strategy.

We have also not explored any variation on the functional form of the statistical model to explore the possibility of further interactions between risk factors (Wikström, 2006) beyond those we have tested for gender. We note that the logistic model is inherently multiplicative and it is often not possible to find well-determined estimates of interactions.

The model needs to be applied to other administrative data-material, where the distribution of risk may be either more concentrated or more dispersed than in the data used here. It also needs to be supplemented by non-administrative information such as personal interviews which include questions about self-reported criminal behaviour. International comparisons may also add to our knowledge on crime prevention strategies, though few countries outside Scandinavia have such rich linked register data.

Thirdly, the present study used administrative data to predict future criminal behaviour and the results revealed some limitations in this method, but also possible guidelines for choosing between crime prevention methods and measures. In accordance with the Rose paradox we illustrate the difficulty of using the high-risk approach when predicting a low base-rate event. We found the riskiest 9% of the population accounted for nearly half (44%) of the people with a police record, but this leaves 56% of the criminals outside the high risk group.

If there were a population-wide measure or set of measures preventing crime in the same way as seat-belts protect people from injury in car accidents, the wide base from which these young people were drawn into crime would indicate it should be deployed in a Rose-style 'population' strategy. The results suggest that supporting young people gaining qualifications in or after school could be part of such a strategy. However the paradox does not mean that particular identifiable groups- such as the children of convicted or mentally ill parents, or those with mental health problems themselves- should be ignored, just that there are not enough of them upon whom to rest prevention efforts. In most of the very high risk situations males are at greater risk than females, and gender-specific interventions may be appropriate if feasible.

It is recommended that early delinquency prevention measures only include (a) measures with convincing demonstration of causal and preventive effects; (b) measures regarded as positive by the participants; and (c) should have a dual focus, targeted and universal. The 'high risk' group might be the relatively high risk group in the population from which the majority of those involved in criminal behavior originate (here, say 20%), or a higher risk and smaller minority who only account for a minority of the crime. Universal measures would in any case reach these individuals and for some of the effective early interventions, they could not be identified in advance with certainty. The results support both a selective strategy on a high risk group and a population strategy of measures lowering the low or moderate risk in the majority of the population. Future research needs to find the causal links between risk factors, criminal activities and cost-effective population intervention measures in order to lower risk across the board.

Acknowledgements

Brian Francis, Keith Soothill, Bo Vinnerljung, Gorm Gabrielsen and Gerda Engholm are thanked for their comment and advice to a previous version of the paper. The anonymous reviewers and the editor are thanked for their helpful comments.

References

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. In S.Leinhardt (Ed.), *Sociological Methodology (*61-98). San Francisco: Jossey-Bass. http://dx.doi.org/10.2307/270718

Barth, R. P., Hacking, S., & Ash, J. R. (1988). Preventing child abuse: An experimental evaluation of the child parent enrichment project. *The Journal of Primary Prevention, 8,* 201-217. http://dx.doi.org/10.1007/BF01695023

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *The Journal of Political Economy, 76,* 169-217. http://dx.doi.org/10.1086/259394

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). *Changed lives: the effects of the Perry Preschool program on youths through age 19*. Ypsilanti, Mich: High/Scope Press, cop.

Boligministeriet (1993). *Første rapport fra Byudvalget*. København: Indenrigsministeriet.

Chambless, D. L. & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52,* 685-716. http://dx.doi.org/10.1146/annurev.psych.52.1.685

Christoffersen, M. N., Soothill, K., & Francis, B. (2003). An upbringing to violence? Identifying the likelihood of violent crime among the 1966 birth cohort in Denmark. *Journal of Forensic Psychiatry and Psychology, 14,* 367-381. http://dx.doi.org/10.1080/1478994031000117830

Christoffersen, M. N., Soothill, K., & Francis, B. (2005). Who Is Most at Risk of Becoming a Convicted Rapist? The Likelihood of a Rape Conviction among the 1966 Birth Cohort in Denmark. *Journal of Scandinavian Studies in Criminology & Crime Prevention, 6,* 39. http://dx.doi.org/10.1080/14043850410000839

Christoffersen, M. N., Soothill, K., & Francis, B. (2007). Violent Life Events and Social Disadvantage: A Systematic Study of the Social Background of Various Kinds of Lethal Violence, Other Violent Crime, Suicide, and Suicide Attempts. *Journal of Scandinavian Studies in Criminology & Crime Prevention, 8,* 157-184. http://dx.doi.org/10.1080/14043850701498469

Christoffersen, M. N., Skov Olsen, P., Vammen, K. S., Sander Nielsen, S., Lausten, M., & Brauner, J. (2011). *Tidlig identifikation af kriminalitetstruede børn og unge: risiko- og beskyttelsesfaktorer*. København: SFI - Det Nationale Forskningscenter for Velfærd.

Clarke, R. V. G. (1980). "Situational" Crime Prevention: Theory and Practice. *British Journal of Criminology, 20,* 136-147.

Clarke, R. V. G. (1997). *Situational crime prevention*. (Second Edition ed.) Criminal Justice Press.

Cornish, D. B. & Clarke, R. V. (2003). Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention. *Crime prevention studies, 16,* 41-96.

Farrington, D. P. (1994). Early developmental prevention of juvenile delinquency. *Criminal Behaviour and Mental Health, 4,* 209-227.

Farrington, D. P. (2013). Longitudinal and experimental research in criminology. *Crime and Justice, 42,* 453-527. http://dx.doi.org/10.1086/670396

Farrington, D. P., Jolliffe, D., Loeber, R., Stouthamer-Loeber, M., & Kalb, L. M. (2001). The concentration of offenders in families, and family criminality in the prediction of boys' delinquency. *Journal of Adolescence, 24,* 579-596. http://dx.doi.org/10.1006/jado.2001.0424

Farrington, D. P., Ohlin, L. E., & Wilson, J. Q. (1986). *Understanding and controlling crime toward a new research strategy*. New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-4940-5

Farrington, D. P. & Welsh, B. (2007). *Saving children from a life of crime: early risk factors and effective interventions*. New York: Oxford University Press.

Glueck, E. T. (1962). Toward improving the identification of delinquents. *Journal of Criminal Law, Criminology & Police Science, 53,* 164-170. http://dx.doi.org/10.2307/1141070

Glueck, E. T. (1963). Toward further improving the identification of delinquents. *Journal of Criminal Law, Criminology & Police Science, 54,* 178-180. http://dx.doi.org/10.2307/1141159

Glueck, S. (1950). *Unravelling juvenile delinquency*. London: Harvard Univ. Press.

Gottfredson, D.C., Wilson, D.B., & Najaka, S.S. (2002). School-based crime prevention. In Sherman, L.W., Farrington, D.P., Welsh, B.C. & MacKenzie, D. (Eds.): *Evidence-based crime prevention.* London: Routledge, s. 56-164. http://dx.doi.org/10.4324/9780203166697_chapter_4

Graversen, B. K., Hummelgaard, H., Lemmich, D., & Nielsen, J. B. (1997). *Residential mobility in Danish problem housing estates*. Copenhagen: AKF Amternes og Kommunernes Forskningsinstitut.

Gray, J. D., Cutler, C. A., Dean, J. G., & Kempe, C. H. (1979). Prediction and Prevention of Child Abuse and Neglect. *Journal of Social Issues, 35,* 127-139. http://dx.doi.org/10.1111/j.1540-4560.1979.tb00805.x

Hahn, A. (1999). Extending the time of learning. In D.J.Besharov (Ed.), *America's disconnected youth: toward a preventive strategy* (pp. 233-265). Washington, D.C.: CWLA Press.

Hope, T. (2000). *Perspectives on crime reduction*. Aldershot, Hants: Ashgate.

Hummelgaard, H., Graversen, B. K., Lemmich, D., & Nielsen, J. B. (1997). *Udsatte boligområder i Danmark*. København: AKF Amternes og Kommunernes Forskningsinstitut.

Justitsministeriet (2009). *Indsatsen mod ungdomskriminalitet. Betænkning nr. 1508*. København: Justitsministeriet.

Kyvsgaard, B. (2002). *The criminal career: The Danish longitudinal study*. Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511499463

Leschied, A., Chiodo, D., Nowicki, E., & Rodger, S. (2008). Childhood Predictors of Adult Criminality: A Meta-Analysis Drawn from the Prospective Longitudinal Literature. *Canadian Journal of Criminology and Criminal Justice, 50,* 435-467. http://dx.doi.org/10.3138/cjccj.50.4.435

Liberman, A. (2008). *The long view of crime: a synthesis of longitudinal research*. New York: Springer. http://dx.doi.org/10.1007/978-0-387-71165-2

Loeber, R. & Dishion, T. (1983). Early predictors of male delinquency: a review. *Psychological Bulletin, 94,* 68-99. http://dx.doi.org/10.1037/0033-2909.94.1.68

Loeber, R. & Blanc, M. L. (1990). Toward a Developmental Criminology. *Crime and Justice, 12,* 375-473. http://dx.doi.org/10.1086/449169

Loeber, R., Dishion, T. J., & Patterson, G. R. (1984). Multiple Gating: A Multistage Assessment Procedure for Identifying Youths at Risk for Delinquency. *Journal of Research in Crime and Delinquency, 21,* 7-32. http://dx.doi.org/10.1177/0022427884021001002

Manuel, C. & Klint Jørgensen, A. M. (2013). *Systematic review of youth crime prevention interventions: Published 2008-2012*. SFI - Det Nationale Forskningscenter for Velfærd.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological review, 100,* 674-701. http://dx.doi.org/10.1037/0033-295X.100.4.674

Murray, J., Farrington, D. P., & Eisner, M. P. (2009). Drawing conclusions about causes from systematic reviews of risk factors: The Cambridge Quality Checklists. *Journal of Experimental Criminology, 5,* 1-23.

Murray, J. & Farrington, D. P. (2010). Risk factors for conduct disorder and delinquency: Key findings from longitudinal studies. *The Canadian Journal of Psychiatry, 55,* 633-642.

Olds, D. L., Henderson, C. R., Chamberlin, R., & Tatelbaum, R. (1986). Preventing Child Abuse and Neglect: A Randomized Trial of Nurse Home Visitation. *Pediatrics, 78,* 65-78.

Olds, D., Henderson, C. R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., & Powers, J. (1998). Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior. *JAMA: The Journal of the American Medical Association, 280,* 1238-1244. http://dx.doi.org/10.1001/jama.280.14.1238

Olds, D. L., Henderson, J., Tatelbaum, R., & Chamberlin, R. (1988). Improving the Life-Course Development of Socially Disadvantaged Mothers: A Randomized Trial of Nurse Home Visitation. *American Journal of Public Health, 78,* 1436-1445. http://dx.doi.org/10.2105/AJPH.78.11.1436

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2009). Risk Assessment With Young Offenders. *Criminal Justice and Behavior, 36,* 329-353. http://dx.doi.org/10.1177/0093854809331457

Olweus, D. (1994). Bullying at School: Basic Facts and Effects of a School Based Intervention Program. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 35,* 1171-1190. http://dx.doi.org/10.1111/j.1469-7610.1994.tb01229.x

Olweus, D. (1995). *Bullying at school: What we know and what we can do*. (Reprinted ed.) Oxford: Blackwell.

Olweus, D. (2005). A useful evaluation design, and effects of the Olweus Bullying Prevention Program. *Psychology, Crime and Law, 11,* 389-402. http://dx.doi.org/10.1080/10683160500255471

Pease, K. (1997). Crime Prevention. In M.Maguire, R. Morgan, & R. Reiner (Eds.), *The Oxford Handbook of Criminology* (second edition ed., pp. 963-995). Oxford: Oxford University Press.

Pfiffner, L. J., R.A.Barkley, & G.J.DuPaul (2006). Treatment of ADHD in School Settings. In R.A.Barkley (Ed.), *Attention-deficit hyperactivity disorder: a handbook for diagnosis and treatment* (3. ed. ed., pp. 547-589). New York: Guilford Press.

Piquero, A. R., Farrington, D. P., & Blumstein, A. (2007). *Key issues in criminal career research: New analyses of the Cambridge Study in Delinquent Development*. Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511499494

Romelsjö, A. & Danielsson, A.-K. (2012). Does the prevention paradox apply to various alcohol habits and problems among Swedish adolescents? *The European Journal of Public Health, 22,* 899-903. http://dx.doi.org/10.1093/eurpub/ckr178

Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *British medical journal (Clinical research ed.), 282,* 1847-1851. http://dx.doi.org/10.1136/bmj.282.6279.1847

Rose, G. (1992). *The strategy of preventive medicine*. Oxford: Oxford Medical publications.

Rose, G., Khaw, K. T., & Marmot, M. (2008). *Rose's strategy of preventive medicine*. Oxford University Press, USA. http://dx.doi.org/10.1093/acprof:oso/9780192630971.001.0001

Rossow, I. & Romelsjö, A. (2006). The extent of the 'prevention paradox' in alcohol problems as a function of population drinking patterns. *Addiction, 101,* 84-90. http://dx.doi.org/10.1111/j.1360-0443.2005.01294.x

Sampson, R. J. & Laub, J. H. (1992). Crime and Deviance in the Life Course. *Annual Review of Sociology, 18,* 63-84. http://dx.doi.org/10.1146/annurev.so.18.080192.000431

Sampson, R. J., Morenoff, J. D., & Raudenbush, S. (2005). Social anatomy of racial and ethnic disparities in violence. *American Journal of Public Health, 95,* 224-232. http://dx.doi.org/10.2105/AJPH.2004.037705

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: the High-Scope Perry preschool study through age 27*. Ypsilanti, MI: High/Scope Press.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: the High/Scope Perry preschool study through age 40*. Ypsilanti, Mich.: High/Scope Press.

Sherman, L. W. (1997a). Communities and crime prevention. In Sherman, L.W., Gottfredson, D. C., MacKenzie, D., Eck, J. E., Reuter, P. & Bushway, S. (Eds.), *Preventing crime: what works, what doesn't, what's promising*. University of Maryland at College Park: Department of Criminology and Criminal Justice University of Maryland.

Sherman, L. W. (1997b). *Preventing crime what works, what doesn't, what's promising: a report to the United States Congress*. Washington, DC: U.S. Dept. of Justice, Office of Justice Programs.

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (2006). *Evidence-based crime prevention*. (Rev. ed. ed.) London: Routledge.

Soothill, K., Christoffersen, M. N., Hussain, M. A., & Francis, B. (2010). Exploring Paradigms of Crime Reduction: An Empirical Longitudinal Study. *British Journal of Criminology, 50,* 222-238. http://dx.doi.org/10.1093/bjc/azp076

Welsh, B. C. (2007). *Evidence-based crime prevention scientific basis, trends, results and implications for Canada*. Ottawa: National Crime Prevention Centre.

Welsh, B. C. & Hoshi, A. (2006). Communities and crime prevention. In Sherman, L.W. & Farrington, D. P. (Eds.), *Evidence-based crime prevention* (Rev. ed, pp. 165-197). London: Routledge. http://dx.doi.org/10.1007/1-4020-4244-2

Wikström, P-O. H. (2006). Individuals, settings, and acts of crime: Situational mechanisms and the explanation of crime. In *The explanation of crime: Context, mechanisms and development.* Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511489341.004

Wikström, P-O. H. (1998). Communities and crime. In Tonry, M. (Ed.). *The Handbook of Crime and Punishment*. New York: OUP.

Williams, J. & Sickles, R. C. (2002). An analysis of the crime as work model: Evidence from the 1958 Philadelphia birth cohort study. *Journal of Human Resources, XXXXVII,* 479-509. http://dx.doi.org/10.2307/3069679

Woodward, M. (1999). *Epidemiology: Study design and data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Appendix A: The outcome, risk factors and their definitions

Outcome factors		***Definition***
First contact with the police		The Police Archives include persons who have been confined or charged with crimes under the Danish Penalty Code. Confinement includes arrest, pre-trial detention, incarceration and imprisonment. For the period under review this applied to persons over the age of 15. The Road Traffic Act, the Euphoriants Acts (drug abuse), and the (rarely violated) Weapons Act are not recorded in the Penalty Code.
Risk factors **Social background**		
Parental substance abuse	(Type III)	Alcohol abuse or drug abuse (see below)
Parental inpatient mental illness	(Type II).	One or both parents admitted to a psychiatric ward according to the Danish Psychiatric Nationwide Case Register
Parental violence	(Type III)	Battered adults according to hospitals admissions. Parent exposed to assault or injuries of undetermined intent. Victims of violence which led to hospitalisation and professional assessment that the injury was willfully inflicted by other persons. *Parent convicted for violence:* The Criminal Statistic Register records persons convicted for violence. This category comprises a wide range of criminal behaviour of various degrees of seriousness: manslaughter, grievous bodily harm, violence, coercion and threats. This category does not include accidental manslaughter in combination with traffic accidents, or rape, which belongs to the category of sexual offences.
Parental suicidal behaviour	(Type III)	Parents' suicide attempts according to the National Patient Register and the Danish Psychiatric Nationwide Case Register, or suicide according to the Causes of Death Register. Intentional self-harm according to hospitals admissions is also included.
Child abuse or neglect	(Type II)	The young person having ever been a victims of violence, abuse or neglect which led to hospitalisation and professional assessment of the injury being willfully inflicted by other persons
Family background		
Child ever in care	Type II)	The child is living with the parents under caseworker supervision according to the children's acts section, or the child is placed outside the home living in an institution or in a foster home. Information from the population based register of social assistance to children in care
Family separation	(Type II)	Information on all children who had experienced divorce, separation and or the death of a parent before they were 18 years old, taken from the Danish Central Population Register (CPR) that connects children to their parents whether they are married or not.
Mother teenager	(Type II)	The mother had been a teenager herself when she gave birth to the child in focus.
Parent convicted (mother/father)	(Type I)	Convicted violations of The Danish Criminal Code in the previous year.
Vocational qualification (mother/ father))	(Type I)	Whether each parent has a vocational or professional training (e.g. bricklayer, carpenter, dentist, lawyer, or teacher in a kinder garden). This does not include semi-skilled workers. Information is based on Education Statistics or the educational classification database which is population-based, including schooling and educational training for the highest education achieved by the person each parent in focus.
Parental employment and poverty		
Parental unemployment >21 weeks	(Type II)	Unemployment for at least one parent: The number of days unemployed (more than 21 weeks) during a calendar year. From registers of Income Compensation Benefits, Labour Market Research, and Unemployment Statistics. Parental unemployment for one or both parents.
Poverty (<40% of median income)	(Type III)	Family income was less than 40% of median income in at least one of the years since the child's birth. In this study the income concept is equivalent annual household income after transfers and taxes. An individual's poverty status is decided by the level of consumption possibilities - approximated by equivalent disposable income i.e. adjusted for household composition and size. Gross income is the sum of labour earnings, asset flows, imputed value of owner occupied housing, private transfers and public transfers such as sickness benefits, unemployment insurance benefits, pensions and social assistance. Asset flows include income from rent, dividends and value of house ownership.

Individual resources		
Disadvantaged Area	(Type 1)	A governmental board has pointed identified the most disadvantaged housing areas. These are a part of the subsidized housing sector, consisting of 135 areas. About 4% of the population (200,000 persons) live in these areas. Each area has 1,500 inhabitants, on average, ranging from. 30 to 14,000 persons (Hummelgaard, Graversen, Lemmich, & Nielsen, 1997; Boligministeriet, 1993; Graversen, Hummelgaard, Lemmich, & Nielsen, 1997). These disadvantaged housing areas were divided into quintiles and the two most disadvantaged quintiles were identified as disadvantaged areas in the present by this dichotomized variable. These most disadvantaged areas would thus cover about 80,000 inhabitants or 1.6% of the total population.
Rented housing	(Type I)	The house or flat is rented, not owned by family
Danish/non-Danish Citizenship	(Type I)	. The definition is based on fulfilling one of the following conditions: • If at least one of the parents have Danish citizenship and is born in Denmark. • If there is no information in the registers about any of the parents and the child himself/herself has Danish citizenship and is born in Denmark. All others are defined as non-Danish.
Own Unemployment	(Type I)	The number of days unemployed (more than 21 weeks) during a calendar year according to registers of Income Compensation Benefits, Labour Market Research, and Unemployment Statistics.
Basic secondary schooling only	(Type III)	This corresponds to not staying at school beyond lower secondary level, which corresponds to the 9 years of compulsory schooling.
Not in process of training or education	(Type I)	Not in school, gymnasium (high school), or other education; nor in vocational training.
Not graduated from high school	(Type III)	Passed basic, but had not gone on from school to university, not at least graduated, or ever been in high school (gymnasium)
Current poverty (<50% of median level)	(Type I)	Current income of family or household less than 50% of median income the previous year.
Own inpatient mental illness	(Type II)	Admitted to a psychiatric ward according to the Danish Psychiatric Nationwide Case Register.
Own attempted suicide	(Type II)	Self-inflicted harm according to hospitals admissions. The definition of suicide attempts also included behaviour that conformed to the following conditions: (i) Suicide attempts that had led to hospitalisation, (ii) assessment of the trauma being an act of self-mutilation according to the international statistical classification of injuries when discharged from hospital, (iii) the trauma had to be included in a specified list of traumas traditionally connected with suicide attempts: cutting in wrist (carpus), firearm wounds, hanging, self-poisoning with drugs, pesticide, cleaning fluids, alcohol or carbon monoxide. This does not include non-suicidal self-harm
Drug abuse	(Type II)	Addiction or poisoning by drugs according to hospitals admissions. Mental and behavioural disorder due to use of drugs (e.g. opioids, cannabinoids, cocaine). Dependence on morphine was not included if associated with diseases of chronic pain
Alcohol abuse	(Type II)	According to hospital admissions the following diagnoses were expected to be associated with long-term alcohol abuse: Alcoholic psychosis, alcoholism, oesophageal varices, cirrhosis of liver (alcoholic), chronic pancreatitis (alcoholic), delirium, accidental poisoning by alcohol. Mental and behaviour disorder due to use of alcohol also included

Appendix B: Statistical model

The data have been analysed by the discrete time-Cox-model (Allison, 1982). A procedure was carried out to select significant risk factors to give the best possible prediction. Only first contacts with the police are analysed in the Cox-model. The available event history data contains information on events that fell within each calendar year from 1984 to 2006. Individuals' event history is broken up into a set of discrete time units (a calendar year) in which an event either did or did not occur. An event is first contact with the police (arrest, pre-trial detention or charges of crimes).

When the discrete time unit is a calendar year, it is difficult to use continuous-time methods. Problems arise when the time intervals are large enough that more than one individual experiences an event in the same time interval (Allison, 1982). A discrete-time model is more appropriate for the estimation of parameters as it treats each individual history as a set of independent observations. It has been shown that the maximum likelihood estimator can be obtained by treating all the time units for all individuals as though they were independent, when studying first-time events (Allison, 1982).

Each individual is observed until time *t*, at which point either an event occurs or the observation is censored, by reaching the age limit, because of death, or the individual is lost to observation for other reasons. Consequently, individuals were excluded from the case group and controls after the first event. Pooling the non-censored years of all individuals, the person-years, made the numbers at risk. The person-years at risk were constructed for the total birth cohort of 27,840 men and 26,618 women who were living in Denmark in 1998 when they were about 14 years old.

The estimated hazards of first time criminality within the following year are estimated by following equations:

$$(1) \qquad \pi(Y=1) = \frac{e^{\beta_0+\beta_1+\beta_2+\dots}}{1+e^{\beta_0+\beta_1+\beta_2+\dots}}$$

Weights are estimated by the Greek letter β, and *e* is a constant, which also is the base of the natural logarithm approximately: 2.71828. These estimations are done using the whole database and compared to actual observed criminality based on Police archives. The beta-coefficients are assumed to be constant within the relatively short time-span (1999-2006).

The weights (or parameters) are estimated according to the 'maximum-likelihood' method which gives the best possible prediction based on the most informative risk factors among the available significant factors.

And for each person and each calendar year e.g. 2000, the first time hazards are calculated and named:

$$\pi_{2000}.$$

The hazards for not being 'caught' in year 2000, given that the person had not been 'caught' before, are therefore:

$$(1-\pi_{2000}).$$

The hazard for being 'caught' at least once over the years 1999 to 2006 is one minus the hazards of not being 'caught' any of the years:

$$(2) \qquad 1-(1-\pi_{1999})(1-\pi_{2000})(1-\pi_{2001})\dots(1-\pi_{2005}).$$

This will be an estimate of the hazard of reaching the 22nd birthday having been at least once caught by the police.

Appendix C: Estimates of Discrete Cox Model 2 (table 2) of first-time crime events when interactions with gender included

	Odds ratio	95% Wald Confidence Limits	
20 year	0.41	0.37	0.46
19 year	0.57	0.51	0.63
18 year	0.80	0.74	0.88
17 year	0.82	0.76	0.89
16 year	0.85	0.78	0.91
Parental background:			
Parental inpatient mental illness	1.27	1.11	1.46
Parental violence	1.32	1.14	1.53
Non-Danish	1.59	1.35	1.88
Mother has no vocational qualification	1.15	0.99	1.34
Parental unemployment >21 weeks	1.29	1.11	1.49
Child abuse and neglect	1.79	1.47	2.19
Family separation	1.72	1.52	1.96
Mother teenager	1.31	1.08	1.58
Mother convicted	2.56	1.72	3.80
Father convicted	1.59	1.14	2.22
Location:			
Rented housing (not self-owner)	1.26	1.11	1.43
Disadvantaged area	1.25	0.98	1.60
*Individual resource*s			
Not in process of training or education	1.41	1.22	1.62
Not graduated high school	1.77	1.51	2.07
Current Poverty (< 50 % of median)	1.23	1.05	1.45
Own unemployment >21 weeks	1.49	1.09	2.03
Focus child in care	1.49	1.28	1.74
Substance abuse (alcohol. drugs)	2.24	1.75	2.88
Own attempted suicide	1.93	1.45	2.55
Gender (1=boy; 0=girl)	4.01	3.15	5.10
Interaction term:			
Male*(Parental mental illness)	0.83	0.71	0.97
Male*(did not pass basic secondary level)	1.66	1.23	2.25
Male*(Mother convicted)	0.56	0.33	0.93
Male*(Attempted suicide)	0.67	0.45	1.00

Note: All interaction terms were included in the model, but only significant interaction terms shown in the table.

Appendix D: Estimated odds ratios: first-, second- to fifth-time events (arrest, confinement or charged according to penalty code)

	Odds ratio: First event	Odds ratio: second event	Odds ratio: third event	Odds ratio: fourth event	Odds ratio: Fifth event
Parental background:					
Parental substance abuse	ns	ns	Ns	ns	ns
Parental in-patient mental illness	1.1	ns	Ns	ns	ns
Parental violence	1.5	1.6	1.7	1.7	1.8
Non-Danish	1.6	1.7	1.9	2.1	2.3
Mother has no vocational qualification	1.2	1.2	1.2	1.2	1.3
Father has no vocational qualification	1.3	1.2	1.3	1.2	1.2
Parental suicidal behaviour	Ns	ns	Ns	ns	ns
Poverty (<40% of median income)	Ns	1.1	Ns	1.1	ns
Parental unemployment >21 weeks	1.4	1.5	1.6	1.7	1.7
Child abuse and neglect	1.9	2.0	2.1	2.1	2.2
Family separation	1.6	1.6	1.6	1.6	1.6
Mother teenager	1.3	1.3	1.3	1.3	1.2
Mother convicted	1.8	1.5	ns	Ns	ns
Father convicted	1.3	1.4	ns	Ns	ns
Location					
Rented housing (not self-owner)	1.2	1.3	1.3	1.3	1.3
Disadvantaged area	1.3	1.4	1.3	1.3	1.4
Individual resources					
Didn't pass basic secondary education	1.5	1.8	2.0	2.0	2.1
Not in process of training or education	1.1	1.2	1.2	1.3	1.4
Not graduated high school	1.8	2.0	2.4	2.5	2.6
Current poverty (< 50 % of median)	1.2	1.2	1.2	1.2	1.3
Unemployment >21 weeks	1.6	1.8	1.7	1.7	1.8
Ever in care	1.4	1.5	1.8	1.9	2.1
Own attempted suicide	1.6	1.5	1.7	1.7	1.7
Substance abuse (alcohol. drugs)	2.0	1.9	1.9	2.1	2.2
Own inpatient mental illness	ns	ns	ns	Ns	ns
Gender (1=boy; 0=girl)	3.8	4.8	6.3	7.9	8.6
Number persons with police-contact	6,075	4,189	2,867	2,228	1,814
Number person-years	300,591	310,350	316,330	318,875	320,432

Note: Number of first-time events n=6,075. Total number of individuals in the study 54,458, while the total number of person-years is 300,591. 'ns' stands for Non-Significant. Age terms not shown

A comparison of approaches for assessing covariate effects in latent class analysis

Jon Heron School of Social and Community Medicine, University of Bristol, UK
jon.heron@bristol.ac.uk
Tim J. Croudace School of Nursing and Midwifery, University of Dundee, UK
Edward D. Barker Institute of Psychiatry, King's College London, London, UK
Kate Tilling School of Social and Community Medicine, University of Bristol, UK

(Received October 2014 Revised March 2015) http://dx.doi.org/10.14301/llcs.v6i4.322

Abstract

Mixture modelling is a commonly used technique for describing longitudinal patterns of change, often with the aim of relating the resulting trajectory membership to a set of earlier risk factors. When determining these covariate effects, a three-step approach is often preferred as it is less computationally intensive and also avoids the situation where each new covariate can influence the measurement model, thus subtly changing the outcome under study. Recent simulation work has demonstrated that estimates obtained using three-step models are likely to be biased, particular when classification quality (entropy) is poor. Using both simulated data and empirical data from a large United Kingdom(UK)-based cohort study we contrast the performance of a range of commonly used three-step techniques. Bias in parameter estimates and their precision were determined and compared to new bias-adjusted three-step methods that have recently become available. The bias-adjusted three-step procedures were markedly less biased than the simpler three-step methods. Proportional Maximum Likelihood (ML), with its complex-sampling robust estimation, suffered from negligible bias across a range of values of entropy. Whilst entropy was related to bias for all methods considered, there was evidence that class-separation for each pairwise comparison may also play an important role. Under some circumstances a standard three-step method may provide unbiased covariate effects, however on the basis of these results we would recommend the use of bias-adjusted three-step estimation over these standard methods.

Keywords

ALSPAC, latent class analysis, trajectories, bias, three-step

Introduction

The use of mixture models in epidemiological research has increased markedly in recent years, partly due to developments in statistical software packages such as Mplus (Muthén & Muthén, 2012) and Latent Gold (Vermunt & Magidson, 2013) that have brought these complex, computationally intensive techniques within the grasp of the average applied researcher. Mixture models come in various forms; some designed specifically for longitudinal data e.g. Latent Class Growth Analysis or Growth Mixture Models (Muthén & Muthén, 2000) and others such as standard Latent Class Analysis appropriate in either a longitudinal or cross-sectional setting. All models share one feature, the estimation of an underlying categorical latent variable (hereafter referred to as *X*) which is theorized to be the reason for some or all of the patterns of association observed within the dataset. The procedure will estimate the likely distribution of *X*, namely the number of classes and their prevalence, as well as individual probabilities of class membership, which describe the allocation of each participant/observation to each latent class under the estimated model. Many stopping rules, e.g. entropy (Ramaswamy, DeSabro, & Robinson, 1993), Bayesian Information Criterion (BIC) (Schwarz, 1978), Bootstrap Likelihood Ratio Test (BLRT) (Nylund, Asparouhov, & Muthén, 2007) have been utilized with the goal of determining an adequate number of classes.

In some cases *X* itself is of little interest, for instance its inclusion may be purely to help with some deviation from normality within the data. However, more often estimating *X* is a key focus as it may represent underlying subpopulations who have different characteristics or who may respond differently to some intervention. The analyst will typically estimate *X* on the basis of a few 'class-indicators', such as repeated measures of enuresis (Croudace, Jarvelin, Wadsworth, & Jones, 2003) or cross-sectional symptoms of psychosis (Shevlin, Murphy, Dorahy, & Adamson, 2007) before offering up *X* for further investigation e.g. to understand which early-life factors distinguish between the classes or what is the long-term prognosis of members of each group. It is during this secondary stage where no firm rules have been established with regard to best practice and a number of analytical approaches have been adopted across the applied literature. Despite the relative ease with which one may determine covariate effects within a "one-step" model where the measurement model for *X* is estimated at the same time as the covariate odds-ratios for class-membership, a number of "three-step" procedures are commonly used.

The term "three-step" (Vermunt, 2010) refers to the sequential stages of firstly estimating the mixture model, secondly exporting the salient features of the model to a different statistical package, before finally analysing some derived indicator of class membership in further analysis, e.g. as the outcome in a multinomial logistic regression model. Popular second-step procedures include assigning each participant to their most likely class (Modal Assignment) or incorporating class-assignment uncertainty either by making multiple draws from each participant assignment probabilities (Pseudo-Class Draws, PCD) or using the probabilities themselves as regression weights (Proportional Assignment). All methods aside from the one-step fall under the banner of three-step methods, even if the second step merely involves exporting the data from step one.

Recent simulation work (Clarke & Muthén, 2009) has demonstrated a number of shortcomings of these three-step methods, including substantial parameter bias and over- precise estimates. However, as described by Clarke & Muthén and also Vermunt, the three-step strategy brings a number of advantages including reduced model complexity as well as avoiding the situation where the form (and potentially interpretation) of *X* may alter depending on the covariates/outcomes included in the model. As is often the case, a single mixture model which defines a sub-division of the study population may give rise to a series of related papers so there is clear benefit to having a consistent, unchanging assignment of the study participants.

In a recent paper, Vermunt (Vermunt, 2010) has brought applied analysts a new alternative by devising a pair of refined three-step procedures. Using standard mixture-modelling output which describes the agreement between the estimated and underlying latent measure, the third step of a three-step procedure can be adjusted to remove the measurement error induced through estimation of the latent measure in step two. Bias and precision are seen to be improved, but crucially the latent class assignment is unchanged, thus a succession of different models can be examined without impacting on the formulation of *X*.

The aim of the current paper is to investigate how these estimation approaches perform in practice, when applied to the analysis of trajectories of conduct problems in childhood (Barker & Maughan, 2009) derived using data from the Avon Longitudinal Study of Parents and Children (ALSPAC), a UK-based birth-cohort. The latent grouping produced in the original manuscript has since been utilized in a number of follow-up publications (Barker, Oliver, & Maughan, 2010; Heron et al., 2013a; Heron et al., 2013b; Kretschmer et al., 2014; Oliver, Barker, Mandy, Skuse, & Maughan, 2011; Stringaris, Lewis, & Maughan, 2014) in which a range of one- and three-step procedures have been employed in order to examine further risk factors for non-normative development or to study late problematic outcomes in those exhibiting different patterns of conduct problem behaviour. In the current manuscript we select a single covariate (gender) in order to compare results obtained using the range of methods now available. Observations are subsequently verified through simulation.

Methods

Participants

The sample comprised participants from the Avon Longitudinal Study of Parents and Children (ALSPAC) (Boyd et al., 2013; Fraser et al., 2013; Golding, Pembrey, & Jones, 2001). ALSPAC is an ongoing population-based cohort study in the South-West of England. Pregnant women resident in the former Avon Health Authority (which included the city of Bristol), who had an estimated date of delivery between 1 April 1991 and 31 December 1992, were invited to take part, resulting in a cohort of 14,541 pregnancies which resulted in 13,796 singletons and first-born twins who were alive at one year of age. Detailed information about ALSPAC is available online (http://www.bris.ac.uk/alspac) and the study website also contains details of all the data that is available through a fully searchable data dictionary (http://www.bristol.ac.uk/alspac/researchers/data-access/data-dictionary/). Ethical approval for the study was obtained from the ALSPAC Law and Ethics Committee and local Research Ethics Committees.

Outcome - Conduct Problem (CP) trajectories during childhood

The derivation of CP trajectories has been reported previously (Barker & Maughan, 2009). Briefly, Latent Class Growth Analysis was applied to six assessments of mother-reported CP, spanning the age period from four to 13 years, using the 'Conduct Problem' subscale of the Strengths and Difficulties Questionnaire (Goodman, 2001; Goodman & Scott, 1999) The sum-score at each wave was dichotomized at the standard threshold of four or more (Goodman, 2001), yielding six binary indicators. The four resulting trajectories were described as "Low" (72.4%), "Childhood Limited" (CL, 11.8%), "Adolescent Onset" (AO, 7.8%) and "Early-Onset Persistent" (EOP, 8.0%). Proportions quoted are for the complete-case sample (n = 4,659) following modal assignment. Entropy for this model was 0.730.

Exposure

For these models we will focus on offspring sex, which is coded 0 'female', 1 'male' so that parameter estimates indicate the extent to which boys have greater log-odds compared with girls of being in the comparison class.

Statistical methods

Whilst "C" is often used when referring to the latent variable within a latent class model, here we adopt the notation used in Vermunt (2010). We use X to denote the underlying latent variable and W for any predicted classification obtained during the second step of a three-step estimation method. Latent class indicators for subject i are denoted by $\boldsymbol{Y}_i$ and a covariate (predictor of class-membership) by Z_i (i.e. sex in the empirical example).

Empirical models

The effect of sex on latent class variable X (conduct trajectory class) was assessed using a range of one- and three-step methods, each time treating X as a four-category multinomial outcome. Of interest was both the magnitude of the main effects of sex, given by log-odds ratios, and their standard errors. As it is customary to approach these models with the mind-set that these classes are all inherently different in some way, we chose to make comparisons between all classes rather than just deriving parameter estimates with reference to the normative (Low) group. For each comparison we examine percentage deviation from the one-step results, defined to be the difference between each three-step result and those derived from the one-step method, expressed as a percentage of the one-step estimates. We note here that we are making the assumption that the one-step results are correct and for our empirical models we do not know this to be the case.

The following methods were compared:

One-step estimation - The direct effect of sex on X was estimated by incorporating this independent variable into the original mixture model. Estimation was carried out using Mplus version 7.1 (Muthén & Muthén, 2012).

Three-step methods - With all three-step methods the first step entails the estimation of an unconditional mixture model, i.e. a measurement model for latent class X in the absence of any potential covariates. The output from this first step consists of a set of class-assignment probabilities – denoted $P(X = t \mid \boldsymbol{Y}_i)$ – for each respondent. Respondents with the same set of responses for class indicators $\boldsymbol{Y}_i$ are given an identical set of class-assignment probabilities, however depending on the three-step method chosen, such respondents may not all be assigned to the same class. During step-two these data are used to derive the nominal variable *W,* which is then used as the dependent variable in the final step. Here the methods chosen adopt one of two alternative step-two procedures – Modal Assignment and Proportional Assignment. We first discuss their standard use before describing the bias-adjusted approaches.

Modal Standard - Perhaps the most commonly-used three-step method, the second step entails assigning each respondent to their most likely class (the class for which $P(X = t \mid \mathbf{Y}_i)$ is greatest). In step three this classification *W* becomes the nominal dependent variable in a multinomial logistic regression analysis. Whilst we use Latent Gold for all three-step models described, this model can be estimated in mainstream statistical software such as Stata and SPSS.

Proportional Standard - In contrast to modal assignment, three-step methods based on proportional assignment incorporate the class-assignment probabilities. Proportional Assignment involves stacking ones' class-assignment probabilities so that each respondent has multiple rows of data (one row per class). An additional column is created which indexes these classes. For step-three a multinomial logistic regression model is estimated with this class-index as the dependent variable and the column of assignment probabilities used as regression weights (this method is also known as "Probability Weighting"). This model is also estimable in Stata with the assignment probabilities defined to be "importance weights" and in SPSS through the use of frequency weights.

Modal ML and Proportional ML - The three-step methods Modal Standard and Proportional Standard suffer from two limitations. Firstly they assume a perfect relationship between the classification *W* derived in step two and the unmeasured latent variable *X*, and secondly they fail to account for the fact that *X* is latent so its true values are unknown. Vermunt (2010) devised a pair of bias-adjusted estimation methods, referring to these as "Modal ML" and "Proportional ML". The estimation of these methods requires the appropriate "D-matrix" containing classification probabilities that describe the relationship between *W* and *X*, or put another way, they quantify the measurement error in *W*. Through the use of this classification matrix, a subsequent latent class estimation - well established as a method for dealing with measurement error in categorical variables - is able to reproduce the quantity of interest, namely the effect of covariate Z_i on *X*. As a consequence of the need for a second latent-class analysis, software options for estimating step three are more limited.

Through simulation work, Proportional ML was observed to produce parameter estimates closer to the one-step (true) results, whilst Modal ML gave more accurate standard errors (SE) - SE's for Proportional ML were slightly too large. Vermunt demonstrated how one might estimate these models in Latent Gold, however Modal ML is also estimable in Mplus, and, since version 7.1, has been simplified through use of the "auxiliary" command. See the supplementary material for further details on the derivation of the D-matrix and the estimation of these models in Latent Gold and Mplus. Finally we note that when the D-matrix for either Modal or Proportional Assignment is equal to the identity matrix the Modal Standard or Proportional Standard estimates are reproduced. In other words, as stated above, standard methods make the assumption that there is no measurement error in W.

Modal ML (robust) and Proportional ML (robust) - In a follow-up publication to Vermunt (2010), Bakk and colleagues (Bakk, Oberski, & Vermunt, 2014) revised the estimation methods for both Modal and Proportional ML. By using a complex-sampling robust estimator to allow for within person clustering (in our empirical example the stacked dataset has four rows per respondent) and a Taylor expansion to better allow for the classification-error uncertainty inherent in the third step estimation, improvements on the original bias-adjusted estimates have been demonstrated, particularly for Proportional ML. Modal ML (robust) and Proportional ML (robust) are both available in Latent Gold version 5.0 however neither can be estimated currently in Mplus (version 7.3).

Simulation models

We sought to replicate the findings from the empirical analysis using a simple simulation study. This enabled us to take control aspects of the model such as entropy and class separation, and furthermore ensure that our chosen one-step model was the appropriate one for the data.

Simulation #1: Relationship between bias and entropy

Had we simulated from a model containing a mixture derived from repeated binary indicator variables it would have been difficult to vary entropy/class-separation in a controlled manner. Consequently, the class indicator used here was a single multimodal continuous variate *Y*. Latent class *X* was then to be regressed on a single binary covariate Z_i giving rise to a pair of log-odds ratios describing the Z_i-by-*X* relationship. The Monte Carlo routine in Mplus was used to simulate the necessary data with further details given below.

Defining the relationship between observed class indicator Y and latent class X

Continuous variate *Y* was simulated to be a mixture of three normal distributions of equal size, located at values -1 (class 1), 0 (Class 2) and 2 (Class 3) as illustrated in Supplementary Figure 1. Variances were constrained equal for all three distributions and were increased incrementally from 0.05 to 0.5 in steps of 0.05 yielding ten different simulation models. A (within-class) variance of 0.05 produces a near-perfect value of entropy (~1.0) and very good class separation. As variance is increased, class-separation is reduced initially for the two closer classes (classes 1 and 2) and ultimately all three classes will be poorly separated. Within-class variance was the only aspect of the model to be varied between simulations. 500 replications were produced for each of the ten models with a constant sample size of 5,001. Preliminary work indicated acceptable coverage and bias for the one-step model when using this number of replications.

*Defining the relationship between Covariate Z_i and latent class **X***

The association between binary covariate Z_i and three category nominal outcome *X* can be described as a six-cell contingency table. Consequently, five quantities (in addition to the sample size) are required to fully describe these data. For the set-up used in Mplus, the following details were needed: the proportion of people in the Z_i = 0 group; two log-odds ratios defining the relationship between Z_i and *X*; and two logits to define the class distribution *X* in the unexposed group (Z_i=0). Here we opted for three classes of equal size (n = 1,667). The proportions exposed to Z_i within each class were as follows: class 1 (517/1,667 = 31.0%), class 2 (417/1,667 = 25%), class 3 (317/1,667 = 19%). This results in a covariate Z_i with 25.01% prevalence and log-odds ratios of 0.649 for class 1 and 0.351 for class 2 (with reference to class 3), giving a log-odds ratio of 0.298 for class 1 with reference to class 2. In other words, relative to class 3, exposure to covariate Z_i would convey moderately increased log-odds of being in class 2, and a greatly increased log-odds of being in class 1. Finally, the chosen cell counts imply a class-distribution of *X* of 30.67%/33.33%/36.0% among those unexposed to $Z_{i,}$ which can be described as two logits: -0.160 and -0.077.

Analysis of simulated data

Each of the one-step and three-step methods were used to estimate the effect of Z_i on *X* for each simulated dataset. This was facilitated through use of the brew package (Horner, 2011) in R (R Core Team, 2014). All parameter estimates were imported into Stata version 13.1 (StataCorp., 2013) where the –simsum– routine (White, 2010) was employed to derive the measure of bias relative to the true regression parameters (0.649, 0.351 and 0.298). We also compared estimate precision by calculating the SD in each parameter estimate across the 500 simulated datasets.

Simulation #2: Relationship between bias and pairwise class separation

Analysts tend to focus on entropy as a single summary measure of class assignment uncertainty for the whole model, however it is often the case that some large classes are well defined with other smaller classes being less so. In this case, it will be the large classes driving entropy, and not all class-comparisons will have the same degree of accuracy. Maitra and Melnykov provide equations (equation 2.1 in Maitra & Melnykov, 2010) for deriving what they refer to as cluster-overlap when estimating a Gaussian mixture model. For each pair of classes, the cluster-overlap is defined as the sum of two misclassification probabilities for the overlap with class *i* when considering class *j*, and vice versa. Hence a pairwise measure of cluster-overlap is readily available and is given by the sum of the [i,j] and [j,i] elements of the “D-matrix”. This formally defined measure of cluster-overlap is essentially the opposite of what we have been referring to more loosely as class-separation. For a pair of classes with good separation, overlap will be close to zero. In contrast, independence between *X* and *W* would yield overlap of 2/(# classes), with a more complex *X-W* relationship producing potentially greater values, though ultimately bounded by 2.

We sought to investigate the role that cluster-overlap has on the bias of our estimates. Here, we focus on the first comparison (class 1 versus class 3) for which the covariate had the largest effect in the original simulation (log odds = 0.649). For a given value of entropy, the association between parameter bias and pairwise class-overlap is confounded by the magnitude of the covariate effects. Consequently we re-simulated the data after permuting the ordering of the classes. This was done keeping both entropy AND the covariate-effects constant and only works because our three classes were simulated to be of equal size (otherwise the permutation would alter entropy). If we label the original simulation model as “123” reflecting the ordering of the classes at locations -1, 0 and 2, then permuting the classes to orders “312” and subsequently “231” enables us to vary class-separation as shown in figure 3. Note that there are three other possible class orderings, “132”, “213” and “321”, which produce the same three measures of cluster-overlap and the same levels for bias (“123” is equivalent to “321” etc.). Following the simulation of these new data, the same analytical steps were performed as for Simulation #1. Parameter estimate bias was calculated and its relationship with cluster-overlap was examined.

Results

Empirical example

Estimated sex effects for each pair of latent classes are shown in table 1. Figures in parentheses show percentage deviation from the one-step results. As the entropy for the original mixture model was not particularly high (0.730), previous simulation work would predict that standard three-step methods would be inaccurate, typically under-estimating the effects of sex and also being overly-precise since these methods do not capture the uncertainly in estimated class assignment.

Parameter estimates

For all class comparisons, the standard three-step methods produce estimates closer to the null than the one-step results. Estimates obtained using Proportional ML are consistently within 1 or 2% of the one-step results. Modal ML estimates are more variable, and are substantially higher than the one-step for the comparison of classes Childhood Limited and Early Onset Persistent. Unsurprisingly, the use of robust SE’s has no effect here.

Standard errors

Again, as expected, the standard three-step methods are overly precise with SE’s up to 32% and 58% lower that the one-step for Modal and Proportional Standard respectively. Proportional ML severely over-estimates SE, however the new complex-sampling robust variance estimator demonstrates a marked improvement here. The robust estimator has little effect on Modal ML, with all SE’s being moderately raised compared to one-step and Proportional ML (robust).

Summary of empirical findings

The three-step methods chosen produced a wide range of estimates for the parameters and their standard errors. What is apparent is that deviations relative to the one-step values are typically lower, particularly for the standard errors, when comparing pairs of classes which have better separation. Like many longitudinal mixture models, the analysis of conduct problems produced patterns of trajectories which have been described previously as a soldier’s bed or cat’s cradle (Sher, Jackson, & Steinley, 2011) in other words high and low relatively flat trajectories and a pair of trajectories which cross midway through the time period. Here the classes which cross (AO and CL) are less well separated, whilst the two persistent classes (Low and EOP) have little overlap. This appears to be reflected in the consistency of their estimates across the methods.

Simulation #1: Relationship between bias and entropy

Unconditional three-class mixture models estimated on each simulated dataset reported the following entropy values (averaged across 500 datasets): 0.98, 0.91, 0.85, 0.79, 0.75, 0.70, 0.67, 0.63, 0.61 and 0.58. Figure 1 shows the relationship between entropy and the percentage bias obtained in the parameter estimates and figure 2 shows estimated precision (SD of estimate across datasets) for each method.

When comparing results from bias-adjusted methods our findings were consistent with recent simulation work (Bakk et al., 2014). Modal ML and Modal ML (robust) results were almost identical in both bias and precision, likely due to the large sample size in our examples. In contrast (as expected), there was a marked increase in precision with Proportional ML (robust). Standard errors for Proportional ML (robust) were within 3% of the one-step values for all values of entropy whereas for non-robust Proportional ML the standard errors were in one instance 86% higher than those obtained using a one-step approach. On the basis of these results we would caution against the use of Proportional ML without robust standard errors. Here we report results only for the two more recent methods – Modal ML (robust) and Proportional ML (robust) – however a full set of results are available on request. To facilitate a clearer comparison of these two methods, we have reproduced the figures after removing the standard methods to enable the y-axis to be restricted (see supplementary material).

Parameter estimate bias

Due to the location of the three classes, reduction in entropy initially impacts on the comparison of class 1 versus class 2 (third comparison) followed by the other two comparisons. We observe both positive and negative bias in this example, however we note that estimates affected by positive bias will be bounded by the maximum value of the true log-odds ratios – in this case 0.649 (Bolck, Croon, & Hagenaars, 2004). The standard three-step methods are badly affected by the reducing entropy, with Modal Standard fairing slightly better but still producing unacceptable levels of bias unless entropy is very high. Both bias-adjusted three-step methods produce estimates with a low level of bias for all three class comparisons and across the wide range of entropy values considered.

We see that for the second comparison the bias for standard three-step methods appears to decrease for lower values of entropy. This phenomenon is merely an artefact of our chosen simulation. As entropy reduces, the distinction between classes 1 and 2 is the first to become affected such that class 1 becomes more similar to class 2 and vice versa. Since class 1 is more strongly associated with the covariate, our second comparison (class 2 versus class 3) is boosted, partially offsetting the negative-bias present in both standard methods.

Standard Errors

Decreasing entropy should increase uncertainty and accordingly we observe a reduction in precision for the (correct) one-step model. Standard errors for Proportional ML (robust) closely match the one-step values with Modal ML (robust) giving slightly higher values. What is most apparent from these figures is that the standard three-step approaches are failing to capture the increasing uncertainty, in fact in this example Proportional Standard becomes more precise as the level of assignment uncertainty increases.

Simulation #2: Relationship between bias and pairwise class separation

Table 2 shows the resulting biases for this second set of simulations. Output is restricted here to the five highest values of entropy – typically the range in which an analyst might be considering the use of a standard three-step method. These results are split into two since methods using Modal and Proportional assignment will have a different D-matrix and hence a different value for class-separation for the same dataset. We see that for very high levels of entropy (>> 0.9) there is little detriment to using any modelling approach. However unacceptable (>10%) levels of bias in the parameter estimate is present when entropy is still extremely high (0.91) if the class overlap is moderate, and in contrast, *less* bias for *lower* entropy (0.75 – 0.80) when a particular pair of classes has a good degree of separation. Whilst these results are limited in scope, they suggest that a decision based solely on entropy may be unwise.

Discussion

Using an empirical example from a large UK birth cohort and a limited set of simulations we have compared the estimate effect of a single covariate on latent class membership using various three-step approaches commonly used in applied papers from the fields of psychology, epidemiology and medicine. Our findings are consistent with previous simulations showing that standard three-step methods can produce results which are both biased and overly precise, particularly when entropy is poor. What this study adds is the suggestion that entropy, a single-summary measure of classification quality, is only part of the story and we would advise caution regarding a modelling strategy based solely on its value, for instance whether it exceeds an arbitrary threshold such as 0.8 or 0.9.

We have demonstrated that for extremely high values of entropy it remains possible for individual class comparisons to be biased if the separation between those classes is poor. In contrast, when entropy is low, some class comparisons may be unbiased if their separation is good relative to the rest of the model. When faced with the worst-case scenario of a combination of low entropy and poorly separated classes, only proportional ML (robust), of the three-step methods, appears to fare well, however previous simulations suggest that for extremely low entropy all three-step methods may be flawed (Bakk, Tekle, & Vermunt, 2013; Vermunt, 2010) leaving the one-step method as the only option for obtaining unbiased estimates. Our simulation focussed on what would be regarded as a large sample size for this type of analysis and this is likely to be an explanation for the strong performance of proportional ML (robust) across the whole range of entropy considered.

It is clear from our results that pairwise class-separation may play an important role in determining the level of bias in the standard three-step methods, although we are unable to make recommendations with regard to acceptable thresholds. There is a strong link between separation and entropy, and separation will be also affected by the number of classes present and their relative positioning. Thus, derivation of thresholds for class-separation will be challenging. In our view further efforts would be better directed at facilitating the use of bias-adjusted three-step methods within mainstream statistical software.

In our empirical example we focussed on the respondents with a full set of class indicators. Whilst we observed good agreement between the one-step and the robust ML three-step methods our sample used for analysis consists of merely one third of ALSPAC hence our estimates may not generalise to the broader sample of those who enrolled. Here we make a number of observations in relation to this since the topic of missing data in the context of three-step estimation is currently unexplored.

Firstly, Full Information Maximum Likelihood (FIML) permits the inclusion of partial respondents based on the missing-at-random (MAR) assumption. However, as entropy for such a model would be expected to be lower due to additional uncertainty surrounding these incomplete observations, there is the potential for this to offset gains made through the use of a larger, more representative sample. Alternative approaches include focussing on a sample for which a rich set of class-indicators are available and using a weighting method, e.g. Inverse Probability Weighting (IPW), to adjust for any potential selection bias. IPW has recently been shown to be a useful technique when used in combination with other missing data methods (Seaman, White, Copas, & Li, 2012). Secondly, when using likelihood-based methods to deal with missing data, one may condition on predictors of missingness to strengthen the MAR assumption. Were covariate Z_i to be an important predictor of dropout as well as being an exposure of interest, one would surmise that only the one-step method would achieve an unbiased result. Finally, FIML-based mixture modelling can only deal with missing covariate information (incomplete Z) in a rather simple setting and by making potentially undesirable distributional assumptions. A clear advantage of the treat-as-observed approach of Modal Standard is the ease with which one may then incorporate classification W into a multiple imputation model where any covariate missingness can be dealt with. Future developments could focus on a toolkit for the applied researcher that allows bias-adjusted estimation of the Z_i-by-X association with a range of currently state-of-the-art missing data treatments.

Acknowledgements and funding

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Jon Heron and Kate Tilling will serve as guarantors for the contents of this paper. This research was specifically supported by the Medical Research Council [grant number G1000726]. Kate Tilling works in a unit that receives funding from the UK Medical Research Council and the University of Bristol (MC_UU_12013/5).

References

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis, 22*, 520-540. http://dx.doi.org/10.1093/pan/mpu003

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology, 43*, 272-311. http://dx.doi.org/10.1177/0081175012470644

Barker, E. D., & Maughan, B. (2009). Differentiating early-onset persistent versus childhood-limited conduct problem youth. *American Journal of Psychiatry, 166*, 900-908. http://dx.doi.org/10.1176/appi.ajp.2009.08121770

Barker, E. D., Oliver, B. R., & Maughan, B. (2010). Co-occurring problems of early onset persistent, childhood limited, and adolescent onset conduct problem youth. *Journal of Child Psychology and Psychiatry, 51*, 1217-1226. http://dx.doi.org/10.1111/j.1469-7610.2010.02240.x

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. *Political Analysis, 12*, 3-27. http://dx.doi.org/10.1093/pan/mph001

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort Profile: The 'Children of the 90s'-the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology, 42*, 111-127. http://dx.doi.org/10.1093/ije/dys064

Clarke, S. L., & Muthén, B. (2009). Relating Latent Class Analysis Results to Variables Not Included in the Analysis. Retrieved from www.statmodel2.com/download/relatinglca.pdf

Croudace, T. J., Jarvelin, M. R., Wadsworth, M. E., & Jones, P. B. (2003). Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology, 157*, 834-842. http://dx.doi.org/10.1093/aje/kwg049

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., & Lawlor, D. A. (2013). Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology, 42*, 97-110. http://dx.doi.org/10.1093/ije/dys066

Golding, J., Pembrey, M., & Jones, R. (2001). ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatric and Perinatal Epidemiology, 15*, 74-87. http://dx.doi.org/10.1046/j.1365-3016.2001.00325.x

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*, 1337-1345. http://dx.doi.org/10.1097/00004583-200111000-00015

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: is small beautiful? *Journal of Abnormal Child Psychology, 27*, 17-24. http://dx.doi.org/10.1023/A:1022658222914

Heron, J., Barker, E. D., Joinson, C., Lewis, G., Hickman, M., Munafo, M., & Macleod, J. (2013a). Childhood conduct disorder trajectories, prior risk factors and cannabis use at age 16: birth cohort study. *Addiction, 108*, 2129-2138. http://dx.doi.org/10.1111/add.12268

Heron, J., Maughan, B., Dick, D. M., Kendler, K. S., Lewis, G., Macleod, J., Munafo, M., & Hickman, M. (2013b). Conduct problem trajectories and alcohol use and misuse in mid to late adolescence. *Drug and Alcohol Dependence, 133*, 100-107. http://dx.doi.org/10.1016/j.drugalcdep.2013.05.025

Horner, J. (2011). *Templating Framework for Report Generation. R package version 1.0-6.* . Retrieved from http://CRAN.R-project.org/package=brew

Kretschmer, T., Hickman, M., Doerner, R., Emond, A., Lewis, G., Macleod, J., Maughan, B., Munafo, M. R., & Heron, J. (2014). Outcomes of childhood conduct problem trajectories in early adulthood: findings from the ALSPAC study. *European Child & Adolescent Psychiatry, 23*, 539-549. http://dx.doi.org/10.1111/add.12268

Maitra, R., & Melnykov, V. (2010). Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. *Journal of Computational and Graphical Statistics, 19*, 354-376. http://dx.doi.org/10.1198/jcgs.2009.08054

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism, clinical and experimental research, 24*, 882-891. http://dx.doi.org/10.1111/j.1530-0277.2000.tb02070.x

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide. 7th Edition*: Los Angeles, California: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modelling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 535-569. http://dx.doi.org/10.1080/10705510701575396

Oliver, B. R., Barker, E. D., Mandy, W. P., Skuse, D. H., & Maughan, B. (2011). Social cognition and conduct problems: a developmental approach. *Journal of the American Academy of Child and Adolescent Psychiatry, 50*, 385-394. http://dx.doi.org/10.1016/j.jaac.2011.01.006

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Ramaswamy, V., DeSabro, W., & Robinson, W. (1993). An Empirical Pooling Approach for Estimating Marketing Mix Elasticities with PIMS Data. *Marketing Science, 12*, 103-124. http://dx.doi.org/10.1287/mksc.12.1.103

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics, 6*, 461-464. http://dx.doi.org/10.1214/aos/1176344136

Seaman, S. R., White, I. R., Copas, A. J., & Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics, 68*, 129-137. http://dx.doi.org/10.1111/j.1541-0420.2011.01666.x

Sher, K. J., Jackson, K. M., & Steinley, D. (2011). Alcohol use trajectories and the ubiquitous cat's cradle: cause for concern? *Journal of Abnormal Psychology, 120*, 322-335. http://dx.doi.org/10.1037/a0021813

Shevlin, M., Murphy, J., Dorahy, M. J., & Adamson, G. (2007). The distribution of positive psychosis-like symptoms in the population: a latent class analysis of the National Comorbidity Survey. *Schizophrenia research, 89*, 101-109. http://dx.doi.org/10.1016/j.schres.2006.09.014

StataCorp. (2013). *Stata Statistical Software: Release 13.*: College Station, Texas: StataCorp LP.

Stringaris, A., Lewis, G., & Maughan, B. (2014). Developmental pathways from childhood conduct problems to early adult depression: findings from the ALSPAC cohort. *British Journal of Psychiatry, 205*, 17-23. http://dx.doi.org/10.1192/bjp.bp.113.134221

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450-469.

Vermunt, J. K., & Magidson, J. (2013). *Latent GOLD 5.0 Upgrade Manual*: Belmont, Massachusetts: Statistical Innovations Inc.

White, I. R. (2010). simsum: Analyses of simulation studies including Monte Carlo error. *The Stata Journal, 10*, 369-385.

Table 1. Parameter estimates for the effect of gender on the four-class multinomial outcome describing trajectories of conduct problems through childhood.

			Methods based on modal assignment			Methods based on proportional assignment		
Reference class	Comparison class	one-step	Modal standard	Modal ML	Modal ML (robust)	Prop standard	Prop ML	Prop ML (robust)
Parameter estimates for effect of sex								
Low	CL	0.388	0.290 (-25.3)	0.407 (4.9)	0.407 (4.9)	0.197 (-49.2)	0.383 (-1.3)	0.383 (-1.3)
Low	AO	-0.125	-0.062 (-50.6)	-0.158 (26.4)	-0.158 (26.4)	0.019 (-115.0)	-0.127 (1.6)	-0.127 (1.6)
Low	EOP	0.303	0.220 (-27.5)	0.279 (-7.9)	0.278 (-8.3)	0.232 (-23.6)	0.301 (-0.7)	0.301 (-0.7)
CL	EOP	-0.084	-0.070 (-16.4)	-0.128 (52.4)	-0.129 (53.6)	0.034 (-141.0)	-0.083 (-1.2)	-0.083 (-1.2)
AO	EOP	0.429	0.281 (-34.4)	0.437 (1.9)	0.436 (1.6)	0.213 (-50.4)	0.427 (-0.5)	0.428 (-0.2)
AO	CL	0.513	0.352 (-31.5)	0.566 (10.3)	0.565 (10.1)	0.178 (-65.2)	0.510 (-0.6)	0.510 (-0.6)
Standard error for above parameter estimate								
Low	CL	0.125	0.093 (-25.9)	0.132 (5.6)	0.132 (5.6)	0.085 (-32.0)	0.176 (40.8)	0.121 (-3.2)
Low	AO	0.151	0.111 (-26.7)	0.169 (11.9)	0.168 (11.3)	0.099 (-34.6)	0.236 (56.3)	0.151 (0.0)
Low	EOP	0.127	0.109 (-13.9)	0.130 (2.4)	0.130 (2.4)	0.109 (-14.3)	0.145 (14.2)	0.124 (-2.4)
CL	EOP	0.171	0.135 (-21.2)	0.179 (4.7)	0.179 (4.7)	0.128 (-25.0)	0.219 (28.1)	0.166 (-2.9)
AO	EOP	0.203	0.148 (-27.2)	0.225 (10.8)	0.225 (10.8)	0.138 (-32.0)	0.305 (50.2)	0.201 (-1.0)
AO	CL	0.200	0.136 (-32.1)	0.222 (11.0)	0.221 (10.5)	0.085 (-57.6)	0.328 (64.0)	0.199 (-0.5)

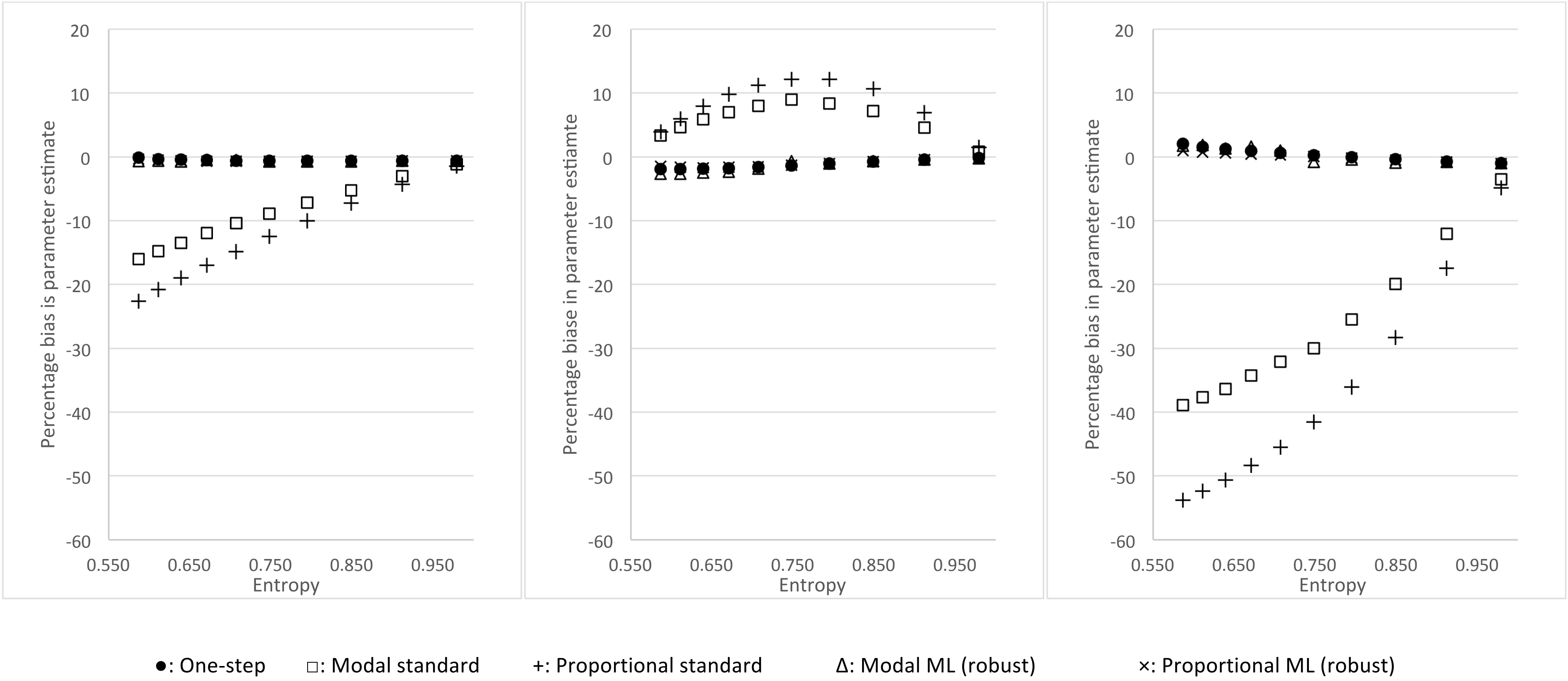
Figures in brackets indicate percentage deviation from the one-step results
CL: Childhood Limited, AO: Adolescent Onset, EOP: Early Onset Persistent

Table 2. The relationship between bias and class-separation for the simple and bias-adjusted three-step methods (effect of covariate Z on class 1 relative to class 3)

		Methods based on modal assignment					Methods based on proportional assignment				
Entropy	Class order	Class overlap	Modal standard		Modal ML (robust)		Class overlap	Proportional standard		Proportional ML (robust)	
			Estimate	% bias	Estimate	% bias		Estimate	% bias	Estimate	% bias
0.979	123	0.00	0.642	-1.1%	0.646	-0.6%	0.00	0.640	-1.4%	0.646	-0.6%
	231	0.00	0.639	-1.6%	0.644	-0.8%	0.00	0.637	-2.0%	0.644	-0.8%
	312	0.03	0.628	-3.2%	0.645	-0.7%	0.04	0.620	-4.5%	0.645	-0.6%
0.912	123	0.00	0.630	-3.1%	0.646	-0.6%	0.00	0.622	-4.3%	0.646	-0.6%
	231	0.00	0.620	-4.6%	0.643	-1.0%	0.00	0.609	-6.2%	0.644	-0.9%
	312	0.11	0.571	-12.1%	0.646	-0.5%	0.17	0.535	-17.7%	0.646	-0.5%
0.849	123	0.00	0.615	-5.2%	0.645	-0.7%	0.00	0.602	-7.2%	0.645	-0.6%
	231	0.01	0.598	-7.9%	0.642	-1.1%	0.01	0.579	-10.8%	0.644	-0.9%
	312	0.20	0.516	-20.5%	0.648	-0.3%	0.29	0.457	-29.7%	0.647	-0.4%
0.795	123	0.00	0.603	-7.2%	0.645	-0.7%	0.00	0.585	-10.0%	0.645	-0.7%
	231	0.03	0.576	-11.2%	0.643	-0.9%	0.04	0.547	-15.7%	0.644	-0.8%
	312	0.26	0.469	-27.8%	0.646	-0.6%	0.38	0.394	-39.3%	0.648	-0.3%
0.748	123	0.00	0.592	-8.9%	0.645	-0.7%	0.00	0.568	-12.5%	0.645	-0.7%
	231	0.05	0.554	-14.7%	0.644	-0.8%	0.07	0.515	-20.6%	0.645	-0.7%
	312	0.32	0.430	-33.7%	0.644	-0.8%	0.45	0.344	-47.0%	0.648	-0.2%

Estimate = average point estimate across 500 replications. % bias = percentage bias relative to true value of 0.649. i.e. (100%*estimate – true-value)/true-value)

Figure 1. Estimated parameter percentage bias = 100%*((estimate – true-value)/ true-value)



First comparison

Effect of covariate Z on class 1 relative to class 3.

True log-odds ratio= 0.649

Second comparison

Effect of covariate Z on class 2 relative to class 3.

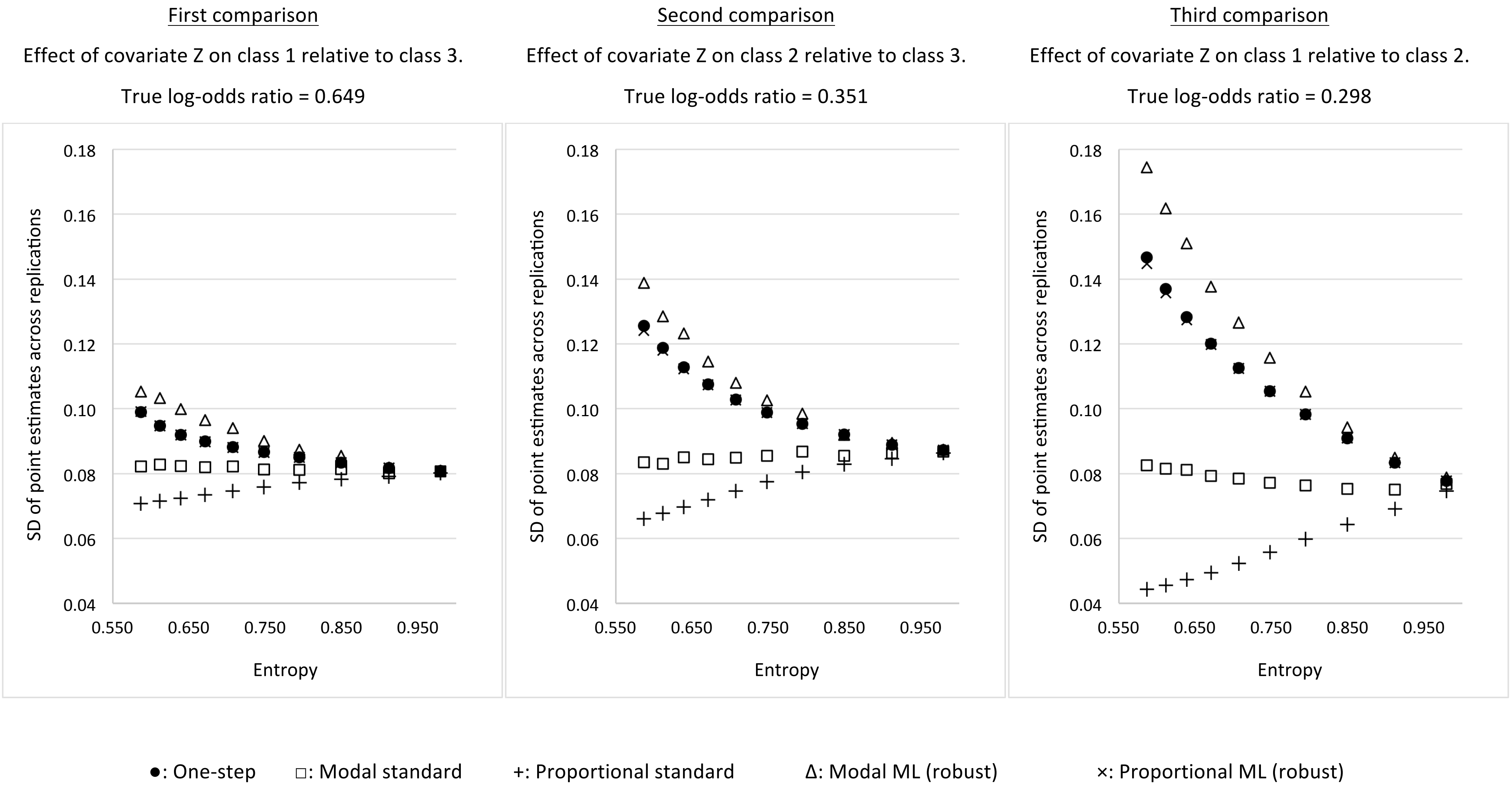
True log-odds ratio = 0.351

Third comparison

Effect of covariate Z on class 1 relative to class 2.

True log-odds ratio = 0.298

Figure 2. Estimated empirical SE (Standard Deviation of the point estimates across 500 replications) for each method



First comparison

Effect of covariate Z on class 1 relative to class 3.

True log-odds ratio = 0.649

Second comparison

Effect of covariate Z on class 2 relative to class 3.

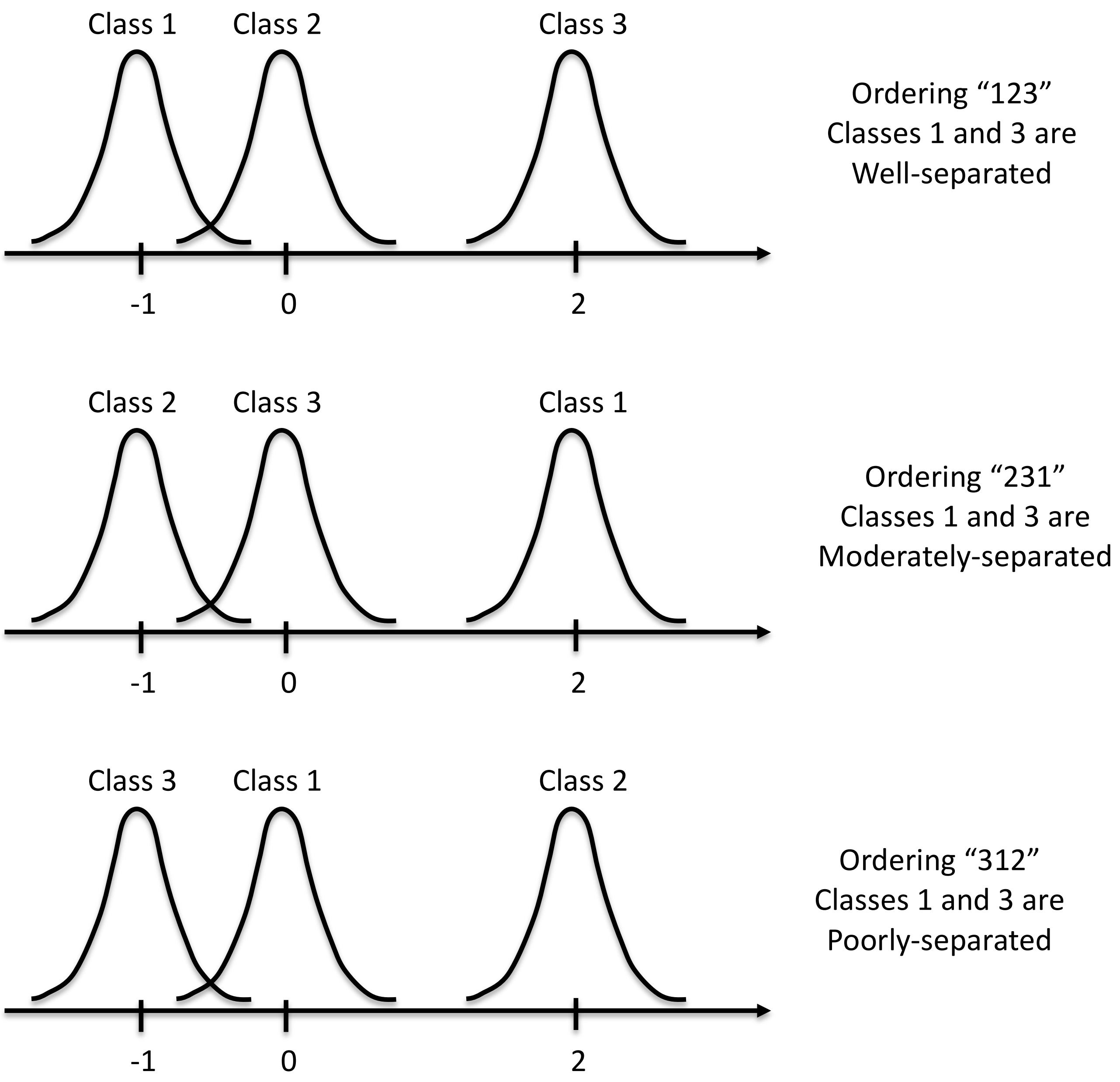
True log-odds ratio = 0.351

Third comparison

Effect of covariate Z on class 1 relative to class 2.

True log-odds ratio = 0.298

Figure 3. Permutation of the class ordering to control class separation



STUDY PROFILE

Zurich Longitudinal Study 'From School to Middle Adulthood'

Nicolas Schmaeh University of Applied Science of Special Needs Education, Switzerland
Kurt Häfeli University of Applied Science of Special Needs Education, Switzerland
haefeli.kurt@teachhfh.ch
Claudia Schellenberg, University of Applied Science of Special Needs Education, Switzerland
Achim Hättich University of Applied Science of Special Needs Education, Switzerland

(Received November 2014 Revised April 2015) http://dx.doi.org/10.14301/llcs.v6i4.330

Abstract

The Zurich Longitudinal Study 'From School to Middle Adulthood' (ZLSE) is a longitudinal study which, to date, encompasses ten surveys from various projects. The study covers a life span from the age of 15th to the 49th year of life and started in 1978 when the participants attended their last compulsory school year; another survey is planned for spring 2015 at the age of 52. The focus lies, on the one hand, on a broad coverage of various personality dimensions supplemented by sociobiographical information in adolescence, and, on the other hand, on the professional and non-professional development from adolescence to adulthood. At this moment, data of 485 people representative of the German-speaking part of Switzerland are available. The aim of this article is to give an overview of the study and to explain in detail the individual surveys.

Keywords

Professional development, longitudinal study, career, personality, adolescence, adulthood, Switzerland

Introduction

The transition from school to work has received much attention over the last few decades, not only on the practical but also on the scientific level (OECD, 2000; Schoon et al., 2009). In countries with a Vocational and Education Training (VET) system this transition starts much earlier and is more strongly related to the economy and the enterprise system than in countries relying on a general education system with a strong academic track. In the dual VET system as it is practiced in countries such as Switzerland, Germany or Austria, a strong emphasis on training in a company is supplemented by teaching in a vocational school which usually lasts three or four years. Therefore a process of matching the interests of an adolescent with a company starts to take place at around age 15-16. This first phase of the transition from school to work, which began in the 1970s for ZLSE participants, marks the start of the study to be presented here. Although not planned, favourable circumstances made it possible to continue the study at irregular intervals to track career development through the ages of 20, 36 and up until 49, from adolescence to middle adulthood. This provides the chance to analyse how this generation of late babyboomers born in around 1963 has dealt with the economic, societal and political developments of the past few decades. Among other things, a shift from an industry-based to a service-based economy took place, along with advancing economic globalization, changes in traditional gender roles and changes in demography. How did today's middle-aged generation (approximately 50 years old) cope with these developments? This generation is now mostly active in professional and family life, and many have children in the educational system. This cohort completed vocational education and training or academic education at the end of the 1970s or the beginning of the 1980s and had, in the following decades, to cope in an active or passive way with many of the aforementioned changes (Leemann & Keck, 2005; Sheldon, 2005).

Study objectives

The Zurich Longitudinal Study 'From School to Middle Adulthood' (ZLSE), which began in 1977, was initially only planned as a short longitudinal study on vocational choice of adolescents. It was later expanded to include a study on personality development during apprenticeship and continued into early and now middle adulthood (Schallberger & Spiess Huldi, 2001). In the meantime ten survey waves (B1-B10, B11 is planned) were carried out in Switzerland at irregular intervals. At this moment the survey spans more than 30 years and covers a life span from age 15 to age 49. The last survey was carried out in summer 2012, and provides data for 485 people. In the following section the study objectives of the main phases are briefly described and summarised in figure 1.

Phase 1: Vocational choice

In the 1970s there was a shortage of qualified young people in Switzerland aiming for an apprenticeship, even though this was still the most popular choice. However, increasing numbers of young people (and their parents) aspired to the gymnasium and the academic track. This was of great concern for the Swiss Trade and Crafts Association and so a research project was initiated. The two universities of Lausanne and Zurich were asked to conceptualize a study "Vocational and professional choice and training of apprentices in Switzerland". It was led by Francis Gendre and Jean-Blaise Dupont from the University of Lausanne and François Stoll from the University of Zurich and was financed by the Swiss Ministry of Economic Affairs. The main goal of the project (1977-1982) was to investigate the determinants and the course of the career choice process (Gendre, 1987; Gendre & Dupont, 1982; Häfeli, 1983). The research was based on the theories and empirical work of several authors and a broad conceptual framework was used (Blau, Gustad, Jessor, Parnes, & Wilcock, 1956; Holland, 1973; Super, 1980). The results demonstrate the importance of the individual (with cognitive, affective and evaluative characteristics), the family, the socio-cultural environment and working environment in predicting the vocational choices of adolescents (Gendre & Dupont, 1982).

Phase 2: Vocational education and training and personality development

Following the studies of Kohn and Schooler (1983) on reciprocal effects of job conditions and personality, the question was if and in what way personality development between the ages 15 and 19 is connected with the working and training situation of the adolescents. The results show a complex interaction of selection and socialisation influences on personality traits such as intelligence, self-esteem and masculinity/femininity, thus supporting Kohn and Schooler's position of reciprocal effects (Häfeli, Kraft, & Schallberger, 1988; Schallberger, 1987; Schallberger, Häfeli, & Kraft, 1984). This research, conducted between 1980 and 1984 was under the leadership of Urs Schallberger (Psychological Institute, University of Zurich) and was financed by the Swiss National Foundation.

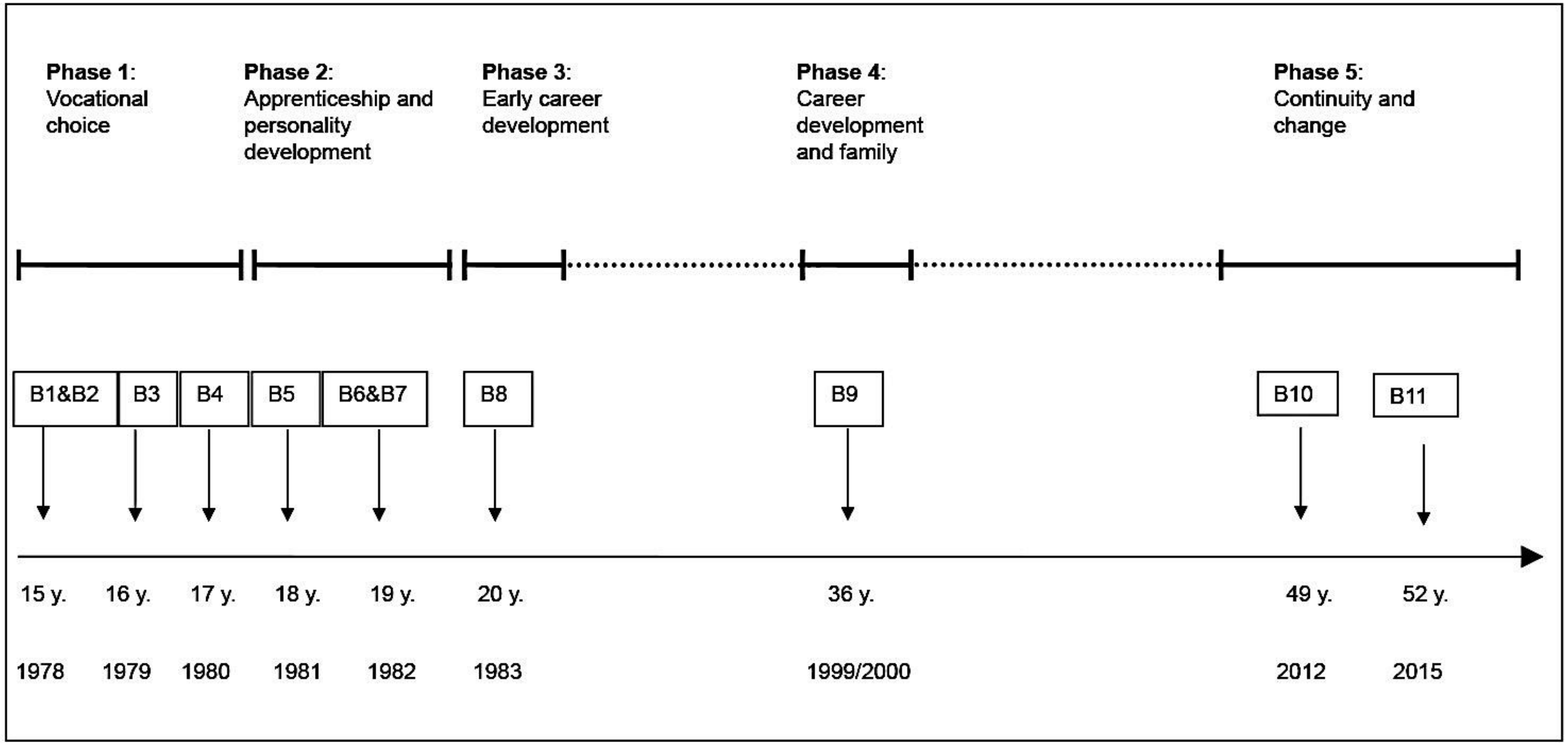
Phase 3: Early career development

The subject of this phase of the survey was the first career steps and adaptation processes of the young adult participants after finishing their vocational training or their next steps after finishing a general education in a gymnasium. The results support the developmental theory of occupational aspirations by Gottfredson (1971) whereby gender roles and social class restrict, to a large degree, the range of acceptable occupations for young people (Gendre, 1987; Gottfredson, 1971). This project (1982-1985) was again conducted by Francis Gendre of the University of Lausanne and financed by the Swiss National Foundation.

Phase 4: Career development and family

The aim was to consult the now 36-year-old participants of phase 2 about their professional career development and their actual situations. This information could then be related to adolescent factors, e.g. the influence of risk and protective factors in youth on satisfaction and success in young adulthood thus supporting Werner's work on resilience (Spiess Huldi, Häfeli, & Rüesch, 2006; Werner & Smith, 2001). In another analysis titled "The power of personality" (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007) traits such as conscientiousness or emotional stability could be demonstrated for the attainment of occupational status in adulthood (Spiess Huldi, 2009). This study (1998-2002) was under the leadership of Urs Schallberger and Claudia Spiess Huldi of the University of Zurich (Schallberger & Spiess Huldi, 2001).

Figure 1. Synopsis of the five phases


Phase 1: Vocational choice
Phase 2: Apprenticeship and personality development
Phase 3: Early career development
Phase 4: Career development and family
Phase 5: Continuity and change
B1&B2
B3
B4
B5
B6&B7
B8
B9
B10
B11
15 y.
16 y.
17 y.
18 y.
19 y.
20 y.
36 y.
49 y.
52 y.
1978
1979
1980
1981
1982
1983
1999/2000
2012
2015

Phase 5: Continuity and change

Following the research of Super (1980) and Holland (1973) on career development and Schoon et al. (2009) on life-span development, we are interested in this phase to investigate the career and personality development from adolescence into middle adulthood. How much change and continuity can be observed? What are possible influences on horizontal and vertical career mobility? How can persistent gender segregation be explained? To answer these questions the participants of the last survey were questioned 13 years later, shortly before reaching the age of 50 (B10). Our findings show that wide-spread vertical gender segregation (Charles & Bradley, 2009) is a result not only of personality dimensions (measured in adolescence) but also of traditional gender roles in adulthood (Häfeli, Hättich, Schellenberg, & Schmaeh, 2015). We also find much continuity in career development during more than 30 years, as the majority of the sample is still in the same occupational field (using Holland’s typology) – despite the massive economic changes during this period (Schellenberg, Schmaeh, Häfeli, & Hättich, 2015). An additional, expanded survey (B11) is planned in 2015 at the age of 52 (see “Outlook”). The project (conducted between 2011 and 2017) is financed by the Swiss State Secretariat for Education, Research and Innovation (SERI) and is directed by Kurt Häfeli and Claudia Schellenberg (University of Applied Sciences of Special Needs Education Zurich) and Alexander Grob (Psychological Institute, University Basel).

Survey content

Next we present the main topics and dimensions that were covered in each phase (see also table 1).

Phase 1: Vocational choice

The main goal of the first phase was to investigate in a broad terms the determinants and the course of the vocational and professional choice process. Therefore sociobiographical indicators, such as gender, age, and family background, were included. Standardized methods to measure cognitive abilities come from the intelligence structure test IST-70 (Amthauer, 1970) and the vocational and professional ability test BET (Schmale & Schmidtke, 1967). For the measurement of the ‘Big Five’ (extraversion, neuroticism, agreeableness, conscientiousness, openness to experience) and other personality dimensions a short version with 155 items of the Adjective Check List / ACL (Gough & Heilbrun, 1980) was used. Other dimensions (see table 1) included locus of control (Reid & Ware, 1974), attitudes toward gender roles (Häfeli, 1983), personal and professional values, occupational interests, achievement motivation, self-esteem (Gendre & Dupont, 1982), leisure time activities and parent-child relations (Roe & Siegelman, 1963). Finally, the adolescents in B1 and the first and second follow-up (B3, B4) were asked in detail about their career search activities.

An outside perspective was gained by asking the classroom teachers (B2) to rate their students individually on 19 different aspects (personality, work attitudes, school grades and abilities, career prognosis).

Phase 2: VET and personality development

To investigate the reciprocal effects of training conditions and personality development from the 15th to the 19th years of life, study participants were questioned about their occupational histories in the fifth and sixth survey. As a central part, a repeated measurement of the most important personality dimensions from B1 was carried out in B6 (intelligence, personality, values, gender roles etc.). Different aspects of the work and training conditions were also measured, such as content and complexity of the work task (Kohn & Schooler, 1983), motivational work dimensions (Hackman & Oldham, 1975) and social climate (Moos, 1979). In the seventh round of data collection, experts with a broad occupational knowledge were asked to assess the 44 professions and schools represented in the sample survey with regard to 20 dimensions relevant for personality development (Häfeli & Schallberger, 1983).

Phase 3: Early career development

The aim of the eighth survey was to record the actual life and working situation as well as the wellbeing of the now 20-year-old young adults. To do this, information about the study participants’ activity since schooldays and their actual living situation (inclduding how they spent their leisure time) was collected. In addition, questions were asked about the following areas: mental and physical health, self-concepts and satisfaction with the various aspects of life. Finally, the young adults were asked about their professional plans.

Table 1. Summary of study variables

		Surveys										
Topics	**Constructs/dimensions**	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**	**B11**
Sociodemographic aspects	age, gender	X							X	X		
	family of origin (structure, family climate)	X	X				X					
	partnership/family	X					X		X	X	X	X
Abilities and skills	cognitive skills	X					X					
	career adaptability											X
Values, attitudes, interests	personal values	X					X		X			X
	occupational values	X					X		X			X
	self-esteem	X					X		X			
	achievement motivation	X		X			X		X			
	sex-role attitudes	X					X					X
	locus of control/self-efficacy	X					X					X
	occupational interests	X										
Personality in a narrow sense	personality (“Big Five“)	X	X				X				X	X
	masculinity, femininity	X					X					X
Career search and finding	career plans	X		X	X	X	X		X			X
	procedure in first career choice	X		X								
	assessment of first career choice			X	X	X	X					
Professional activities and trainings	professional activities, training, continuing education and training CET, professional development	X			X	X	X	X	X	X	X	X
Work and training characteristics	work and training contents					X	X	X	X			
	working conditions					X	X	X	X			X
	commitment to household, family									X	X	X
Well-being, life satisfaction	health	X					X		X		X	X
	satisfaction	X			X	X	X		X	X	X	X
Answering behaviour	request to get a feedback	X								X	X	X
	readiness for future participation									X	X	X

Phase 4: Career development and family

The main focus of the ninth survey centered upon career development and actual life situation. In this context, the on-average 36-year-old cohort members were questioned about their occupational histories since their 18th year of life (a listing of all activities). Also included were details about the degree of employment, their particular function in the company and their actual wages. The interaction between the professional and the private (partnership, children) areas of life was also explored in this survey, as well as details about leisure time activities. In order to record satisfaction levels, the participants were invited to describe their degree of satisfaction with regard to profession, family and so on.

Phase 5: Contuinity and change

In the tenth survey, all data relevant to job-related and non job-related development from the 36th to the 49th years of life were of interest. For this reason, the occupational histories, since the last survey, were questioned. Moreover, participants were asked, as in B9, about their employment situation (wages, function etc.), activities in leisure time, partnership and children. In the tenth survey, some personality dimensions were also investigated (Rammstedt & John, 2007). Since wellbeing was an important topic, satisfaction at work was recorded in a detailed way, as was general satisfaction with life (Diener, Emmons, Larsen, & Griffin, 1985). Furthermore, some questions regarding mental and physical health were added. Finally, participants were invited to give some information about their vocational or personal intentions in the years to come.

Reference population, sample and data collection

The target group for this study was students in their last compulsory school year (ninth grade, approximately age 15) in Switzerland. As the first data collection took place in 1978, a large part of the sample was born in the year 1963. In order to get a representative sample and to make interregional comparisons possible, Switzerland was split up in to 88 regions according to geographical (urban / middle land / mountain) areas and economic (primary / secondary / tertiary economic) sectors (Werczberger, 1964; Wronksy, 1967). Based on these criteria, 18 representative regions were chosen in ten (out of 25) cantons. In the German-speaking part of Switzerland these were Basel (representing the urban region), various regions in Central Switzerland (from the cantons Aargau, Berne, Glarus and Saint Gallen) and the mountain region (Bernese Oberland). For the French speaking part of Switzerland regions from the cantons Geneva, Vaud, Valais and Neuchatel were selected. The small Italian- and Romansch-speaking parts of Switzerland (6%), however, were excluded.

Within the selected regions communities, and within these communities classes, of the ninth grade were chosen on a random basis. This resulted in 2,357 students from 123 classes, namely 1,706 from the German-speaking and 651 from the French-speaking parts of Switzerland (see table 2). This sample was used for phase 1 (vocational choice) and 3 (early career development). For economic reasons, from phase 2 onwards, only the participants from the German-speaking part of Switzerland (N=1706) were contacted. In phase 2 the sample was also further reduced for practical reasons (N=504) as personality development was the focus and this required extensive testing and questioning in small groups (see phase 2 below). The same smaller sample was kept for phase 4. However, in the latest phase 5, at middle adulthood, we tried to expand the sample by going back to the original broader Swiss-German speaking sample (see table 1). For the last survey (B10), addresses of 84% of the target sample could be found and data of 485 people (76%) was collected. This sample is a good representation of the original sample B1, and therefore the age group born around 1963, in terms of gender, social background and type of secondary school visited. During their 36th year of life, 76% were employed (40% of the women interrupted their career because of children/family. At 49 years, 92% of the participants reported being employed (men mostly full-time, women mostly part-time).

Phase 1: Vocational choice

For the first survey (B1) in summer 1978 - at the beginning of their last compulsory school year (ninth grade, age 15) - the participants were questioned and tested in class (see table 2). This lasted for one school day (or six hours). The testing was administered and supervised by advanced psychology students. For the second survey (B2) which took place one month after B1, the teachers received a one-page questionnaire in order to be able to rate their respective students individually.

Shortly before the end of the ninth school year in February/March 1979 the third survey (B3) took place. During this process the students received a questionnaire of seven pages which had been sent to the class teachers and was distributed in class (87% response rate). The fourth survey (B4) took place in September/October 1979, six months after the end of the ninth school year. The participants received a questionnaire of three pages sent to their home address (72% response rate).

Phase 2: VET and personality development

The fifth survey (B5, see table 2) took place in March 1981. The 12-page questionnaire was sent by post to the 1,706 former ninth graders from the German-speaking part of Switzerland (return rate 75%, 1,284 people). For the sixth survey (B6), a repeated measuring of the distinctive features of the people from B1 was sought. This called for tests and questionnaires with standardized conditions which could not be handled by post. For economic reasons the random sample had, therefore, to be reduced. In this context, 691 adolescents were chosen from the initial random sample, who could present a fairly stable career pattern and who came from the most popular 36 occupations. In addition two groups were chosen: full-time students from the Gymnasium or Teachers College as well as adolescents without further education. The adolescents were asked by telephone to take part in survey sessions in small groups (seven – 12 people) lasting 2.5 hours. Finally, 504 adolescents at the end of their upper secondary level (average age 19) agreed to participate. The seventh survey (B7) consisted of an expert rating of 20 aspects of the 44 most frequent occupational professions/schools. For this, 28 experts with a broad professional knowlege were contacted in order to rate each occupation and the schools in the form of a Q-sort. For this task the experts needed on average half a day.

Phase 3: Early career development

The eighth survey (B8) took place in October 1983. A questionnaire of 19 pages was sent by post to the 2,357 people from the initial random sample. The return rate was 65% of the total group.

Phase 4: Career development and family

The ninth survey took place in the autumn 1999 (see table 2). The 504 adolescents selected in phase 2 (B6) were once more asked to participate after a gap of 16 years. The on-average 36-year-old participants received a questionnaire of four pages by post, after their addresses had been updated. This updating was quite successful as 443 (88%) addresses could be verified after such a long time (for 54 (11%) people no address could be found; 7 (1%) people had died). 394 people answered the questionnaire which corresponds to a return rate of 89% (Schallberger & Spiess Huldi, 2001).

Table 2. Data collection and sample

Survey	Time	Age	Method	Length	Target group	Particip	% Return
Phase 1 Vocational choice							
B1	1978 May/June 1978	15 (9th grade)	Classroom (survey, tests)	6 hrs /26 quest./tests	2357 (a.1706 German; b. 651 French)	2357	100%
B2	1978 June/July	15	Teacher rating	1 page	Teachers rated students B1	2048	87%
B3 1st follow-up	1979 Febr/March	15;10	Classroom survey	7 pages	B1 2357	2168	92%
B4 2nd follow-up	1979 Fall	16;6	Postal survey	3 pages	B1 2357	1704	72%
Phase 2 Vocational education and training and personality development							
B5	1981 March	18	Postal survey (telephone)	12 pages	B1a German speaking	1284	75%
B6	1982 Spring-Fall	19	Small groups (face-to-face)	2.5 hr, 25 quest./tests	Selected group of B5: 691	504	73%
B7	1982 Fall	-	Postal survey/Q-sort	4 hours	31 Professional experts	28	90%
Phase 3 Early career development							
B8	1983 October	20	Postal survey	19 pages	B1 2357>>2205 Addresses found (94%)	1428	65% of 2205 (61% of 2357)
Phase 4 Career development and family							
B9	1999/2000 Sept-March	36	Postal survey (telephone)	4 pages	B6: 504 >>443 Addresses found (88%)	394	89% of 443 (78% of 504)
Phase 5 Continuity and change							
B10	2012 April-July	49	Postal survey (telephone)	8 pages	B6 (504) plus target sample B1a (250) >>637 Addresses found (84%)	485	76% of 637 (64% of 754)
B11	2015 May-July	52	Postal survey (telepone)	Appr. 20 p.	B10 plus rest of B5 (N=1284)		

Phase 5: Contuinity and change

The last survey (B10) so far took place in 2012 (see table 2). As the sample of the ninth survey was not representative in all points, an under-represented subsample of 125 people from the initial random sample was drawn in order to counteract this. For this subsample, women with lower educational levels from the German-speaking parts of Switzerland were selected. In addition, another random sample was drawn out of the initial sample (B1) in order to increase the sample size. Altogether 754 persons were chosen to participate in the tenth survey. After a time-consuming search, the addresses of 637 (84%) former participants were found (Schmaeh, Hättich, Häfeli, & Schellenberg, 2013). The search for these addresses was carried out with the help of an online program specializing in looking for addresses by making inquiries at the last known municipality. Nevertheless, 117 cases could not be contacted: 21 people (3%) had died and for 96 people (13%) the current address could not be found. The survey was carried out, in most cases, by a six-page questionnaire. Altogether 485 out of the 637 people contacted completed the questionnaire, which equates to return rate of 76%.

Panel maintenance and incentives

With two exceptions the participants asked did not receive any remuneration for their participation in the study. Nevertheless, a considerable response rate (see table 2) was achieved thanks to repeated enquiries (for the most part after two written reminders and additional phone calls). With return rates of 89% for the ninth survey and 76% for the tenth survey, it can be spoken of as a success in the maintanance of the sample. Despite the fact that, in the case of the tenth survey, some of the participants had not taken part in the study for 30 years, many were still motivated to take part in a further survey. A monetary remuneration for participation only took place for the fifth and sixth survey. In B5 the young participants could win three rewards of CSF 200 (approximately $ 200) in a lottery. In B6 the adolescents who participated received an amount of CSF 50 (approximately $ 50), because the questioning lasted half a day and took place in leisure time, sometimes necessitating a journey.

The success may also be partly due to the fact that before and after each survey all participants were informed in short letters (several times in the form of comics when the participants were adolescents) about the goals of the study and some selected results. At B5 the adolescents received personalized ability and interest scores. For the last two surveys (B9 and B10), the participants were informed about the new survey and its aims before participating. In addition, the importance of the participation of every single person was emphasized. The questionnaire was then sent to them, and after the deadline had expired, the participants received two reminders requesting that they complete the questionnaire. If after that, no response was forthcoming, they were contacted by the phone and a short version of the questionnaire was filled out.

Following the survey, all people whose addresses could be found received an informative booklet detailing the initial results. In addition, the participants were constantly referred to the homepage of the study (www.zlse-hfh.ch) which informed them about the latest results.

Outlook and data availability

To sum up, the Zurich Longitudinal Study ‘From School to Middle Adulthood’ is in many ways successful. Thanks to the longitudinal character of the study, determinants for the vocational course of a life span of over 30 years can be identified which is unique for Switzerland. With the help of the broad gathering of personality variables and sociobiographical indicators in adolescence, the predictors which influenced the further vocational course and status can be identified. In contrast to many countries with a system of general education at the secondary level, this study is situated in the context of an apprenticeship system with early vocational choices. Thanks to the meticulous gathering of the occupational histories between the 15th and the 49th year of life, statements regarding continuitiy and discontinuity of vocational careers can be made (Häfeli et al., 2015; Schellenberg, Häfeli, Schmaeh, & Hättich, 2013; Schellenberg et al., 2015). In addition, a stronger focus on health aspects (physical and mental health, exercise behaviour, substance abuse), in the more recent surveys also helps to investigate the influence of risk and protection factors in the vocational and personal development on health aspects in adulthood.

Due to the representative sampling of Swiss school classes in the ninth grade, there are data of approximately 2,400 young people available that capture their vocational start, abilities and personality in the broader sense as well as their sociobiographical background. Thanks to the additional surveys in adulthood (B9 and B10), important supplementary data for the further vocational development and the private situation of a selected random sample exist. Above all, detailed information about the careers of individuals over a life span of 30 years answer many exciting questions. Consequently, at present there is a sample of 485 people each with altogether approximately 3,500 variables covering the whole period of time on hand.

As a result of of limited funding the data has not been as fully analysed or published as widely as we would have hoped to date. Due to its complexity, the data set has not yet been described and prepared in a way that it can be made available for other researchers. A proposal to finance this work is planned for 2016 whereby the data would be made available via FORS, the Swiss Centre of Expertise in the Social Sciences.

The planned eleventh survey will serve to widen the random sample in order to facilitate more specific investigation of careers in different occupational groups. Furthermore, with the planned questioning in the year 2015, another important life stage will be highlighted. On reaching the age of 50, questions about any career development still possible and future retirement become particularly relevant for the participants. It will be interesting to find out how they will respond to these topics. The 2015 survey will also repeat some of the personality measurement from adolescence to study the reciprocal effects of job conditions and personality over a long period. In addition, information regarding partners and children will be collected. Through this, statements regarding the co-development of careers will also become possible.

Even though the ZLSE study concerns a representative random sample, it is about a specific cohort born in the year 1963 in the context of Switzerland. For this reason, explicit comparisons with other cohorts must be made in order to be able to judge the relevance of the results. In Switzerland comparable projects, such as TREE or the COCON study, started a few years ago with similar questions and will make comparisons possible with younger cohorts (Bergman, Hupka-Brunner, Keller, Meyer, & Stalder, 2011; Buchmann & Kriesi, 2009, 2012).

References

Amthauer, R. (1970). *Intelligenz-Struktur-Test IST-70*. Göttingen: Hogrefe.

Bergman, M. M., Hupka-Brunner, S., Keller, A., Meyer, T., & Stalder, B. E. (Eds.). (2011). *Transitionen im Jugendalter. Ergebnisse der Schweizer Längsschnittstudie TREE*. Zürich: Seismo.

Blau, P. M., Gustad, J. W., Jessor, R., Parnes, H. S., & Wilcock, R. C. (1956). Occupational choice: a conceptual framework. *Industrial Laboor Relations, 9*, 531-543. http://dx.doi.org/10.1177/001979395600900401

Buchmann, M., & Kriesi, I. (2009). Escaping the gender trap: young women's transition into non-traditional occupations. In I. Schoon & R. K. Silbereisen (Eds.), *Transition from School to Work* (pp. 193-215). Cambridge/New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511605369.009

Buchmann, M., & Kriesi, I. (2012). Geschlechtstypische Berufswahl Jugendlicher. Begabungszuschreibungen, Aspirationen und Institutionen.*Kölner Zeitschrift für Soziologie und Sozialpsychologie*(Sonderheft), 256-280. http://dx.doi.org/10.1007/978-3-658-00120-9_11

Charles, M., & Bradley, K. (2009). Indulging our gendered selves? Sex segregation by field of study in 44 countries. *American Journal of Sociology, 114*, 927-976. http://dx.doi.org/10.1086/595942

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment 49*, 71-75. http://dx.doi.org/10.1207/s15327752jpa4901_13

Gendre, F. (1987). L'orientation et le devenir des apprentis: dix ans de recherches à l'IPAUL. *Schweizerische Zeitschrift für Psychologie, 46*(3/4), 155-172.

Gendre, F., & Dupont, J.-B. (1982). Structures de l'individu et choix professionnel *Rapport de recherche* (pp. 627). Lausanne: Université de Lausanne, Institut de Psychologie Appliquée.

Gottfredson, L. D. (1971). Circumscription and compromise: A developmental theory of occupational aspirations. *Journal of Counseling Psychology, Monograph 28*(6), 545-579. http://dx.doi.org/10.1037/0022-0167.28.6.545

Gough, H. C., & Heilbrun, A. B. (1980). *The Adjective Check List Manual*. Palo Alto: Consulting Psychologists Press.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*, 159-170. http://dx.doi.org/10.1037/h0076546

Häfeli, K. (1983). *Die Berufsfindung von Mädchen: zwischen Familie und Beruf*. Frankfurt, Bern, New York, Paris, Wien: Peter Lang.

Häfeli, K., Hättich, A., Schellenberg, C., & Schmaeh, N. (2015). Gründe für zunehmende vertikale Geschlechtersegregation im Erwachsenenalter. *Schweizerische Zeitschrift für Bildungswissenschaften 37*(2).

Häfeli, K., Kraft, U., & Schallberger, U. (1988). *Berufsausbildung und Persönlichkeitsentwicklung. Eine Längsschnittstudie*. Bern, Stuttgart, Toronto: Huber.

Häfeli, K., & Schallberger, U. (1983). Merkmale von Berufen. Eine Befragung von berufskundlichen Experten. *Berufsberatung und Berufsbildung, 68*, 281-294.

Holland, J. L. (1973). *Making Vocational Choices. A Theory of Careers*. New York: Englewood Cliffs.

Kohn, M. L., & Schooler, C. (1983). *Work and Personality. A inquiry into the impact of social satisfaction*. Norwood, NY: Ablex.

Leemann, R. J., & Keck, A. (2005). *Der Übergang von der Ausbildung in den Beruf. Die Bedeutung von Qualifikation, Generation und Geschlecht*. Neuchâtel: Bundesamt für Statistik.

Moos, R. H. (1979). *Evaluating educational climates*. San Francisco: Jossey-Bass.

OECD. (2000). *From Initial Education to Working Life. Making Transitions Work*. Paris: OECD.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version oft he Big Five Inventory in English an German. *Journal of Research in Personality, 41*, 203-212. http://dx.doi.org/10.1016/j.jrp.2006.02.001

Reid, D., & Ware, E. E. (1974). Multidimensionality of internal versus external control: Addition of a third dimension and non-distinction of self versus others. *Canadian Journal of Behavioral Science, 6*(131-142). http://dx.doi.org/10.1037/h0081862

Roberts, B. W., Kuncel, N., Shiner, R., Caspi, A., & Goldberg, L. (2007). The power of personality. The comparitive validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives in Psychological Science, 2*, 313-345. http://dx.doi.org/10.1111/j.1745-6916.2007.00047.x

Roe, A., & Siegelman, M. (1963). A parent-child relations questionnaire. *Child Development, 34*, 355-369.

Schallberger, U. (1987). Berufsarbeit und Persönlichkeit - Aspekte einer komplexen ökologischen Problemstellung. *Schweizerische Zeitschrift für Psychologie, 46*(1/2), 91-104.

Schallberger, U., Häfeli, K., & Kraft, U. (1984). Zur reziproken Beziehung zwischen Berufsausbildung und Persönlichkeitsentwicklung. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 4*, 197-210.

Schallberger, U., & Spiess Huldi, C. (2001). Die Zürcher Längsschnittstudie "Von der Schulzeit bis zum mittleren Erwachsenenalter". *Zeitschrift für Soziologie der Erziehung und Sozialisation, 21*(1), 80-89.

Schellenberg, C., Häfeli, K., Schmaeh, N., & Hättich, A. (2013). Auswirkungen von erschwerten Startchancen auf den beruflichen Erfolg im mittleren Erwachsenenalter: ein Längsschnitt über 34 Jahre. *Schweizerische Zeitschrift für Heilpädagogik, 19*(11-12), 26-35.

Schellenberg, C., Schmaeh, N., Häfeli, K., & Hättich, A. (2015). Horizontale und vertikale Mobilität in Berufsverläufen vom Jugendalter bis zum 49. Lebensjahr: Ergebnisse einer Längsschnittstudie. In K. Häfeli, M. P. Neuenschwander & S. Schumann (Eds.), *Berufliche Passagen im Lebenslauf* (pp. 305-333). Wiesbaden: VS Verlag für Sozialwissenschaften. http://dx.doi.org/10.1007/978-3-658-10094-0_12

Schmaeh, N., Hättich, A., Häfeli, K., & Schellenberg, C. (2013). *Technischer Bericht zu Adressrecherche, Rücklauf und Fragebogen. Bericht Nr. 1 aus dem Projekt "Kontinuität und Wandel: Determinanten* *der beruflichen und persönlichen Entwicklung"*. Zürich: Interkantonale Hochschule für Heilpädagogik (HfH).

Schmale, H., & Schmidtke, H. (1967). *Berufseignungstest - BET*. Bern: Huber.

Schoon, I., Salmela-Aro, K., Silbereisen, R. K., Eccles, J., Schneider, B., Trautwein, U., & Bergman, L. (2009). *Pathways to adulthood: Towards a unifying framework*. London: Institute of Education, University of London.

Sheldon, G. (2005). *Der berufsstrukturelle Wandel der Beschäftigung in der Schweiz 1970-2000. Ausmass, Ursachen und Folgen*. Neuchâtel: Bundesamt für Statistik.

Spiess Huldi, C. (2009). *Erfolg im Beruf. Zum Einfluss von Persönlichkeit und psychosozialem Umfeld auf die berufliche Entwicklung Jugendlicher*. Zürich/Chur: Rüegger.

Spiess Huldi, C., Häfeli, K., & Rüesch, P. (2006). *Risikofaktoren bei Jugendlichen und ihre Auswirkungen auf das Leben im Erwachsenenalter. Eine Sekundäranalyse der Zürcher Längsschnittstudie "Von der Schulzeit bis zum mittleren Erwachsenenalter" (ZLSE)*. Luzern: Edition SZH/CSPS.

Super, D. E. (1980). A life-span, life-space approach to career development. *Journal of Vocational Behavior, 16*, 282-298. http://dx.doi.org/10.1016/0001-8791(80)90056-1

Werczberger, E. (1964). Untersuchung über die Pendlerregionen und die Einteilung der Schweiz in Arbeitsmarktregionen. Zürich: ORL-Institut der ETH.

Werner, E. E., & Smith, R. S. (2001). *Journeys from Childhood to Midlife. Risk, Resilience and Recovery*. Ithaca, N.Y.: Cornell University Press.

Wronksy, D. (1967). Industriestandorte. Zürich: ORL-Institut der ETH.

COMMENT AND DEBATE

Population sampling in longitudinal surveys

Harvey Goldstein University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk
Peter Lynn University of Essex, UK
Graciela Muniz-Terrera University of Edinburgh, UK
Rebecca Hardy University College London, UK
Colm O'Muircheartaigh University of Chicago, US
Chris Skinner London School of Economics, UK
Risto Lehtonen University of Helsinki, Finland

http://dx.doi.org/10.14301/llcs.v6i4.345

When and why do we need population representative samples?

Harvey Goldstein University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk

(Received February 2015 Revised April 2015)

Abstract

The paper questions the need for observational studies to achieve representativeness for real populations, in particular for longitudinal studies. It draws upon recent debates and argues for the need to distinguish scientific inference from population inference.

Keywords

Observational studies, longitudinal, representativeness.

Introduction

In a recent issue of the International Journal of Epidemiology (2013, vol 42, 1012-1028) there was a debate about whether analysts have overrated, in epidemiology and social and medical science more generally, the importance of having representative samples from well-defined 'real' populations. In this paper the arguments are summarised and developed to understand how they might affect, in particular, longitudinal studies.

Setting out the arguments

The lead paper in this collection by Rothman, Gallacher and Hatch (2013a) argues that efforts to obtain samples that are representative of real populations are often misplaced and that scientific research questions in epidemiology (and the human sciences more generally) are usually better tackled by sampling purposively. By this they mean selecting groups for study that are directly relevant for the comparisons or relationships of interest, rather than attempting to estimate such relationships within any specific 'real' population. They claim that the key scientific criterion should be the attempt to replicate (generalise) findings across different populations and groups. Any failure to replicate can then lead to a study of those factors that differ among groups and which might explain varying relationships. Thus, for example, replication across different ethnic groups, need not involve representative samples from a population containing such groups, but rather ensuring that data are representative of the groups in question and not subject, for example, to selection bias.

They suggest that traditional emphasis on statistical significance and obtaining population-unbiased estimates downplays the importance of the scientific need for generalisation and replication. As an example they talk about sampling equal numbers in age groups rather than attempting to match the distribution to the distribution within a population. This particular argument, however, seems weak since, in fact, like the example of ethnic groups, this can be regarded simply as a stratified population sample which, combined with suitable weights, can also be used to make population inferences. They also appear to be concerned largely with the situation where there are pre-existing hypotheses or comparisons of interest, whereas in reality populations are often representatively sampled in order to allow exploratory analyses that rely on sufficient diversity and heterogeneity within the population.

They also seek to make a clear distinction between descriptive statistics that require representative samples and analytical statistics that attempt to address scientific hypotheses. In fact, this distinction is often far from clear and I shall return to this point later where I also discuss what exactly is meant by a 'population'.

Four sets of authors provide responses to Rothman's paper, three of whom are broadly supportive (Elwood; Nohr & Gleen; Richiardi, Pizzi & Pearce, 2013). I shall deal with these first, then look at the paper (Ebrahim & Davey-Smith, 2013) that takes a somewhat different view and then refer to a rebuttal by Rothman and colleagues (2013b).

Elwood (2013) makes the point that that any real population is a historical entity, and when inferences about it are available it may have changed in important ways. Of course, for enumeration purposes, this may still be the best information available. For scientific purposes, however, the real population serves as an instance of an underlying process that generates a data set at a particular time, and where inference is to all possible instances. This is often referred to as a superpopulation approach and the actual real population is treated as if it were a sample from such a conceptually infinite population. Thus, the actual population serves as a useful data set for exploratory purposes or to test hypotheses within a heterogeneous sample.

All these three respondents point to the importance of taking account of possible confounders and see this as a key concern for scientific purposes. There is some discussion about choosing unrepresentative samples with a high response rate as being preferable to choosing representative samples with a low response rate. The idea of a purposive sample that can achieve a high response rate is an interesting one, but its success depends crucially on knowing the relevant characteristics of the sample. Examples where this might be the case are the use of internet-based surveys and in some cases of clinical trials. In longitudinal studies it is similar to the way in which attrition may be handled. Such studies often settle down to having a fairly stable sample that has a high response rate in repeated waves. Because the initial sample is often fairly representative the characteristics of these initial respondents can be used to 'adjust' subsequent analyses to avoid attrition biases.

The contribution by Ebrahim and Davey-Smith (2013) seeks to disagree with Rothman and colleagues on several points. They discuss the cases where non-representative samples, in particular randomised controlled trials (RCT), give results different to those from representative samples. They suggest that 'volunteer bias' may distort non-representative studies, including RCT's, and that representative sample inferences may be more trustworthy. They point to the example of the United Kingdom biobank which is not only unrepresentative but also has a very low response rate of 6%. They claim that this will not matter in terms of genetic associations since these are unlikely to be associated with selection and not susceptible to influence by confounders such as, for example, social class. Both of these statements, however, seem disputable, especially in terms of gene-environment interactions, and would require strong supporting evidence for general acceptance.

The final rebuttal by Rothman et al. (2013b) reiterates many of the original points. They use the example of the Doll/Hill smoking and lung cancer study to emphasise the importance of representativeness, although this is really an argument about observational studies versus RCTs and doesn't add anything new. They also discuss the meaning of the statistical term 'bias' and the importance of being clear what this refers to. This is an important issue and I will return to it below.

Defining populations

It is pertinent to ask what is meant by the term 'population' and the associated issue of what is meant by 'bias'. From a statistical viewpoint these are technical terms. Statistical analysis aims to provide estimates for a collection of units (people, institutions etc.) that, at least notionally, can be resampled. Any particular sample is regarded (perhaps conditional on particular variable values, such as belonging to a given age group) as randomly selected in the case of classical inference or as being 'exchangeable' in terms of Bayesian inference. This collection of units is a population. It may be real in the sense that it can repeatedly be sampled or conceptual in the sense that any realised sample is considered to be drawn at random from it (exchangeable with respect to all other possible draws) – a superpopulation. For example, we can define the population of women who smoke in the second trimester of pregnancy as all women who have, or could ever be observed to have, this characteristic. Any scientifically generalisable statement will be one about the distribution of any of their characteristics and relationships. The term 'bias' is defined in terms of the extent to which the estimates obtained from any particular sample differ from the (unknown) distribution in this population. Thus, from a statistical viewpoint, the population does have to be well defined in terms of being able to describe its characteristics, but it does not have to correspond to any actual 'real' population. Unfortunately, there is sometimes confusion between these uses of the term population, but here I use it in the sense of a well-defined collection of units rather than any human population that actually exists or has existed.

In the case of longitudinal data there is a special problem. Suppose we sample randomly from a real population, for example all births in a given country. After the first contact with respondents, the relationship with this real population will change. Thus, some individuals will emigrate and when reporting on relationships across time, in terms of population representativeness we will need to choose whether the relevant population consists of those individuals present in the country at a subsequent occasion, including immigrants, or those who were present at the start and did not emigrate. If it is the latter then we may anticipate that as time goes on the relationships estimated are less and less appropriate for the individuals who currently make up the population (including immigrants). If the former, then we may try to obtain current representativeness, treating unknown early data on immigrants as missing. The problem is that in general such earlier data values may have different distributions from the earlier data values of those present at the start of the study. This issue will be especially important if immigration status is one of the factors under study. In the light of this, thinking about specific comparison groups would seem to be a more useful focus than attempting to decide how to define population representativeness.

The argument about the lack of need for a representative sample has considerable strength. From an analytical (scientific) perspective what is required are statements that are generalisable to specific groups, including of course those people living within a given society or environment at any moment and who happen to constitute a 'real' population, such as is measured by a census. The distinction between scientifically driven data analysis and analysis directed at making estimates for real populations, however, is not always clear. For example, if interest is in prevalence differences between ethnic groups within age categories, there may be scientific interest in whether these are changing over time within the same geographically defined population, and whether any changes can be explained by other factors. In this case successive representative samples would be needed. What would be gained scientifically from such a comparison is information on potentially causal factors that mediate or explain the prevalence differences. The use of 'real' populations for this purpose in effect is to take advantage of 'naturally occurring' changes in such factors that may be happening over time. On the other hand it may be more efficient to choose a heterogeneous sample that allows the same exploration based on having sufficient variation for those factors. Thus, if we were interested in the relationship between pregnancy smoking and neonatal mortality, we would not generally wish to derive estimates for a real population where the structure of that population affected the size of the relationship or the power to detect any effects. Thus, for example, in a population with high average birth weight this relationship is known to be weak with a very large sample size needed to have reasonable power to detect it (see for example, Goldstein, 1977). Selecting a sample that does not represent a real population but has a high degree of heterogeneity in terms of birth weight, may provide much more power to investigate the hypotheses of interest.

We can illustrate this particular point from an analysis of early studies that looked at the relationship between maternal smoking and neonatal or perinatal mortality. Goldstein (1977) showed that, for different studies representing different populations of pregnant women, the difference in (or ratio of) mortality rates between smokers and non-smokers increased steadily as the average birth weight in the population decreased. Table 1 shows this for six different studies. The simplest explanation for the relationship is that smoking acts on mortality through an average 160g reduction in birth weight. The relationship between mortality and birth weight is nonlinear, with the relationship becoming steeper as birth weight decreases, and this implies that we will observe a greater difference for those populations with more low birth weight babies. In fact, for the two populations with the highest average birth weight, the difference is negligible.

Thus, if we had confined ourselves to the 'marginal' relationship between smoking and mortality, then our inferences would have differed according to the 'real' population studied. From a scientific perspective however, such inferences, especially in terms of a causal relationship, would be inadequate. It illustrates the point that, from a scientific perspective, the real population is of secondary importance: what we need is to understand those factors that could mediate the relationship of interest.

Table 1. Maternal smoking in pregnancy and neonatal/perinatal mortality

Population (1950-1970)	**% low birth weight (<2500g)**	**Mortality ratio: smokers/non-smokers**
US private health	3.2	1.03
Sweden	3.5	1.01
US naval wives	4.3	1.32
Ontario	4.5	1.27
UK	5.4	1.28
US general	5.9	1.40

A case where both specific population estimates are required and there is sufficient power to explore scientifically interesting hypotheses, is the British birth cohort known as 'Life Study' (Dezateux et al., 2013). This has a design that studies all 60,000 mothers over a period of time during pregnancy within relatively small but heterogeneous geographic clusters, treated effectively as a random sample from a superpopulation for those geographic strata, together with a UK random sample over the same time period, of some 20,000 live births, treated as a random sample from the superpopulation defined over the whole country. Both components of the study are followed up during the first year of life (and potentially beyond) with considerable overlap in terms of the information collected. The pregnancy component aims to collect genetic and other biological data not collected in the birth component. The advantage of such a design is that for population estimates using variables collected in the birth component there is additional information available from the numerically larger pregnancy component to improve the accuracy of these, for example using suitable weights that can be computed from nationally available birth data. For many scientific hypotheses the data available from the pregnancy component alone will often suffice, but power can also be increased by using the data from the birth component, within a combined analysis. Furthermore, informative selection, notably as a result of non-response, can be addressed by the existence of comprehensive population birth registry data against which the characteristics of those responding can be checked. This is in effect a special case of purposive sampling.

The ability to exploit such a design requires appropriate software tools that can 'borrow strength' across the two components. Providing such tools for routine data analysis is highly desirable, although it may be practically challenging. The point, however, is that it helps to understand the debate over whether a sample should be purposive or representative since in this case it can efficiently be both.

Conclusions

The idea that population studies, especially longitudinal ones, should strive to be representative of 'real' populations may not always be helpful. While, for certain purposes associated with enumeration and administrative policies, real population representativeness is required, from a scientific perspective this may well be unnecessary. Scientific inferences are concerned with uncovering relationships that can be tested across different contexts and that may eventually attain the status of causal explanations. To ensure validity researchers need to pay attention to selection factors that may lead to biased estimates, where 'bias' is defined in terms of a clearly defined *statistical* (super)population, and much of applied statistical methodology is devoted to this issue. To enhance the effectiveness of any analysis, heterogeneity is generally desirable, and this will often imply purposive sampling that is non-representative of any particular real population. In practice, as is the case with Life Study, an optimum design may well be one that combines such purposive sampling with population representativeness, so serving both enumeration and scientific aims.

Acknowledgements

My thanks for comments are extended to the following: Carol Dezateux, Francesco Sera and Rachel Knowles. The paper is based upon one delivered at the SLLS 2014 conference in Lausanne. This work was supported in part by the Economic and Social Research Council [Grant number ES/L002353/1]. Life Study is supported by the Economic and Social Research Council (ESRC), the Medical Research Council (MRC) and University College London (UCL) and is part of the Birth Cohort Facility Project, which receives funding from the UK Government's Large Facilities Capital Fund.

References

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vignoles, A., & Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. *The Lancet, 382,* S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Ebrahim, S. & Davey-Smith, G. (2013). Commentary: should we always be deliberately non-representative? *International Journal of Epidemiology, 42*, 1022-1026. http://dx.doi.org/10.1093/ije/dyt105

Elwood, J.M., (2013). Commentary: on representativeness. *International Journal of Epidemiology, 42,* 1014-1015. http://dx.doi.org/10.1093/ije/dyt101

Goldstein H. (1977). Smoking in Pregnancy: some notes on the Statistical Controversy. *British Journal of Preventive & Social Medicine 31* 13-17. http://dx.doi.org/10.1136/jech.31.1.13

Nohr, E.A., & Gleen, J. (2013). Commentary: Epidemiologists have debated representativeness for more than 40 years – has the time come to move on? *International Journal of Epidemiology, 42,* 1016-1017. http://dx.doi.org/10.1093/ije/dyt102

Richiardi, L., Pizzi, C.F. & Pearce, N. (2013). Commentary: representativeness is usually not necessary and often should be avoided. *International Journal of Epidemiology, 42,* 1018-1022. http://dx.doi.org/10.1093/ije/dyt103

Rothman, K.J., Gallacher, J.E.J., & Hatch, E.E. (2013a). Why representativeness should be avoided. *International Journal of Epidemiology, 42,* 1012-1014. http://dx.doi.org/10.1093/ije/dys223

Rothman, K.J., Gallacher, J.E.J., & Hatch, E.E. (2013b). When it comes to scientific inference, sometimes a cigar is just a cigar. *International Journal of Epidemiology, 42,* 1026-1028. http://dx.doi.org/10.1093/ije/dyt124

Commentary by

Peter Lynn University of Essex
plynn@essex.ac.uk

The need for representative survey samples

Introduction

In any field of scientific endeavour it is healthy to challenge orthodoxy. Standard practice should not be assumed to be best practice without question. Representative sampling is the orthodoxy in many applied fields of survey research and it is pleasing that this special section of *Longitudinal and Life Course Studies* is questioning when and why this should be the case. Let us be clear what this debate is *not* about. It is not about *how* to select a representative sample. There is a long history of debate on that subject, going back at least as far as the foundation of modern survey sampling theory with Kiaer (1897) and Neyman (1934), given prominence following the 1948 United States Presidential Election polling disaster (Mosteller, Hyman, McCarthy, Marks & Truman, 1949), and periodically revisited in various forms ever since. My thoughts on the role of non-probability sampling are recorded in Lynn (2005). That debate is again topical currently, particularly due to the rise of relatively cheap and fast online access panels in the social and political sciences (Bosnjak, Das & Lynn, 2015). However, the topic here is not *how* to select a representative sample but rather *when* and *why* it should be our objective to do so.

What should a sample represent?

Survey samples are rarely if ever of inherent interest. Rather, a sample is used to make broader inferences. Therefore, survey samples should be representative of something broader. But what? Goldstein's article touches upon this question by drawing distinctions between descriptive and analytical statistics and highlighting the role of confounding (or mediating) variables. I would suggest that if the analytical objective is to estimate the association between a particular set of variables, then the sample should be representative of that association. If the objective is to estimate a population distribution of some kind (be that univariate or multivariate) then the sample should be representative of that distribution. And so on. If the sample is not representative of the set of parameters to be estimated, whether those are causal, associative or descriptive, then we risk biased estimation, in the statistical sense outlined by Goldstein. It could therefore be argued that the representativeness objectives for a survey sample should depend on the analytical objectives[1].

To take an extreme example, suppose we want to estimate the association between two variables, when we already know (or assume) this association to be linear and already know (or assume) that there are no (important) confounding variables. If there are truly no confounding variables, the association should hold in any population, so it matters not whether our sample represents any particular population. In fact, we only need two non-identical observations in order to be able to perfectly estimate the bivariate association. This is obviously an unrealistic example for survey research (though it is exactly the type of estimation that takes place in school physics classes, for example), so it should be instructive to consider the ways in which it is unrealistic. First, it is ambitious to suppose that we know in advance the exact form of the association. Sampling just a few observations from each extreme of the distribution should be adequate to estimate a linear association, but if the true association has some curvature, this may be missed unless we have observations from throughout the distribution. Second, a complete absence of confounding variables is unlikely. Thus, to estimate the (conditional) association between our two variables of interest, we need also to identify (and obtain good measurements of) each confounding variable. One could argue, then, that a representative sample is not necessary provided that we can identify in advance all confounding variables of the relationship of interest, and measure them with our survey, and provided we ensure that the sample broadly covers the distribution of interest. However, this begs the question: which distribution? To be able to truly generalise our findings, we surely mean the distribution of values that could exist in any population to which we wish to claim that our results apply. Thus, we cannot completely get away from the notion of populations.

These criteria for being able to rely on a non-representative sample are quite demanding. It is hard to envisage a realistic social science research example where we can be confident of knowing in advance all possible confounding variables (let alone being able to measure them all well). When the causal mechanism of interest is, say, biological or chemical, one may be able to get closer to meeting these criteria - and that is a possible reason for epidemiologists to have a different take on this debate to social scientists - but the fundamental issues are the same.

Most social surveys – even those tightly focused on a single topic – have multiple analysis objectives. Large numbers of estimates of different kinds are typically required, making it unlikely that all confounding mechanisms are known for all analyses. In this situation, as pointed out by Goldstein, a population representative sample will at least provide a means of identifying the form of unexplained variation, testing in an exploratory way the association of this variation with other variables, and thereby moving towards the advancement of knowledge about hitherto unidentified causal factors. The primary purpose of some surveys – and secondary purpose of many – is to provide a data resource for research by secondary analysts. It is impossible for such research to have been specified prior to the original design of the survey and therefore to have influenced the survey design. In this situation, having a population representative sample can be thought of as a safety mechanism that ensures that the population distribution of the phenomena of interest is covered and also permits estimation of the extent and nature of unexplained variation. Of course, it remains up to the researcher to decide whether the particular population covered is suitably similar to, or representative of, the kind of population to which inferences should be made. I return to this issue below.

Which Population?

The ultimate objective of most survey-based research is to inform policy or practice of some kind. With this in mind, my earlier statement about wanting a sample to be representative of the parameters of interest can be re-cast. The parameters of interest are those in the population(s) that will be affected by policy or practice. Let's refer to this population as the *policy population*[2]. So, broadly, we want our survey sample to be representative of the policy population in terms of the parameters to be estimated. How can we be sure that this is the case? We can't. Not least because the policy population is always, by definition, a future population and we can never perfectly predict the future. But there are two things we *can* do:

a) try to minimise the risk that our parameters of interest differ greatly between the study population and the policy population, by defining the study population appropriately;
b) try to predict or model relevant ways in which the policy population may differ from the study population and incorporate this into our estimation.

Step a) is typically achieved by studying the most recent available equivalent of the relevant future population. Thus, in 2015 we may be able to analyse data from a representative sample of the 2014 population of Great Britain, for example, in order to infer the likely effects of a policy that might be implemented in 2016. Our assumption is that the 2016 population will be broadly similar to the 2014 one in terms of the relevant (causal) parameters. However, we do not expect the population structure to be identical: based on recent trends, we may expect some net ageing and some net immigration, for example, in which case we can implement step b) by projecting our estimated parameters onto the predicted 2016 population structure.

The example of the previous paragraph is an optimistic scenario, where the study population and policy population have a very large overlap, though even in this case the overlap may not be as large as it seems. Policies often remain in place for many years, and can have long-lasting impacts, so the true policy population perhaps consists of people resident in Britain at any time over the subsequent several years or decades. And often study and policy populations are even further disconnected. For example, if a good survey-based study has been carried out in one country, should researchers and policy-makers in another country assume that the findings will apply to their situation too? This is a common dilemma.

Funders must decide whether it is worth investing considerable resources to replicate a study carried out in a different context. They should be guided by the principles set out above. It is only worth funding the replication study if there is a sufficiently strong probability that the key parameters of interest are substantially different. Interpreting concepts such as "sufficiently strong probability" and "substantially different" will of course be subjective, but can be guided by knowledge of pertinent differences between the two populations and, particularly, by study findings regarding important confounders and unexplained variance.

Relevant policy populations can be very different for different types of research. Medical researchers may often hope that their findings could be generalisable to almost all current and future human populations (barring changes in the underlying etiology), whereas public bodies concerned with administering healthcare, education, housing, social support and so on are generally responsible for populations that are clearly defined by geography, usually at a national, regional, or local level. In the latter case, researchers may use survey samples that are representative of a recent equivalent of the same geographically-defined population or may resort to similarity-of-parameters arguments in using data from a different population (for example, arguing that national findings should apply in each region of the country).

Longitudinal Surveys

The arguments that I have presented so far are rather general and should apply to any sample-based scientific endeavour. However, longitudinal studies in the social sciences have at least three additional distinct characteristics that should influence the answer to the question posed in the title of Goldstein's paper:

a) Longitudinal estimates by definition refer to longitudinal populations;
b) The time interval between data collection and policy impact can be particularly great;
c) During the course of the study, new research agendas can emerge that were not envisaged when the study was initially designed.

I discuss here each of these three points in turn.

Any human population ('real' population, in Goldstein's terms) is dynamic; people will join or leave the population over time. Analysts of cross-sectional surveys tend to ignore this uncomfortable fact and instead claim that their estimates relate to a well-defined population that existed at a moment in time. This may be a reasonable approximation to reality for many purposes, but the longer the period of time over which elements were sampled or data collected, the less accurate the approximation will be.

Longitudinal surveys cannot duck this issue. An estimate of, say, the relationship between a treatment or baseline measurement and an outcome ten years later can only be based on a sample of people who were in the 'real' population at both points in time. People who entered the 'real' population subsequent to the baseline measurement (e.g. through birth, migration or status change) or who left the 'real' population prior to the outcome measurement cannot contribute to the estimate. The study population can therefore be defined as persons who were members of the 'real' population at both time points. Longitudinal parameters are properties of longitudinal populations (Smith, Lynn & Elliot, 2009), whether the population is 'real' or a conceptual superpopulation. The distinction between cross-sectional and longitudinal representativeness is important (Lynn, 2011).

Research based on long-term longitudinal studies is incredibly powerful for understanding dynamics and causality over long periods. The down side of this is that some of the data underpinning the research will be rather old. A study of the influence of infant feeding practices on, say, educational and employment outcomes by age 30 must rely on feeding practice data that is at least 30 years old. The study population and policy population are therefore separated not by just a couple of years, as in the example of the previous section, but by four decades or more. This makes it harder for the researcher to be confident that key population parameters will remain unchanged: in a rapidly-changing world, not only may feeding practices themselves have changed, but so might the many mediators of their impacts on early-adulthood outcomes.

Research agendas certainly evolve over time, due to new knowledge, new technology, new social problems, and so on. When the sample design for the National Child Development Study (NCDS) was established, in the 1950s, it would have been impossible to envisage the myriad purposes for which researchers would be using the data half a century later. For this reason, the role of population representative sampling in ensuring the sample will contain as much heterogeneity as exists in the population is particularly important. The heterogeneity will be present for any research objective, not just those that were identified when the study was conceptualised.

Conclusion

The omission of the word ‘population’ from the title of this piece is deliberate: survey samples certainly need to be representative, but not necessarily of a conventionally-defined population. To meet scientific objectives, samples should represent the estimation parameters of interest. How this is best achieved will depend largely on how much is known about these parameters prior to the study. When little is known, and particularly when some research objectives cannot be well specified in advance, population representative sampling provides a mechanism for ensuring representation of extant variance. For multi-purpose surveys, population representative sampling is likely to represent an efficient compromise between the diverse optimal sample distributions for different analytical purposes. The sample should represent a population that is as similar as possible to the future policy population(s) that may be affected by study findings. A good choice may be a recent equivalently-defined population, especially when this maximises overlap between the study population and the policy population.

Longitudinal studies are typically characterised by the features that point towards population representative sampling as an appropriate strategy (limited advance knowledge about estimation parameters, inability to specify all estimation requirements in advance, large time interval between data collection and policy implementation).

References

Bosnjak, M., Das, M. & Lynn, P. (2015). Methods for probability-based online and mixed-mode panels: Selected recent trends and future perspectives. *Social Science Computer Review*. Published online 7 April 2015. http://dx.doi.org/10.1177/0894439315579246

Kiaer, A.N (1897). *The Representative Method of Statistical Surveys*. Translation 1976, Norwegian Central Bureau of Statistics, Oslo.

Kruskal, W. & Mosteller, F. (1979). Representative sampling III: the current statistical literature. *International Statistical Review 47*(3), 245-265. http://dx.doi.org/10.2307/1402647

Lynn, P. (2005). Inferential potential of non-probability samples: discussion. *Bulletin of the International Statistical Institute*, Proceedings of the 55th Session. Sydney: International Statistical Institute.

Lynn, P. (2011). Maintaining cross-sectional representativeness in a longitudinal general population survey. *Understanding Society Working Paper* 2011-04, Colchester: University of Essex. https://www.iser.essex.ac.uk/research/publications/working-papers/understanding-society/2011-04

Mosteller, F., Hyman, H., McCarthy, P.J., Marks, E.S. and Truman, D.B. (1949). *The Pre-election Polls of 1948*. New York: Social Science Research Council.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society 97*, 558-606. http://dx.doi.org/10.2307/2342192

Smith, P., Lynn, P. & Elliot, D. (2009). Sample design for longitudinal surveys. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*, 21-33. Chichester: Wiley. http://dx.doi.org/10.1002/9780470743874.ch2

Endnotes

[1] Kruskal and Mosteller (1979) distinguish estimation bias from selection bias. Goldstein notes that unbiased estimators can be constructed from biased samples, provided the biasing selection mechanism is known, as with the case of disproportionate stratified probability sampling. In this brief note I shall fudge this issue: my use of the term population representative sample includes – but is not necessarily limited to – any probability-based sample that covers the whole population.

[2] I deliberately avoid the term *target population*, as this is usually used in a more restrictive sense. However, under an explicit superpopulation model the two concepts converge.

Commentary by

Graciela Muniz-Terrera University of Edinburgh, UK
G.Muniz@ed.ac.uk
Rebecca Hardy University College London, UK

Some thoughts about representativeness

The paper by Goldstein makes an important additional contribution to the ongoing debate about whether and when analytic samples need to be population representative in studies in epidemiology and social and medical research.

The paper outlines the arguments presented by Rothman, Gallacher and Hatch (2013) and the stimulating accompanying commentaries that initiated the recent discussion on the topic. The need to distinguish between a "real" population and a population defined as a statistical concept that refers to any well-defined collection of units, but that may not reflect any actual population is also discussed. Additionally, Goldstein recalls the definition of bias as the difference between estimates obtained from any particular sample and the unknown true parameter of the population under study, emphasising that this population only has to be a statistically defined population and not a "real" population. In this paper, we comment on a number of points which have particular relevance for birth cohort and longitudinal studies.

The discussion of the temporal aspect of the concept of representativeness is, of course, important. Goldstein points out that representativeness is not a static concept that is preserved indefinitely over time, but rather, is a concept affected by the passing of time. Even when all efforts are made to select a representative sample of a given population at the outset of a study, the representativeness of this initial sample is unlikely to be preserved over time as the sample is followed up longitudinally. The real population of which the sample was initially representative will inevitably evolve, while at the same time loss to follow up will alter the characteristics of the study sample. Goldstein cites the example of the 'Life Study', the newest of the British birth cohort studies, where a complex sampling strategy and the use of weighting allows both the estimation of population parameters with adequate accuracy and the investigation of scientific hypotheses in a group with more extensive biological data. Let us now consider the oldest of the British birth cohort studies, the MRC National Survey of Health and Development (NSHD) (Wadsworth, Kuh, Richards & Hardy, 2006). The NSHD followed up a sample of all single births to married women in England, Scotland and Wales which took place in one week in March 1946. This initial sample included all babies born to women with husbands in non-manual and agricultural employment and one in four births to women with husbands in manual employment. This sampling scheme was chosen to keep the national distribution and to achieve a similar proportion of children in each social group (Wadsworth, 1991). Weights have thus been used when calculating prevalence estimates in order to allow for this original sampling. In 2015, the cohort is now aged 69 and the 24th data collection on the whole sample is taking place. Of course, the NSHD sample are no longer representative of the population of individuals aged 69 years old now living in England, Scotland and Wales. Demographic changes have occurred, with both immigration and emigration taking place over the lifetime of the cohort. Hence, any prevalence estimates can only ever be representative of the British-born population of 69 year olds. Furthermore, the diverse origins of immigrants joining the British population will mean that they have been exposed to different early life conditions compared with the British born population. Such differences in early life experience are likely to impact on adult health and mortality patterns and could thus affect estimates of association between early life risk and adult outcomes.

This raises the question of whether national cohort studies should adopt the practice of supplementing the samples to try and maintain study representativeness. Such supplementation was not attempted in the NSHD. In contrast, in the 1958 British Birth Cohort (National Child Development Study) and the 1970 British Birth

Cohort, during childhood, as cohort members could be traced through schools, immigrants born in the reference week were added to the samples. This was no longer possible once cohort members became adults (Power & Elliott, 2006, Elliott & Shepherd, 2006). We appreciate the value of such attempts to retain representativeness, but also see challenges in this practice if the distribution of the subgroups that comprise the original population is also dynamic and vary significantly over time. The innovative design of the 'Life study' (Dezateux et al., 2013) means that the initial sample is both "purposive and representative" and it will be informative to see how appropriate software tools for routine and complex data analysis can be provided. It will also be interesting to see whether representativeness can be maintained as the sample is followed up longitudinally, as loss to follow up and continuous demographic changes to the population occur. Given the richness of the data available in cohort studies and their ability to address unique scientific hypotheses about long term associations, we need to consider whether attempting to retain representatives by sample supplementation or by statistical weighting for investigations of prevalence is the best use of such studies.

In the original exchange between Rothman and others, Elwood (2013) elaborated the concept that any real population is a historical entity and that by the time inferences about the population are available, the initial population may have changed in important ways. We now reflect on how period effects can affect inferences made using historical data. As an example, let us consider the association of smoking and cognitive function in school pupils aged 15. Assume we have data for two samples of children that were representative of the school population aged 15 at the time of data collection, such that one sample comprised of students aged 15 years old in 1982 and the other of students aged 15 in 2013. Smoking prevalence in these two samples born 30 years apart will vary greatly. In 1982, 24 % of pupils aged 15 smoked, a percentage that has been decreasing steadily over time so that by 2013 only 8 % of pupils smoked (www.ash.org.uk) as a consequence of heightened awareness of its negative effects on health and various changes in laws, public health and commercial policies. A lack of power to detect an effect of smoking on cognitive function could therefore result as the prevalence of the risk factor declines. So, even when both samples were chosen to be representative of the population of pupils aged 15, because of a period effect, different conclusions about the association of interest could be drawn. If the researcher is interested in the potential causal association between smoking and cognition, then selecting a population with a higher prevalence of smoking is more important than picking one which is representative. On the other hand a risk factor might become more prevalent over time and thus associations may not be picked up in historical cohorts. For example, the prevalence of childhood obesity was considerably lower in the NSHD compared with cohorts born in the 1990s and later (Johnson, Li, Kuh & Hardy, 2015). It is therefore unclear whether the generally null associations between body mass index (BMI) in early childhood and coronary heart disease (CHD) observed in historical cohorts (Owen et al., 2009) are due to a lack of power. Such historical differences need to be considered and discussed when, for example, synthesizing results in systematic reviews and when implementing evidence based public health policies.

Finally, an interesting argument presented by Goldstein and discussed in the original exchange between Rothman and other commentators is about the value of non-representative samples in the context of replication and generalisation of results across different populations. The importance of a thorough understanding of all the potential sources of heterogeneity across studies, including the representativeness, or not, of samples, and the period effects, as well as differences in data collection methods and analytic methods when evaluating the reproducibility of results is vital. These points are of particular relevance in the heated debate about reproducibility and replicability of results that has entertained the attention of researchers across various scientific areas (Francis, 2012; Ioannidis, Nosek & Iorns, 2012; McNutt, 2014; Mulkay & Gilbert, 1986), particularly when reproducibility is defined as the conceptual replication of experiments as conceived by Drummond (2009). Despite unfortunate publishing practices that discourage publication of reports that aim at testing reproducible research and result in publication biases (Francis, 2012), the concept of reproducible research has, historically, been at the core of scientific discovery.

From that perspective, the need to generate strong evidence about patterns of associations is at the core of the multi-study work fostered by the Integrative Analysis of longitudinal Studies of Ageing network, a network of longitudinal studies of ageing (www.ialsa.org). Researchers affiliated to the IALSA network independently analyse data from multiple studies employing a coordinated approach that involves the consistent use of the same analytical method (identical analytical model where possible and consistent coding of harmonized variables where possible). This coordinated analytical approach maximises the ability to fairly compare results and enables the examination of consistency of patterns and of associations across samples that may differ in a variety of ways, including differences by geographical location, sample composition and representativeness (Piccinin). The use of the same analytical approach reduces the potential sources of heterogeneity across studies that may emerge from the use of different statistical methodologies to answer similar questions. Consistent results generated from diverse samples are reassuring and provide stronger evidence in support of the hypothesis tested. On the other hand, inconsistent results require a thoughtful evaluation of potential reasons that may explain the divergence of results, including differences that may emerge from features of the data (including representativeness), and sample composition and sampling procedures. For example, in an investigation of the association of the effect of education, age and sex on global cognitive function measured using the Mini Mental State Exam in six international longitudinal studies of ageing, Piccinin and colleagues (2012) found that education was positively associated with performance across all six studies, but was only associated with rate of decline in the cohort containing the oldest participants. In five of the six studies, estimates of rate of decline were also found to be similar, but in the cohort of oldest individuals, individuals were found to decline at a much faster rate than in the other samples. The authors report that an investigation of the sample composition and a better examination of the sampling procedure followed in this outlying study helped them understand that dementia cases had been handled differently in the study compared to the other studies. Indeed, in this study efforts had been made to keep individuals who developed dementia in the study, whereas in all the other studies individuals with dementia were not included in the follow up samples. When individuals with dementia were removed from the sample, the estimated rate of decline aligned to the rate of decline estimated in the other five studies.

The general discussion about representativeness and Goldstein's contribution with particular relevance to longitudinal studies and their historical context is very valuable. This discussion is helpful in raising awareness among researchers to think more about when representativeness is a problem, but also to appreciate when to value a lack of representativeness.

References

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vignoles, A., & Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. *The Lancet, 382,* S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Drummond, C. (2009). Replicability is not Reproducibility : Nor is it Good Science. Retrieved from: www.site.uottawa.ca/ICML09WS/papers/w2.pdf

Elliott, J. & Shepherd, P. (2006) Cohort profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology, 35*(4), 836–843. http://dx.doi.org/10.1093/ije/dyl174

Elwood, J.M. (2013). Commentary: On representativeness. *International Journal of Epidemiology, 42*(4), 1014–1015. http://dx.doi.org/10.1093/ije/dyt101

Francis, G. (2012) Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review 19*(6), 975-991. http://dx.doi.org/10.3758/s13423-012-0322-y

Ioannidis, J.P.A., Nosek, B. & Iorns, E. (2012). Reproducibility concerns. *Nature Medicine 18*(12),1736–7. http://dx.doi.org/10.1038/nm.3020

Johnson, W., Li, L., Kuh, D. & Hardy, R. (2015). How Has the Age-Related Process of Overweight or Obesity Development Changed over Time? Co-ordinated Analyses of Individual Participant Data from Five United Kingdom Birth Cohorts. *PLOS Medicine*, *12*(5), e1001828. http://dx.doi.org/10.1371/journal.pmed.1001828

McNutt M. (2014). Reproducibility. *Science*. *343*(6168),229. http://dx.doi.org/10.1126/science.1250475

Mulkay, M.& Gilbert, G.N. (1986) Replication and Mere Replication. *Philosophy of the Social Sciences 16*(1), 21–37. http://dx.doi.org/10.1177/004839318601600102

Owen, C.G., Whincup, P.H., Orfei, L., Chou, Q.A., Rudnicka, A.R., Walthern, A.K. … Cook, D.G. (2009). Is body mass index before middle age related to coronary heart disease risk in later life? Evidence from observational studies. *International Journal of Obesity, 33*(8), 866–877. http://dx.doi.org/10.1038/ijo.2009.102

Piccinin, A.M., Muniz-Terrera, G., Clouston, S., Reynolds, C.A., Thorvaldsson, V. Deary, I.J. …Hofer, S.M. (2012).Coordinated Analysis of Age, Sex, and Education Effects on Change in MMSE Scores. *The Journal of Gerontologly Series B: Psychological Sciences and Social Sciences*, *68*(3), 374-390. http://dx.doi.org/10.1093/geronb/gbs077

Power, C. & Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology, 35*, 34–41. http://dx.doi.org/10.1093/ije/dyi183

Rothman K.J., Gallacher, J.E.J. & Hatch, E.E. (2013). Why representativeness should be avoided. *International Journal of Epidemiology 42*(4), 1012-1014. http://dx.doi.org/10.1093/ije/dys223

Wadsworth, M. Kuh, D., Richards, M. & Hardy, R. (2006). Cohort profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology 35*(1), 49-54. http://dx.doi.org/10.1093/ije/dyi201

Wadsworth, M.E.J. (1991). *The imprint of time: Childhood, History and Adult Life.* Oxford University Press.

Commentary by

Colm O'Muircheartaigh University of Chicago, US
caomuirc@uchicago.edu

Why we need population representative samples

Goldstein questions the need for observational studies to achieve representativeness for well-defined populations, in particular for longitudinal studies. While he recognises the distinction between the notions of representativeness and proportionality, he fails to acknowledge the importance of distinguishing between samples of convenience and targeted samples from special subpopulations. In this note I emphasise the critical significance of probability sampling, in contrast to purposive sampling, and draw special attention to the artificial distinction between descriptive and analytical statistics. Goldstein (correctly) draws attention to the confusion between disproportional sampling and non-representative sampling but fails to recognise the inferential implications of choosing between probability samples and nonprobability samples. A probability sample is in essence a sample in which every element of the population has a (known) non-zero probability of selection; the definition of the population may be such that it does not correspond to a real population. The structure of a probability sample from a (general) population may exclude some domains from the target population and may be modified by design in order to produce appropriate numbers of cases for particular comparisons of subsamples of that target population. Probability samples have particular strength in making inferences, whether for scientific or for policy purposes.

Defining populations

All inference is, by definition, to a population beyond the sample on which the inference is based. Much of the argument in Goldstein, and in the papers he references, has to do with the definition of this inferential population. I concur that the population must be clearly defined; I accept also that it may not correspond to a "real" population at a point in time. However, unless it can be defined in such a way that a sample may be selected from it, there will be no scientific foundation for inferences to it without untestable assumptions about freedom from bias.

Consider first the case where the purpose is to represent a national population; as an example, consider the selection of a sample for the United States (US) National Children's Study (NCS)(Michael & O'Muircheartaigh, 2008). In designing a nationally representative sample for this study, the purpose is not to address every subpopulation of interest in the US. The purpose is to insure that every element in the population has a non-zero probability of being selected into the sample. This is achieved by identifying a survey population that is defined to be as close to the target population as feasible, such that it reflects both measurable and unmeasurable characteristics of that population

Suppose that we are interested in the relationship between an environmental exposure *X* and a health outcome *Y*, which can be modeled (for simplicity) as the linear function $Y=a+bX+e$. If all people in the population have the same *b*, then the nature of the sample does not matter because as long as *X* is accurately measured we will have only random measurement error in *Y*. However, if there are confounding factors *Z*, which affect *Y* and are related to *X*, then our estimate of *b* may be biased unless the elements of *Z* are controlled. If *Z* is known, then model-based estimates of the relationship between *X* and *Y* can be obtained that control for *Z* and yield an unbiased estimate of *b*, again regardless of the sampling design. However, there may also be moderator variables *W*, which interact with *X* in influencing *Y*. Here, different individuals will have different values of *b* depending on the elements in *W*. If *W* is known, then we can include interactions in the model and the separate estimates of *b* will also be unbiased.

Unfortunately in practice *W* and *Z* are at least to some extent unknown and in the case of longitudinal studies like the NCS are likely to evolve over time. Some elements of *Z* and *W* may be known but are unmeasurable and others may simply be unknown at the time. Here, the best that we can do is to provide an average effect *b*. To do so, however, requires that we create a sample that fully reflects the population of interest, a probability sample drawn from the population so that our estimate of *b* is an unbiased estimate of the average effect in the population or in a defined subgroup. The probability sample guarantees that we will (in expectation) cover the range of confounding variables proportionately.

It is also possible that the interest is not in the average effect but in the effect on specific subgroups of this general population (as in the birth weight example below). Thus, in the NCS we might wish to focus on particular ethnic groups or on the comparison of these groups. In this case to maximize power for the comparison we would take equal numbers of cases from the groups of interest, rather than numbers proportional to their distribution in the general population. These subsamples would however be chosen to be representative of the groups of interest; their representativeness would be warranted by the fact that they were probability samples from their respective groups. Only the relative sizes of the subsamples would deviate from the parent population, not the intrinsic nature of the sampling process.

Goldstein's example of the relationship between pregnancy smoking and neonatal mortality provides a further illustration of this principle. The six studies he cites (from an analysis by Goldstein (1977)) demonstrate a non-linear relationship between mortality and birth weight, with a negligible effect for the two populations with the highest average birth weight, and an increasingly steep relationship as the population average birth weight decreases.

Goldstein argues that this example demonstrates the secondary importance of the population. To the contrary, the data demonstrate the opposite. Had the range of birth weights across the US been included in a single US study, the analysts might have been more likely to observe the non-linearity in the relationship; this indicates the importance of covering the full range of variation of X, W, and Z in a population rather than accepting the subpopulation that is most convenient. One might indeed argue that there was a failure of both the theoretical basis and the analysis of the studies in not examining the data for possible interactions with birth weig

ht in the model,

At no point in his disquisition does Goldstein suggest that the samples in any of the studies he cites should be "non-representative". The implicit understanding is that the sample in each is in fact representative of the population from which it is drawn. Were it not, neither the partial generalisation within the study would be justified, nor would its incorporation into Goldstein's 1977 meta-analysis.

Two-phase sampling

The case of the British birth cohort known as the 'Life Study' is also subject to an alternative interpretation from that offered in Goldstein. A geographically clustered sample of 60,000 mothers is selected from a set of relatively small but geographically heterogeneous clusters; the 60,000 mothers are assumed to constitute a random sample from a set of geographic strata; there is a parallel (random) UK sample of 20,000 live births. The two samples can be used together to "borrow strength" from each other for different analyses. Comprehensive national (population) data from birth registries can be used to correct for differential nonresponse.

This combining of samples with different characteristics and different intensity of measurement is well recognised as a powerful design. The classic two-phase sampling design (Neyman, 1938) proposes just this combination of general representation and subsample focus; Neyman visualizes both samples as probability samples. Goldstein proposes this as a special case of purposive sampling, though it is not clear what his argument is. Presumably he does not argue that selecting the geographical areas purposively is superior to a design in which the areas were selected on a probability basis from a properly constructed frame of geographical areas. If indeed the selected areas were for some reason the only areas available, then suspicion must attach to them as being unrepresentative even of areas with ostensibly equivalent characteristics.

The extent to which the combined sample can be justifiably used to make inferences to the whole population depends critically on either (i) both samples being probability samples, or (ii) model-based assumptions that allow generalisation from the purposive component to the whole.

Additional benefits of representation through probability sampling

A platform for scientific discovery

Hypotheses about new exposures and gene-by-environment moderation will arise over the next 20 years, and a probability sample provides the best insurance that the study will provide useful numbers of children with variation in those environments and exposures of interest. The probability design also increases the prospects for serendipity by maximizing the spread of W and Z in the sample.

Maximization of scientific acceptability of data and of discoveries across disciplines

While many disciplines do not require probability samples for their inferences, no discipline considers a probability sample to be inferior to an alternative. Thus data based on a probability sample maximize the potential for cross-disciplinary collaboration and publication.

Public and political/policy acceptance

Resource allocation and acceptability of discoveries will be greater if the data are based on a scientifically warranted representative sample of the population.

Full variation in risks and exposures

A probability sample will produce generalisable risk estimates and the capability to estimate policy/intervention benefits from associations discovered and reported from the study.

Conclusion

Investigations of all kinds can make a contribution to science, and samples that are not representative have a place in scientific research, especially at early stages of exploration. I contend however that the superficial message of Goldstein's excellent article is wrong. Ceteris paribus, for both science and policy a probability sample is superior to a non-probability sample, representation trumps convenience, and the best way to obtain representation of the population of interest is through probability methods.

References

Michael, R. & O'Muircheartaigh C. (2008). Design Strategies and Disciplinary Perspectives: the Case of the US National Children's Study. *Journal of the Royal Statistical Society, Series A*, *171*(2) 465-480. http://dx.doi.org/10.1111/j.1467-985X.2007.00526.x

Neyman, J. (1938). Contributions to the theory of sampling human populations, *Journal of the American Statistical Association, 33*, 101-116. http://dx.doi.org/10.1080/01621459.1938.10503378

Commentary by

Chris Skinner London School of Economics, UK
c.j.skinner@lse.ac.uk

Discussion of 'When and why do we need population representative samples?'

There is much wisdom in this paper by Harvey Goldstein which builds on discussion in a set of papers in the *International Journal of Epidemiology* (IJE), and applies the ideas developed to a new British birth cohort study, the Life Study. I shall focus on his main theme, which rejects the need for representative samples, and on his concluding remarks relating to the Life Study. My comments come particularly from a survey statistics perspective.

I was reminded in looking at the papers in IJE of the observation by Kruskal and Mosteller (1979) (and in their three related articles) that the term 'representative sample' has multiple uses and "because of its ambiguities and imprecision", they "recommend great caution" in the use of this term and "usually a more specific expression will add clarity" (p.13). I shall seek to make greater use of the expressions 'population', 'sample' and 'bias' in my discussion.

As I understand Goldstein's main concern about representative sampling, it is that, for scientific purposes, making inference about 'real' populations is of secondary importance. This is a position which I should like to question. The survey statistics literature does make a distinction between descriptive/enumerative and analytic/scientific uses of surveys/studies. Estimation for a single study population is a common primary objective for the former. For the latter, the focus of Goldstein's paper, I think the notion of population will invariably need further refinement, but I think it can still serve a useful purpose to specify collections of units underlying targets for inference. I do not feel the need to downplay the notion of 'real' population.

Perhaps the simplest definition of populations of interest for scientific purposes is where there are two subpopulations to compare. I conceive of these subpopulations as 'real populations' in Goldstein's terminology. Suppose, for example, we wish to undertake a comparison of an outcome Y, according to values of X, given confounding factors Z (say infant mortality by maternal smoking given birth weight in Goldstein's example). For such conditional analysis, it would be natural to define specific subpopulations by X and Z, between which comparisons are to be made. Thus, in the example, one might choose to compare a low birth weight subpopulation and a normal birth weight subpopulation. Such comparisons have many vital roles in scientific research, as Goldstein notes. They may help to elicit and test causal hypotheses, perhaps through control of confounding factors. They may be valuable in assessing the replicability of findings across populations or to learn about interactions.

Given the specification of such subpopulations, it will often make sense to sample these subpopulations with different sampling fractions. For example, as discussed by Goldstein, the power to investigate the analytic objectives may be improved by sampling the low birth weight subpopulation with a higher relative sampling fraction. But I do not see this observation as any reason why the subpopulations (as real populations) are of 'secondary importance'. Their definition seems fundamental. I also do not see any reason why an analysis embracing a comparison of such subpopulations need be weighted to the population of all births (Skinner, 2005, p.84), let alone any need for the analysis to be confined to the 'marginal' relationship between smoking and mortality.

The simple comparison of subpopulations needs extension in various ways. With a continuous variable like birth weight, the definition of subpopulations via cut-points is arbitrary and we may imagine intervals of values of decreasing width and decreasing population counts. In this context, the notion of superpopulation which Goldstein mentions is useful and enables, for example, a regression relationship with continuous covariates to be specified in usual model terms. The longitudinal setting also introduces complexities, as Goldstein notes. A population like a labour force becomes dynamic with people entering and leaving the labour force over time. Even more complexity arises with, for example, households with the structure of the unit changing over time. In such cases, the term ‘population’ may seem stretched, but I think it is still reasonable to think in terms of what Goldstein calls a ‘well-defined collection of units’. Causal questions cannot be assessed from data on a single case but rather require reference to a set of units. As Holland (1986, p. 947) writes, “the important point is that the statistical solution [to the fundamental problem of causal inference] replaces the impossible-to-observe causal effect of t on a specific unit with the possible-to-estimate average causal effect of t over a population of units”. In my view the relevant populations do define ‘real’ notions of primary not secondary importance, given the need to report scientific findings transparently in terms of the kinds of people or other units to which they apply.

I now turn to the role of sample selection. I have already noted, in agreement with Goldstein’s discussion, that it may often be sensible to allocate the sample differentially according to variables of scientific interest (X and Z above) with a view to improving sampling efficiency (i.e. reducing variance). Consider next the question of bias, as arising from differences between the characteristics of sample units and those in the population (as conceived of in the previous three paragraphs). I have in mind bias arising from purposive and other forms of non-probability sampling, for example the volunteer effects described by Ebrahim and Davey-Smith (2013). Such bias is of major concern to survey methodologists today, with the relentless push to adopt non-probability samples, such as in internet panels, for cost and other non-scientific reasons.

In summary, I do think that in the analysis of longitudinal studies it is desirable to specify collections of units as populations, with a clear scientific rationale, and that the potential biasing effects of sample selection are of primary concern.

My final comments will elaborate on these points in the context of the Life Study. Here the basic study populations from which samples are drawn (leaving aside timing aspects) are (a) k populations of pregnant mothers (and partners) associated with k maternity units and (b) the population of all live births in the UK. I am unclear about the value of k (perhaps it remains to be determined) but suppose that it is small (under 10?). Sampling in (a) is by census and in (b) by a standard probability scheme and so, for the purpose of current discussion and leaving aside non-response considerations, I think we can disregard issues of representative sampling **within** these populations.

In the context of the earlier discussion, the key issue relates to the purposive selection of the maternity units. Following Goldstein’s discussion, it seems natural to ask what is the scientific rationale for the choice of maternity units? From Goldstein’s paper, the rationale seems to be geographic heterogeneity, perhaps associated with differences in distributions of what I have called X and Z variables relevant to the study. This raises the question of how differences in findings between different maternity units are to be interpreted? If, for example mortality ratios vary between units as in table 1 and there is also significant variation between units in a large number of other maternal health and socioeconomic factors, how will the finding be scientifically informative if k is small? Moreover, for some kinds of analyses, interpretation may even be complicated by confounding between the effect of the maternity unit and the nature of the maternal population.

In any case, if the results of analyses of data from a given maternity unit are only to be reported as relating to that population then issues of external generalisability are avoided and I have no concerns about sample selection bias. There do not then seem to be any differences in questions of representativity/generalisability compared to other geographically specific studies, such as the Southampton Women’s Survey (Inskip et al., 2006). The fact that scientific studies have some spatial and temporal specificity seems inevitable.

The more difficult questions relate to how the data will be combined across populations. The statistical methodology for standard comparisons would seem straightforward. Thus, in a regression setting, one may construct a categorical covariate representing both the k maternity populations and the general ‘birth population’, the latter possibly broken down by region or in some other geographical way. I am still unclear how to interpret the coefficients of this covariate and associated interaction terms, but this is just the comparative question I have already asked above.

Much less straightforward seems to me the question of how far it will be possible to increase “the precision of estimates for nationally representative measures” (Dezateux et al., 2013) using the maternity unit data, that is how to use data from a restricted and purposefully selected set of geographical clusters to make inference about the wider UK population? This 'borrowing of strength' across (a) and (b) is intended to provide, as Goldstein refers to it, an optimum design combining purposive sampling with population representativeness.

A review of non-probability sampling was conducted recently by the American Association of Public Opinion Research, with a summary report and discussion appearing in Baker et al. (2013). The combination of a national probability sample with a small number of geographically clustered 100% samples does not appear to be a standard approach. Baker et al. (2003) do provide some discussion of weighting and note that "the main concern with model-based inferences from non-probability samples is that population estimates are highly dependent on model assumptions" (p.97). A combination of a large non-probability sample (161,000+ web respondents) with a smaller 'nationally representative' quota sample (10,000+ respondents) was used in the Great British Class survey (Savage et al., 2013). Savage et al. (2014) recognised that their design is 'unorthodox', in response to criticisms e.g. by Mills (2014), and emphasised that their work should be seen as part of an 'experiment'. This survey is very different from the Life Study but I mention it just to illustrate that such 'combined' designs seem to me still novel and the extent to which reliable and efficient national estimates can be produced by combining the separate data sources seems to me a topic still in need of further study.

'Borrowing strength' is referred to in the small area estimation literature (e.g. Ghosh & Rao, 1994), but in that context borrowing across geographical units comes from fitting a model across a sufficient number of such units for a reasonable model to be fitted and for valid confidence intervals, taking account of geographic heterogeneity, to be constructed. It is not clear to me that k will be large enough for such an approach to be adopted.

Goldstein suggests a weighted approach will be used. One approach would be to weight inversely by the probability of selection, with weights of one attached to members of the maternity unit sample (since 100% are sampled). However, I would assume this would only increase the effective sample size of the birth sample by a small fraction and that this is not what is conceived. The idea may instead be to construct weighting classes using population registry data (but not geography) and then to make the modelling assumption that observations are exchangeable between the maternity unit and birth populations within weighting classes. Such a modelling assumption will depend upon the relevant analysis and the availability of auxiliary information but, in general, it would seem to me heroic. The assumption should, at least, be testable, although its testing would seem to be similar to testing the hypothesis of no maternity unit effect in the kind of regression analysis I noted above, where weighting variables are included as covariates. In summary, the proposed combined design seems to me to be novel (although perhaps I am unaware of similar designs) and I think there are several methodological questions regarding data combination to explore, even before one gets to the question of software tools referred to by Goldstein.

I am grateful to the Editor for the opportunity to discuss this interesting paper.

References

Baker, R., Brick, M.J., Bates, N., Battaglia, M., Couper, M.P. & Dever, J.A. (2013) Summary report of the AAPOR Task Force on non-probability sampling (with discussion). *Journal of Survey Statistics and Methodology*, *1*, 90-143. http://dx.doi.org/10.1093/jssam/smt008

Dezateux, C., Brockelhurst, P., Burgess, S., Burton, P., Carey, A., Colson, D., Dibben, C., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Knowles, R., Leon, D., Lyons, G., Reay, D., Vognoles, A., Walton, S. (2013). Life Study: a UK-wide birth cohort study of environment, development, health, and wellbeing. (2013). The Lancet, 382, Pp S31. http://dx.doi.org/10.1016/S0140-6736(13)62456-3

Ebrahim, S. & Davey-Smith, G. (2013). Commentary: should we always be deliberately non-representative? *Intl. J.* Epidemiol., 42, 1022-26. http://dx.doi.org/10.1093/ije/dyt105

Ghosh, M. & Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science, 9,* 55–93. http://dx.doi.org/10.1214/ss/1177010647

Holland, P.W. (1986) Statistics and Causal Inference, *Journal of the American Statistical Association*, *81*, 945-960. http://dx.doi.org/10.1080/01621459.1986.10478354

Inskip, H.M., Godfrey, K.M., Robinson, S.M., Law, C.M, Barker & Cooper, C. (2006) Cohort profile: the Southampton Women's Survey. *International Journal of Epidemiology*, *35*, 42-48. http://dx.doi.org/10.1093/ije/dyi202

Kruskal, W. & Mosteller, F. (1979) Representative sampling, I: Non-scientific literature, *International Statistical Review*, *47*, 13-24. http://dx.doi.org/10.2307/1403202

Mills, C. (2014) The Great British Class Fiasco: A Comment on Savage et al. *Sociology*, *48*, 437-44. http://dx.doi.org/10.1177/0038038513519880

Savage M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J. … & Miles, A. (2013) A new model of social class? Findings from the BBC's Great British Class Survey experiment. *Sociology, 47*, 219–50. http://dx.doi.org/10.1177/0038038513481128

Savage M., Devine, F., Cunningham, N., Friedman, S., Laurison, D., Miles, A. … & Taylor, M. (2014) On Social Class, Anno 2014. *Sociology*, *48* (forthcoming). http://dx.doi.org/10.1177/0038038514536635

Skinner, C.J. (2005) Introduction to Part B. In R.L. Chambers and C.J. Skinner (Eds), *Analysis of Survey Data* (pp. 75-84). Chichester: Wiley.

Commentary by

Risto Lehtonen University of Helsinki risto.lehtonen@helsinki.fi

I want first to congratulate Harvey Goldstein for his inspiring debate paper titled "When and why do we need population representative samples?" Population representativeness versus sample purposefulness has been recently debated in epidemiology and social sciences literature. Rothman, Gallacher and Hatch (2013a) challenge the dominant role of representativeness in epidemiology and social and health sciences by asking why representativeness should be avoided and arguing that "...studies that control skillfully for confounding variables and thereby advance our understanding of causal mechanisms" offer a proper route ahead (1014). According to Rothman, Gallacher and Hatch (2013b) "representativeness, although it may have a place in health surveys, is not a proper goal for scientific studies" (1027). By "scientific studies" he refers to causal studies about how nature operates.

In his debate paper Goldstein addresses several points that remain unclear in Rothman's writing. I agree with many of Goldstein’s arguments. Both he and Rothman seem to restrict what they call "representative sampling" essentially to "enumeration" or "population inference" purposes, that is, the sample data set is used to estimate the parameters of a well-defined finite population, for example the prevalence of chronic disease in age-sex-groups in a given real population at a given time point. Later in the paper Goldstein however widens his framework beyond that of Rothman. As an example, he describes the follow-up study design of the British "Life study". For that study he proposes the use of additional register-based population information (sometimes called auxiliary data), supplementing the original study data, for both descriptive (enumeration) purposes and for studying scientifically interesting hypotheses. He considers the combined use of data taken from different sources to represent a special case of purposive sampling. Goldstein thus proposes a kind of hybrid solution: "...an optimum design may well be one that combines such purposive sampling with population representativeness, so serving both enumeration and scientific aims". In my opinion, this is a fruitful view and I will try to elaborate this approach further in my commentary.

What is meant by ‘representativeness’ and ‘purposefulness’?

Population representativeness (representativeness for short) and representative sampling are key concepts in Rothman and Goldstein’s papers (29 hits in Rothman and 25 in Goldstein) but the concept itself remains unclear. This is not necessarily a surprise because there is no universally accepted definition of representativeness or representative sampling. In a series of four papers on representative sampling published in the International Statistical Review in 1979 and 1980, William Kruskal and Frederick Mosteller give nine different definitions of representative sampling they have found in scientific literature. All definitions are loose. Freshmen may think population representativeness refers to a miniature population obtained by representative sampling i.e. study subjects are selected from the population with an equal chance of being included. This interpretation is far too simplified because such a design only represents a special case of probability sampling. Even if the term "representativeness" is rarely used in modern survey sampling literature, we might think of population representativeness as a procedure where the study subjects are selected with a specified random mechanism from a well-defined finite population, either with equal or varying probabilities. If drawn with varying probabilities, the structure of the realised sample data set is restored (or forced to be "population representative") by weighting the observations by the inverses of the inclusion probabilities. Obvious benefits of probability sampling are in its flexibility for a controlled selection of the study subjects and in its ability to provide a basis for proper statistical inference under the actual sampling design. For example, oversampling of understudied groups would be covered, as suggested by Rothman. However, the scope of representative sampling in Rothman's paper seems narrower (this also holds for Goldstein's paper). Unequal probability sampling is not explicitly covered, as can be inferred, for example, from Rothmans's rebuttal (2013b p. 1026). When reading both papers it is hard to disagree with Kruskal and Mosteller who suggest avoiding the use of the concept of representativeness. In epidemiological literature, the term is occasionally used without clarification but it is fair to say that in some cases, a reasonable explanation is given (An example is Rothman, Greenland & Lash, 2008 p. 146).

Purposive sampling is another key concept in Goldstein's paper (Rothman does not use the concept of purposive sampling). This concept is problematic as well. Purposive in what specific sense? In Goldstein's paper, purposive sampling refers to a sample that is "non-representative of any particular real population". Now, it remains unclear whether a probability sample from a real population becomes "purposive" because of serious and informative nonresponse or if, instead of probability sampling, a quota sampling method or a self-selection scheme has been used or, alternatively, if the realised sample data set is being interpreted to be "representative" of a fictitious superpopulation. Later on I will come back to purposive sampling from a survey statistics point of view.

As a curiosity, there is a certain discrepancy between representativeness and purposefulness going back to the infancy of probability sampling. In a seminal paper entitled "Den repræsentative Uildersøgelsesmethode" (The representative method of statistical surveys) published in 1897 by a Norwegian statistician Anders Kiær, the term "representative" appears for the first time in survey sampling literature. His main argument was that it is not necessary to implement a census to obtain useful information on a human population but to carry out a "partial investigation". Fulfilling a well-specified type of representativeness on the population structure, would be enough to make inferences on the whole population. But in fact, the method of Kiær is a kind of combination of representativeness and purposive sampling (see e.g. Langel & Tillé, 2011).

The contradictory nature of the two key concepts, representativeness and purposefulness, has given rise to much debate and misinterpretation for decades (and the contradiction is, implicitly or explicitly, visible in both Rothman and Goldstein's papers). The discriminatory power of the terms is weak as is evident from example in the paper by Goldstein. As the climax of his paper, a certain type of data combination appears effectively to be both purposive and representative, indicating a complete overlap of the concepts. So, we are back in Kiaer!

The hybrid solution revisited

Let me elaborate further Goldstein's hybrid solution by using ideas from modern survey statistics. The key idea is to successfully combine, in one way or another, methods used in the sampling phase for the selection of study subjects and the methods used in the analysis of the study data. In both sampling and analysis phases, auxiliary population data taken from administrative registers or censuses and statistical modelling can play a crucial role. For example, in balanced sampling (Deville & Tillé 2004) the sample is forced to fit with the known population distribution of selected auxiliary variables, in effect representing purposive sampling with properly defined inclusion probabilities. In the analysis phase, the effect of varying inclusion probabilities caused by balancing can be adjusted for by weighting the sample observations with inverse inclusion probabilities, which is a standard survey analysis practice. Alternatively, the effect of balancing can be accounted for by including the balancing variables as potential explanatory variables in the statistical model to be fitted to the study data set, representing a possible model-based way of treating sampling complexities. As an extension for the analysis phase, statistical calibration techniques (e.g. Särndal, 2007) offer methods for the construction of calibrated weights that force the sample distribution of selected auxiliary variables (covariates; e.g. demographic, socioeconomic etc.) to fit with a known population distribution. The weights (possibly combined with the original survey weights) are then supplied to the analysis procedure (as weight variables or covariates). Thompson (2015) addresses complex longitudinal surveys from both a survey analysis and model-based analysis point of view. Gelman (2007) discusses weighting in the context of Bayesian analysis.

In my opinion, the hybrid design of combining the study data, the available auxiliary population data and statistical modelling fulfils many of the properties of an optimal design introduced by Goldstein. There are many favourable properties in this approach. The combined methodology offers a useful tool for the balancing of the sample distribution of important confounders against the known distributions at the population level, needed in studies based on purposive sampling and in probability samples that suffer from severe and informative nonresponse and selective attrition. Protection against model mis-specification can be attained for superpopulation-based approaches. If the inferential framework is model-based, the auxiliary variables (or the constructed weight variable) - featuring important aspects of the sampling design and nonresponse patterns - might be included as covariates in the statistical model to be fitted in the analysis phase. Effective adjustment for informative nonresponse and attrition can be attained if the auxiliary variables correlate with the response mechanism. Moreover, improved accuracy is possible if the auxiliary variables correlate with the study variables.

Obviously, the hybrid methodology can be very effective in "enumeration studies" where probability sampling (with equal or unequal inclusion probabilities) plays an important role, even if the inferential frameworks may differ. This is because probability sampling offers a firm basis for statistical inference in any empirical science. With certain restrictions, the methods are applicable for non-probability samples as well. In "scientific studies", the approach can be used for example to protect against the possible selection bias of study subjects. Moreover, the methodology toolbox fades out the unnecessary or even harmful confrontation between “scientific studies” and “enumeration studies”, because with appropriate choices the methodology applies to both.

Requirements for data infrastructure

The power of the hybrid machinery described above depends on the data infrastructure accessible to the researcher. Even if there are huge differences in this respect between countries, aggregate-level auxiliary data on demography, health and social affairs are often available in population censuses, official statistics and administrative registers, fulfilling minimum requirements for the methodology. The British "Life study" described by Goldstein offers a good example. Li, Li and Graubard. (2011) illustrate the importance of accounting for the complexities of the study design (stratified multi-stage sampling involving intra-cluster correlation, informative nonresponse accounted for with weighting and calibration to census totals) in order to obtain valid inference in a genetic study. The study shows the potential of the combined methodology in a data infrastructure where aggregate-level census data are available.

In the so-called register countries, notably in the Nordic countries, including Denmark, Finland, Norway and Sweden, unit-level data on various auxiliary variables are available from statistical register and from administrative sources for scientific research in epidemiology and social and health sciences. Examples of data sources are health registers and registers on socio-economic conditions (see e.g. Gissler & Haukka, 2004). In such an infrastructure, the various administrative register files can be linked cross-sectionally at the unit level and also in a panel fashion. The combination of the administrative data sources into integrated statistical registers at the unit level is based on unique identifiers such as personal identification numbers. In many cases, records from the register databases can be linked with the original study data records at the unit level, giving much flexibility in the combined use of the various data sources. Jousilahti, Salomaa, Kuulasmaa, Niemelä & Vartiainen (2005) provides an example of data linkage and the use of combined information in examining drop-out and attrition structures in a health study conducted in a register-based data infrastructure. Fortunately, in many countries such data infrastructures are becoming accessible for scientific research and public statistics purposes.

Conclusion

From the statistical methodology perspective, the dichotomy between “scientific inference” and “population inference” is restrictive and prevents full utilisation of the potential of modern statistical apparatus and today’s emerging data infrastructures. Alongside relaxing this dichotomy, the confrontation of representativeness and purposefulness becomes unnecessary and can be dropped from the researcher’s terminology toolbox. It will also be necessary to introduce up-to-date materials in university courses in epidemiology on such topics as sampling and data integration and statistical record linkage techniques as well as analysis methods for complex study data. I agree with Goldstein’s comment on the importance of access to suitable statistical software in exploiting a combined study design.

Goldstein seems to neglect somewhat the potential of probability sampling as an important phase of the research process but I think that an obituary for probability sampling is premature.

References

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika 91*, 893–912. http://dx.doi.org/10.1093/biomet/91.4.893

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science 22*, 153–164. http://dx.doi.org/10.1214/088342306000000691

Gissler, M. & Haukka, J. (2004). Finnish health and social welfare registers in epidemiological research. *Norsk Epidemiologi 14*, 113–120.

Jousilahti, P., Salomaa, V., Kuulasmaa, K., Niemelä, M., & Vartiainen, E. (2005). Total and cause specific mortality among participants and non-participants of population based health surveys: a comprehensive follow up of 54 372 Finnish men and women. *Journal of Epidemiology & Community Health 59,* 10–31. http://dx.doi.org/10.1136/jech.2004.024349

Langel, M. & Tillé, Y. (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory. METRON - *International Journal of Statistics LXIX*, 45–65. http://dx.doi.org/10.1007/bf03263549

Li, Y., Li, Z. & Graubard, B.I. (2011). Testing for Hardy Weinberg equilibrium in national household surveys that collect family-based genetic data. *Annals of Human Genetics 75*, 732–741. http://dx.doi.org/10.1111/j.1469-1809.2011.00680.x

Rothman, K.J., Gallacher, J.E.J. & Hatch, E.E. (2013a). Why representativeness should be avoided. *International Journal of Epidemiology 42,* 1012–1014. http://dx.doi.org/10.1093/ije/dys223

Rothman, K.J., Gallacher, J.E.J. & Hatch, E.E. (2013b). When it comes to scientific inference, sometimes a cigar is just a cigar. *International Journal of Epidemiology 42,* 1026–1028. http://dx.doi.org/10.1093/ije/dyt124

Rothman, K.J, Greenland, S. & Lash, T.L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*, 99–119.

Thompson, M.E. (2015). Using Longitudinal Complex Survey Data. *Annual Review of Statistics and Its Application 2*, 305–320. http://dx.doi.org/10.1146/annurev-statistics-010814-020403

Population sampling in longitudinal surveys: a response from Harvey Goldstein

Harvey Goldstein
University College London and University of Bristol, UK
h.goldstein@bristol.ac.uk

I am very pleased that my original piece has stimulated an excellent set of thoughtful responses. Reading through these has given me greater insight into the issues and also persuaded me to clarify some of my views, especially on the role of scientific inference. I shall begin by reflecting on the terms that I used since I think that there may be some misunderstanding of my intentions, no doubt through insufficient elaboration originally on my part. I welcome the opportunity to provide elaboration and am grateful to all the contributors for their responses.

I use the term 'real' population, in the same sense as Kish (1965, chapter 1) to mean a finite set of units that, at least in principle, can be enumerated. The intention behind the use of the term 'purposive' sampling is to reflect sampling from a theoretically defined frame of reference but one that does not necessarily correspond to such a population. Thus, the pregnancy component of Life Study is well defined as all the pregnant women attending a set of maternity units. The sample is chosen as those attending over a given time period of four years. Here, the concept of a 'superpopulation' is a key one, namely that any scientific inference that is based upon a sample chosen at a particular time is intended to apply more generally across a time period, a point elaborated by Peter Lynn. We may also wish any inferences we make to apply across space, and both these concerns need to be addressed within a standard scientific framework as I elaborate below.

Such a sample indeed might consist of all the members of a 'real' population, such as the set of children attending primary school in England in year three of their education. Yet scientific inferences about, for example, the relationships between year three childrens' school performance and background factors such as ethnic group, need to postulate a superpopulation model, and we would apply basically the same modelling procedures whether the full population of year three children or a random sample from it, i.e. with known probability of selection, had been chosen. In Life Study, the chosen women would not conventionally be regarded as constituting a probabilistically selected sample from a real geographic population in terms of a fixed time period and by where they live (rather than where they attend), especially as the criteria for being able to attend may change over time, for example in terms of residence or referrals. Nevertheless, for scientific inference purposes, given suitable statistical adjustments, for example to correct for selection biases, we may apply our standard statistical modelling procedures where we attempt to make inferences conditional on individual characteristics such as ethnic origin etc., and we can see that the distinction between a purposive sample and one derived probabilistically from a real population becomes less clear and certainly less important. Let me be clear also that I certainly do not use the term 'purposive', in one of the senses discussed by Risto Lehtonen, namely as a sample that has become biased through selective non-response. As he mentions, an interesting example of purposive sampling is quota sampling where sample members are selected for certain characteristics they happen to possess. Of course, this is not based upon a clear probabilistic mechanism, *but if we are prepared to assume* that the selection process has not differentially sampled individuals who have other characteristics that mediate the relationships of interest, then we will be justified in applying our models to study the relationships of interest. One task for the data analyst is to try to satisfy such an assumption.

The key idea is that it is the underlying social and biological processes that produce an actual set of individuals, that are the real objects of inference, and we are making use of the biological and social realisation of these at a particular historical time to select a sample from which we may make inferences. Unless we make an assumption of this kind, what we describe is, strictly logically, only of historical interest, although of course, for the purpose of enumeration or, say, resource allocation, this may be appropriate. Furthermore, of course, we also need to assume that the process that generates the actual data is essentially probabilistic in order to make inferences about the parameters in our statistical models, and in addition that we have data that allows us to adjust for factors such as differential non-response that could otherwise lead to biases.

Thus, the Life Study maternity component samples all women over a period attending the maternity units, but it is the superpopulation that 'generates' this group that is of scientific interest and that, conditional on observables, the generation process is assumed random. I accept that there is a vagueness here that contrasts with the strictly defined procedures of the conventional survey framework for selecting probability samples from actual 'real' populations, but it seems to me that we need to accept this in the spirit that whatever inferences we come up with are subject to the strict scientific tests of replication and falsification. These tests are what I was trying to illustrate in discussing the studies of pregnancy smoking and mortality. Thus, I concur with Colm O'Muircheartaigh's remarks about the importance of samples that have a probabilistic basis, since this is fundamental to statistical modelling, but I also contend that such a probabilistic basis is consistent with a superpopulation approach. The points made by Graciela Muniz and Rebecca Hardy about the importance of replication and generalisation are helpful here. I hope that my original intentions may now be clearer, especially in the light of Peter Lynn's useful discussion of real and super - population definitions.

I think my original use of the term 'real population' and a 'purposive' sample may have led to some misunderstandings. Thus Colm O'Muircheartaigh points out that had we taken notice of the study across the whole of the US where there is considerable heterogeneity, rather than the private health one or the Swedish one, we could have observed the positive relationship between percentage low-birth weight and mortality ratio that I presented. Rather than undermining my point, however, that is precisely my contention in that it is the heterogeneity present in the sample rather than the fact that it allows inference to any particular geographically well-defined population, that is of key importance. I particularly welcome Risto Lehtonen's discussion of purposive sampling and how, for example by the use of properly specified inclusion probabilities, weights and covariate adjustment, such samples can be brought within a standard statistical modelling framework.

Turning again to my illustrative example of Life Study, choosing to sample from maternity units was not, as Chris Skinner suggests, based on 'geographic homogeneity', but the practical one that this was the only way to obtain high quality prenatal measurements. He is right that there will no doubt be important differences among maternity units in different parts of the country and one aim of analysis will be to explore and attempt to account for these. This is part of the scientific process of replication. In fact in the case of long term longitudinal studies, apart from the relatively small number of national cohort studies in the UK and elsewhere such as the US, Canada, the Nordic countries, France, Germany, the Netherlands, New Zealand and Australia, most are samples of small geographic regions, institutions or other restricted groups. Indeed, the 1946, 1958 and 1970 British cohort studies sampled all births in just one week, in one sense a real population, but certainly not the target population of interest. The scientific value of such studies lies not primarily in their general representativeness but in their heterogeneity, their ability to explore rich data and ultimately in the possibilities for comparison and replication. In the case of Life Study, access to the national population births register and also to local population data, containing birth and demographic variables, also allows us to post-stratify the sample and to adjust for differential non-response by conditioning on such data. It will also allow us to compute weights so that it can be used together with the parallel national probability sample in Life Study to provide efficient combined analyses, the 'borrowing strength' that Chris Skinner refers to. While he is correct that it adds relatively little *national* information, it will provide the user with a consistent and large combined dataset that contains both sample components. Thus, depending on the purpose of any particular analysis and using appropriate weights, one may certainly treat the overall sample as 'representative' of a real population (over an intended four year period), but one may also treat it as a realisation of a superpopulation process. I do agree that such designs are non-traditional and would benefit from further study, and Risto Lehtonen's remarks under the 'hybrid model' heading provide a useful elaboration of the basic idea, and his illustration from registers constituting a data 'infrastructure' for removing sample bias in Nordic countries is interesting.

I'm grateful to Graciela Muniz and Rebecca Hardy for usefully illuminating all these issues with their discussion of cohort studies and especially how difficult the concept of representativeness of a real population becomes over time. They also elaborate on the need for replication and reproducibility and how this may be achieved, with some well-chosen examples.

Since preparing my original article, an interesting paper has been read to the Royal Statistical Society by Keiding and Louis (2015) that has a detailed exploration of many of these issues and explicitly comments on the articles by Rothman and colleagues (2013). They argue, I think correctly, that in some respects Rothman and colleagues overstate their case. Keiding and Louis particularly draw attention to the problem of informative differential non-response that can threaten the validity of any inferences, and I fully concur with this as a major issue for all types of study. They also take the view that "The real representativity issue is whether the conditional effects that we wish to transport (to other times and places) are actually transportable". This echoes my remarks about conditioning on known population data to avoid selection bias. I think that the Keiding and Louis paper, however, is less clear about the relationship between scientific inference and inferences to a well-specified population. As I pointed out in the case of the smoking in pregnancy studies, the characteristics of some populations may make them quite unsuitable for purposes of scientific explanation.

Despite remaining differences I am encouraged that there is a general agreement that these issues are useful ones to discuss and I have no doubt that there will be plenty more to say in the future.

Acknowledgements

I would like to thank the editor, John Bynner, for setting up and encouraging this debate and for providing helpful comments on various drafts.

References

Kish, L. (1965). *Survey Sampling*. Wiley: New York

Keiding, N. & Louis, T.A. (2015). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society*, series A. To appear (with discussion) Retrieved from http://www.rss.org.uk/Images/PDF/publications/rss-preprint-keiding-august-2015.pdf

Referencing

The debate should be referenced as:

Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R. (2015). Population sampling in longitudinal surveys debate. *Longitudinal and Life Course Studies, 6, 447 – 475.* http://dx.doi.org/10.14301/llcs.v6i4.345

Individual contributions may be referenced as:

[Author(s) name(s)]. [Title of contribution], in Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R. (2015). Population sampling in longitudinal surveys debate. *Longitudinal and Life Course Studies, 6,* 447 – 475*.* http://dx.doi.org/10.14301/llcs.v6i4.345

COMMENT AND DEBATE

Social class differences in early cognitive development: a response from Leon Feinstein

Leon Feinstein Early Intervention Foundation and London School of Economics
leon.feinstein@eif.org.uk

(Received September 2015 Revised October 2015) http://dx.doi.org/10.14301/llcs.v6i4.385

Introduction

The July edition of this journal (Feinstein et al. 2015) included a special comment and debate section comprising five papers, including my own, on how the trajectories of cognitive skill through development of children vary with respect to family of origin and early scores at different times and places, how these trajectories might be modelled and what can be inferred from this about the impact of social structure. This debate impinges on what might be said about the opportunities for policy to address structural inequalities in child development. The focus of the debate was a graph in Feinstein (2003) that is recognised as influential in discussions about early intervention but has been criticised as being flawed in a number of ways (Tu and Law, 2010; Jerrim and Vignoles, 2013). I summarise the general background in Feinstein et al. (2015) so do not repeat that here. The annex to this paper includes figure 1 and figure 2 from Feinstein (2003) and the reader is directed there or to Feinstein et al. (2015) for a description of methods. I am particularly grateful to the authors of the comment and debate papers who have contributed further insight and reflection on the underlying questions and to the editors for letting me respond here to the thoughts they set out in their commentaries.

Since the 1970s structural inequalities of wealth and income have increased very substantially (Atkinson, 2015) as have public and private investment in cognitive development (Willetts, 2011). It would be interesting to know how these and other trends have changed the degree to which social circumstance influences the development and educational achievement of children so it is important to have reliable, comparable developmental and structural measures. It would be useful to test whether cross-national differences in the structure of inequality in a nation are reflected in differences in structural inequalities in children's development, but this requires sound, agreed and comparable ways to measure and model trajectories. Until we have clarity on what differences in trajectories are due to measurement or to the way trajectories are modelled we cannot test what differences are socially structured by time and place, nor what might be thought local and what more universal. So I am grateful to the co-authors of Feinstein et al. (2015) for proposing methods that might allow more comparative study. They all offer exciting potential and I hope they are used widely. Between them they offer the possibility of triangulating one approach against another with the aim of achieving convergent conclusions. My remaining comments are intended mainly to highlight some of the difficulties and uncertainties in thinking about the implications of these methods to the question of the possibility of intergenerational change. I also indicate why I think that despite decades of policy failure there is room for optimism.

Numbers and words

I would like to bring out two themes from these papers. The first is that this is not a debate solely about statistical methods. If the graph published in my 2003 paper and reproduced in figure 2 had only ever been published in an economic journal and never been used in policy debates it would have inspired much less discussion. It is the use of the statistics as much as the statistics themselves that is at issue. This theme is emphasised by Lupton (2015) who stresses the importance of inter-disciplinarity, highlighting that much of the important evidence on why there is a social class achievement gap comes from qualitative and sociological research. She points to the problems resulting from the dominance of variable-based statistics and econometrics that appear to have had a much more prominent profile and influence in the analytical work of government, a theme also picked up in Feinstein and Peck (2008). Lupton asks why qualitative and sociological approaches are so often ignored in policy debate. This is an important contribution because she asks us to consider not just which statistical method is best but what might be the limitations of a statistical understanding, a theme not picked up in the other papers. I would only add that in my experience there are few social policy questions on which practitioners and policy makers would consider information solely from narrow quantitative sources sufficient. However, it does appear to be true that quantitative analysts are paid a premium compared to qualitative researchers in government as in the private sector.

Lupton also highlights the importance of language, of how debates are framed. This is not just because of the differences between and within science and social science in how study is undertaken and what terms are used, but also because of the large gap between this and the public or policy understanding of these same terms.

Measuring, modelling and inference

The second theme and a main focus of the other three commentaries by Goldstein and French, Jerrim and Vignoles and Washbrook and Lee (Feinstein et al. 2015) is ways to measure and interpret the degree to which pathways or trajectories are shaped by early scores and family of origin. Three different modelling frameworks are discussed, all of which offer advances in estimation of a more specific set of questions than addressed by figure 2 of Feinstein (2003).

Some points of consensus

Jerrim and Vignoles (2015) provide a useful summary of five points of consensus. Three relate to what is known statistically about the emergence of gaps between social groups in cognitive development through childhood, although the scope of this assessment in terms of time and place is not defined. First, there are large average gaps by socioeconomic group in cognitive skill that can be observed from a very early age. Second, these do not decline through schooling in absolute or, third, relative terms.

It is clear from their paper that there are some important starting points but also many questions on which we have at best cursory and preliminary understanding and substantial disagreement. We cannot say categorically whether clear gaps in the attainment of children living in the UK currently broaden through childhood. I also note that there is not a clear specification of what is meant by a large gap. For the wider public debate it is also important to emphasise that these assertions are only true about averages; in public debate an average difference between groups is often interpreted as a universal difference between all members. Important also to recognise that these are historically contingent statements, true for specific times and places and not general truths.

The final two summary points made by Jerrim and Vignoles relate to understanding of the graph, thus fourth: “The striking decline between 22 and 42 months should *not* be used by academics or policymakers to stress the importance of the early years, that we are failing ‘bright’ young people from disadvantaged backgrounds, or to highlight the lack of social mobility in the UK.” The point was made best by an official in the Scottish Government at a seminar in Edinburgh in 2012, who remarked that if she had known it was so complicated she would never have used the graph. I agree the shift between 22 and 42 months says nothing about the degree of social mobility and does not support the claim that anyone is being failed. As I pointed out in Feinstein (2015) this was never my claim on the basis of the 22 to 42 month shifts nor in my view is it indicative of how the graph was used in general, not least because the shifts between 22 and 42 months are so large for both high and low socioeconomic status (SES) groups. However, I do recognise there has been substantial confusion about this.

The final point of consensus noted by Jerrim and Vignoles is that there remains no robust and consistent evidence that initially high achieving young people from poor backgrounds are overtaken by low achieving children from affluent backgrounds in terms of their cognitive skills. True, although as Washbrook and Lee (2015) point out, whether this crossover happens or not depends on arbitrary assumptions about the cut-points by which children are allocated to groups at the first age of measurement. In many ways figure 2 is best thought of as the corollary of figure 1, which concerns the change through development in the average gap by social class. If the SES gap broadens through development, more and more children from low SES families who score well early on will fall back relative to high SES children.

Washbrook and Lee (2015) offer a useful general approach that takes us away from the discrete approach of the figure 2 graphic and the debate about a crossover. The real question of interest for them is "whether low SES children systematically underperform relative to higher SES children with identical initial capacities," and, as they go on to show, how this differs by age during development.

Washbrook and Lee treat the difficulty of measuring initial capacities as a measurement error problem and therefore responsive to econometric and statistical methods for dealing with measurement error, in particular the technique of instrumental variables (IV) by which an auxiliary measure unrelated to later scores is used to predict the initial cognitive performance score. Under assumptions set out by Washbrook and Lee the use of IV can address the measurement error aspect of the problem of accurately measuring initial abilities. A particular technical difficulty is the challenge of establishing that the instrumental variable does not contain predictive power for later scores, independently of the other factors such as SES included in the statistical model. They offer alternative approaches based on correcting for the reliability of test scores.

Using a prior cognitive score taken just after kindergarten entry as an instrumental variable for a large US sample of children who entered kindergarten in 1998 they find that a good part of the widening in scores between high and low SES groups with equivalent initial capabilities happens between the ages of seven and 14. This analysis does not claim to solve the problem of explaining the divergence by family background but offers a useful general approach to estimating and describing the degree and timing of such divergence given early abilities, having accounted for measurement error. Others will challenge them on their choice of instrumental variable but their approach is promising. I discuss below the question of defining what is meant by initial ability in the context of a debate about true ability and the potential for social change.

Goldstein and French (2015) make two explicit challenges, asserting "flaws" in both my original analysis and my recent response. They also treat the issue as one of measurement error and propose a general framework based on Bayesian modelling not yet published. In my view these flaws indicate the difficulty of framing a clear, common language for understanding the issues involved in modelling the data presented in figure 2, rather than flaws of analysis, but it is useful to set out their comments and, given their use of language, I am compelled to respond.

The first of these alleged flaws is my statement that I was offering in figure 2 "a descriptive analysis," - as I put it "the aim was to present the actual data rather than… corrected trajectories based on modelling assumptions." Goldstein and French argue that there is no difference between an uncorrected and a corrected presentation of data, "each is meant to convey an inference about the underlying social process." They are right of course that the actual data presented in figure 2 are not simple facts about a social process and inferences about what influenced it were implicit. The dependent variable itself is the rank in the first principal component of a set of age appropriate measures of cognitive development taken at four ages in development based on different underlying measures. I only included in the chart the average scores of children characterised as low and high SES, excluding the larger middle SES group and including only those scoring in the first or fourth quartile in the 22 months tests. My original paper (Feinstein, 2003) presented tables of the full set of transition matrices that showed how the children who scored in each quartile at 22 months scored at subsequent ages. This selection of specific cells for the graph has been called "extreme" (see Feinstein, 2015). I make no claim of pure objectivity in a transcendent world of pure fact.

However, I disagree if Goldstein and French wish to assert that a strong inference about an underlying causal process is a necessary element in all social science. Shame if a social scientist cannot explore without prejudice in an attempt to measure and describe the world, only able to test dimly understood hypotheses. In my introduction I made reference to a number of important hypotheses for policy making that do need testing in an attempt to get closer to understanding the underlying processes. However, scientific progress depends, as Lupton indicates, on a multiplicity of views and approaches and there is scope for descriptive and explorative social science, even provocative social science. Nor do I agree that the act of correction for error based on modelling assumptions, whether Bayesian, based on simulations constructed from general linear modelling or instrumental variables techniques, is a trivial addition to the challenge of measurement. The difficulty of specifying even what is meant by true score indicates the difficulty of framing a clear basis for appropriate inference.

This brings us to the second alleged flaw of my argument, which is in my understanding of regression to the mean and true scores. The former, they assert, "simply occurs when the correlation between two measurements over time is less than one, as is the case with heights of fathers and sons." The reference to an intergenerational transfer of associations is a mere detail in the modelling of correlation between two measures. They continue, "the notion of measurement error is entirely separate." Washbrook and Lee make a similar assertion, although explicitly recognising that measurement error is one component, amongst others, of the problem of regression to the mean as described by Jerrim and Vignoles. Indeed, much of the Jerrim and Vignoles reanalysis focused on the difference between high and low SES groups in the degree of change between 22 and 42 months. They emphasise that in the classification of children as high ability at 22 months, more low SES children are misclassified than are high SES children resulting in a higher degree of what they call regression to the mean for the low SES children in the 22 and 42 month scores. They emphasise the role of measurement error (and luck) in this misclassification.

Goldstein and French also point to a lack of clarity in my discussion of the statistical notion of 'true score' which in their specification "proceeds from the common observation that the actual score that a child obtains on a test will depend on the actual items chosen, plus other factors that might be considered 'transient' (their quotation marks) such as time of day, test environment etc."

This may be a fair if imprecise and general specification of the statistical notion of measurement error but ignores the distinction I make between this imprecise notion of a true level of capability that a social scientist might seek to measure and the potential capability of the child, a distinction elided in the simple use of the phrase 'true ability', and in application of these data to the question of the possibility of social change. This double measurement issue is not addressed by these statistical models, which is why I remain cautious about the use of corrected data presented as facts about children of different ability groupings. However, Goldstein and French are right to admonish me with their regret that in my handling of the trends by sub-group in the 1958 cohort study (Feinstein, 2004) I did not reference Goldstein's important work on this topic. I do not argue that corrections for measurement error are inappropriate or misleading, I merely note that they are not trivial nor are their own limitations always as clear as they might be.

Goldstein and French argue that as this debate has been "a difficult one for policy makers... a more cautious long-term attitude should be taken to research findings." They suggest that policymakers should "promote a wide debate about any findings that appear important, where technical and interpretational issues are debated in terms which are widely accessible." I look forward to this and hope that future contributions to the debate from statisticians offer more aid to accessibility than hitherto.

There is more focus in these papers on how to measure the extent to which cognitive development is moderated by early scores and presumed underlying abilities than on the problem of defining social groups or measuring cognitive skill – weaknesses also of Feinstein (2003). There is no discussion of problems of aggregation in interpreting data on profiles of averages without adequate specification of a coherent multilevel framework, yet many in the public debate struggle to appreciate the difference between aggregate findings reflective of general social averages and the likely experience and outcomes of individual children. The ecological fallacy (see e.g. Diez-Roux, 1998) in statements such as that "it is all over by age 5" based on charts of averages such as figure 2 is widespread and misleading.

As stated above, a focus of both figure 2 and the original paper from which it came is on the instability of scores. I presented transition matrices, which showed a great deal of movement by children in their test scores over time. It was never my intention to imply that there exist fixed groupings of ability. As ever the fact that something can be quantified and measured is not proof of its existence and we should beware the over-determination of the meaning of statistics and estimated coefficients. This is a problem for estimates of all kinds, useful if handled carefully but for which the apparent precision of numbers suggests a certainty that is easily over interpreted, particularly in a policy debate in which balance of interpretation is hard to achieve.

My experience in using the graph with policy makers was that using the uncorrected data enabled me to point both to the issue of score instability and to the subsequent patterns at the later ages which under reasonable assumptions are not explained by measurement error (Feinstein, 2015). Presenting corrected data enables a policy maker to gloss over the difficulty of classifying children to groups of ability as though statistical science has resolved the underlying philosophical and biological issues, introducing further miscomprehensions and over-simplifications to the debate.

Genes

This brings me to an important aspect of explanation and causation raised by Jerrim and Vignoles who note the challenges highlighted in relation to understanding the role of genetics. They point to the important and fascinating work of Robert Plomin and the tradition of structural genetic research based on twins studies which indicates that for many observable features of human development there are variable but often substantial proportions of the difference in outcomes that are explained in a structural, statistical sense by genes. This has included analysis of outcomes like intelligence, social class, aspects of personality as well as more obviously physical phenotypes.

As Lupton describes, this set of findings has been linked by Saunders (2011) with the critique of figure 2 based on modelling of ‘true ability’ to imply that our current system of allocation of wealth and opportunity is both efficient and reflective of an underlying natural distribution of capability and hence fair as well. Clearly, as Jerrim and Vignoles, indicate this is a controversial topic. As Lupton (2015) explains this is said to have undermined the 2010 Coalition Government’s commitment to increasing social mobility so some discussion is necessary. I strongly agree with Jerrim and Vignoles that social science should not shy away from addressing the topic of genetics and biological science, not least to recognise how informative it is, how quickly it is changing and how broad is the opportunity it indicates for policy and practice. However, this does not negate the need for social scientific understanding of social questions.

It will be well known to readers of this journal, but is not yet always known in the world of politics, that there is a lot of dynamic complexity in the interactions between genes and environments. Epigenetics is finding that the interactions between environment and genome are so dynamic that it is falsely simplistic to think this a unilinear, biologically driven phenomenon (Carey, 2012). It is as wrong to overstate estimates of heritability as meaning destiny is fixed at birth as it is to ignore the evidence that genetics plays a role. The question of how much is to play for is not well established but we know, not least from genetically sensitive research designs (Weaver at al., 2004) that there is plenty of opportunity for practice and policy to play a substantial role in influencing intergenerational continuities of achievements and behaviour. There is no reason to look on the findings of structural genetics as implying any currently binding limit to the possibility of social change.

It is clear that epigenetic research is changing the nature of scientific understanding of the relationship between genes and environment. In a wide ranging review of the literature on the heritability of intelligence in the late 1990’s Neisser et al. (1996) found a substantial role for environmental factors and scope for intervention to address social gaps in intelligence, as well as strong indications that intelligence is a multi-dimensional construct with only partial relationship to life outcomes. A recent update of that review (Nisbett at al., 2012) found even more scope for impact of the environment as more has been learnt about the interaction of genes and environment and about how environments impact on outcomes. The sequencing of the human genome has not led to the identification of specific genes that explain intelligence and, given all of this and other evidence, it should be very clear that outcomes such as intelligence, social class or income are not fixed, innate or immutable.

The issue of course is one of degree: How much difference can environments make? What is a high degree of heritability? There are few models that go beyond assessment of the degree of heritability and ask what this means for intervention. A standard finding is that the structural genetic heritability of IQ is somewhere between .4 and .8 (Neisser op cit.). However as Bowles, Ginitis and Osborne Groves (2008) discuss, these estimates do not measure persistence across generations, they measure the proportion of scores of intelligence statistically explained at population level by the genetic inheritance of children. To go from these estimates about differences between individuals at population level to the assertion that the average gap in scores between children from groups defined on the basis of distal characteristics of parents such as occupation or income is genetic to any fixed degree is stretching the science beyond its basis in fact – because the social class of parents is not equal to their intelligence, which is not equal to their children's intelligence, which is not equal to school achievement. To go beyond this to the assertion that social class itself is genetic is a further false extension. Such a strong hypothesis would surely need substantial evidence including detailed information on how the heritability of diverse characteristics such as intelligence, motivation, character, physical health and beauty interact in practice with actual contexts to generate social outcomes that are correlated across generations. Until we have a clear specification of this social scientific question the heritability estimate is a number in search of a theory as far as its application to the average continuity of social position across generations is concerned.

This extension is social scientific rather than biological because of the lack of common heredity; no claim is made that poor children have a common and distinct gene pool. The claim that the transmission of SES across generations is explained to any degree by genes is a statement about how children with different genetic inheritances come to achieve common outcomes that are socially structured in times and places by social processes interacting with biological heredity. Yet there is no theory in structural genetics about social process, about how capabilities interact with resources and contexts at multiple levels to influence outcomes, nor about how this changes over time. Nor are there very much data on these things. It is odd that a supposedly biological underpinning of social outcomes should be put forward without reference to evolution as though genes are fixed and capabilities carry value at all times and places in unchanging ways. The nature of the relationship between the social class of one generation and that of the next depends heavily on the nature of the society in which the two generations are studied, so there are no universal truths in the few studies conducted so far – features of society that include banking systems, laws and schools, amongst much else not reducible to genes. Nor is there a theory of how evolution relates to history, of how diverse capabilities play different roles in changing social structures that interact with genes in the generation of biological and social change. The biological sciences of genetics and epigenetics are fascinating and important. Used carefully they yield important clues for social policy and social science but we must look to social and economic research to understand how societies operate in the sharing of wealth and opportunity, the justice or efficiency of our current allocation and what scope there is for change.

Acknowledgements

I would particularly like to acknowledge comments from Stephen Peck on an earlier draft and the ever-wonderful Judith and Max for patience over too many weekends.

References

Atkinson, A.B. (2015). *Inequality; what can be done*. Harvard. http://dx.doi.org/10.1111/j.2050-5876.2015.00834.x

Bowles, S., Gintis, H., & Osborne Groves, M., (2008) (Eds.) *Unequal Chances: Family Background and Economic Success*. Princeton

Carey, N. (2012*). The Epigenetics Revolution.* Iconbooks.co.uk

Diez-Roux, A.V. (1998). Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health, 88*,(2), 216-222. http://dx.doi.org/10.2105/AJPH.88.2.216

Feinstein, L., (2004). Mobility in Pupils' Cognitive Attainment During School Life. *Oxford Review of Economic Policy, 20*(2) Education. http://dx.doi.org/10.1093/oxrep/grh012

Feinstein, L. & Peck, S. (2008). Unexpected Pathways Through Education: Why Do Some Students Not Succeed in School and What Helps Others Beat the Odds? *Journal of Social Issues 64*,(1), 1-20. http://dx.doi.org/10.1111/j.1540-4560.2008.00545.x

Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6*, 331-376. http://dx.doi.org/10.14301/llcs.v6i3.361

Goldstein, H. & French, R. Differential educational progress and measurement error (2015). In Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6*, 331-376. http://dx.doi.org/10.14301/llcs.v6i3.361

Jerrim, J., & Vignoles, A. (2013). Social mobility, regression to the mean and the cognitive development of high ability children from disadvantaged homes. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(4), 887–906. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Jerrim, J. & Vignoles, A. (2015). Socioeconomic differences in children's test scores: what we do know, what we don't know and what we need to know. In Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6*, 331-376. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Lupton, R. (2015). The practice of policy-related research. In Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6*, 331-376. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Neale, B. (2011). Qualitative Longitudinal Research. In S. Becker, A. Bryman and S. Ferguson (Eds) *Understanding Research for Social Policy and Social Work*. Bristol: Policy Press.

Neisser, U., Boodoo, G., Bouchard, T.J. Jr., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D, Loehlin, J.C., Perloff, R., Sternberg, R. & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101. http://dx.doi.org/10.1037/0003-066X.51.2.77

Nisbett, R.E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D., & Turkheimer, E. (2012) Intelligence: New Findings and Theoretical Developments. *American Psychologist 67*(2), 130 –159. http://dx.doi.org/10.1037/a0026699

Saunders, P., (2011). *Social mobility delusions.* Civitas: London.

Tu, Y.K., & Law, J. (2010). Re-examining the associations between family backgrounds and children's cognitive developments in early ages. *Early Child Development and Care 180,*1243-12. http://dx.doi.org/10.1080/03004430902981363

Washbrook, E. & Lee, R. (2015). Beyond the Feinstein chart: Investigating differential achievement trajectories in a US cohort. In Feinstein, L., Jerrim, J. & Vignoles, A., Goldstein, H. & French, R., Washbrook, E. & Lee, R. & Lupton, R. (2015). Social class differences in early cognitive development debate. *Longitudinal and Life Course Studies, 6*, 331-376. http://dx.doi.org/10.1111/j.1467-985X.2012.01072.x

Weaver, I.C.G, Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., & Meaney, M. (2004). Epigenetic programming by maternal behavior. *Nature Neuroscience 7* 847-854. http://dx.doi.org/10.1038/nn1276

Willetts, D., (2011). *The Pinch: How the Baby Boomers Took Their Children's Future - And Why They Should Give it Back*. Atlantic Books.

Annex

Figure 1: Average rank of test scores at 22, 42, 60 & 120 months, by SES of parents

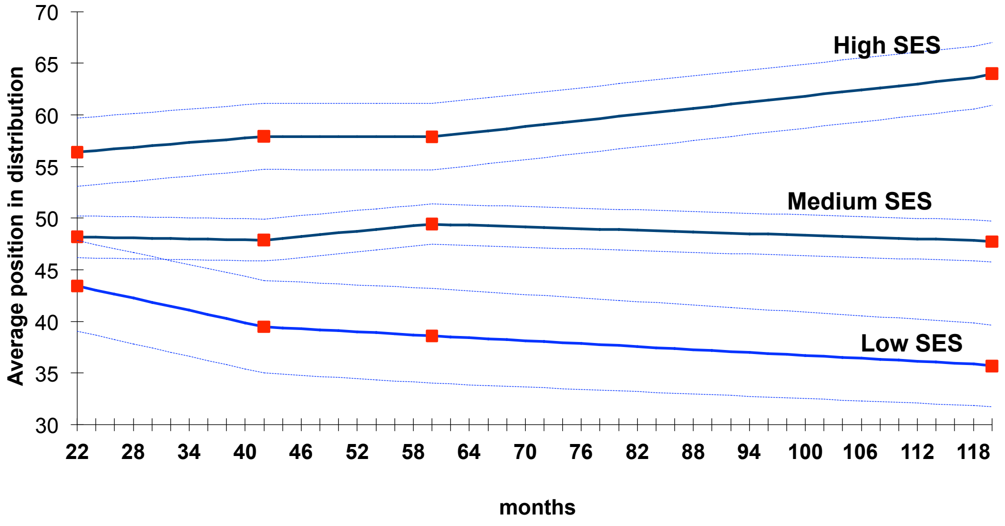
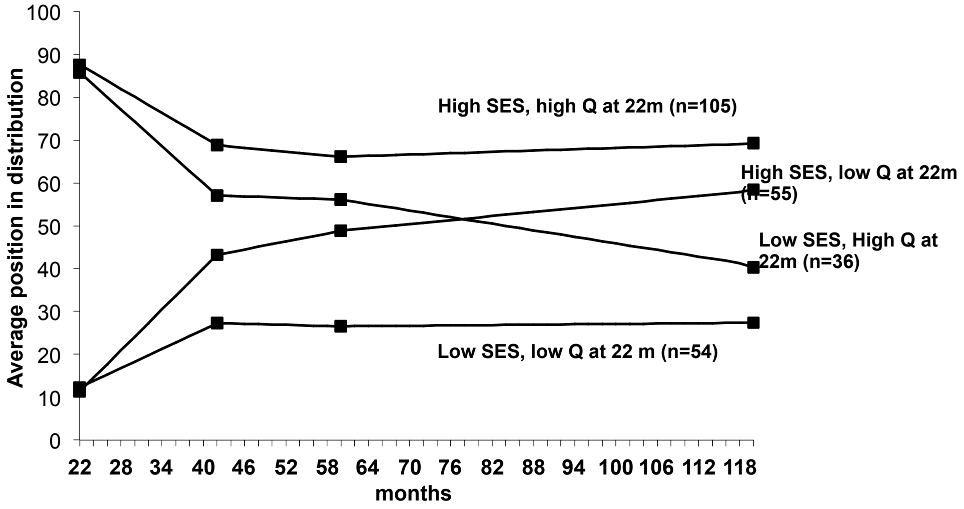


Figure 2: Average rank of test scores at 22, 42, 60 & 120 months, by SES of parents and early rank position



AUTHOR GUIDELINES SUMMARY

Submission of Papers

All papers, written in the English language, should be submitted via the *LLCS* website as a Microsoft Word file. If there is a good reason why this is not possible, authors are requested to contact crandall@slls.org.uk **before** submitting the paper. All subsequent processes involving the author are carried out electronically via the website.

Preparation of Texts

Length. The paper should normally be **approximately 5,000 words**, with longer papers also accepted up to a maximum of 7,000 words. This word count excludes tables, figures and bibliography.

Font and line spacing. Please use Calibri (or similar sans serif) font, 12pt, with 1.5 line spacing in all main text, single space in figure or table captions.

Page layout. All text should be justified to the left hand margin (all margins of at least 2.5cm) with no indents at the start of paragraphs and a line space separating paragraphs. All headings and sub-headings used within the body of the paper should also be left justified. Please do **NOT** use automatic text formatting.

Weblinks. To help our readers to look up cited references or other information available on the web, authors should ensure that all such references are activated.

DOIs. Do NOT include DOIs in the bibliography – they are added during editing as resolvable URLs.

Ensuring an anonymous (blind) review. Please submit papers with a full detailed title page. Once a paper has been submitted to *LLCS* via the website, it will be 'anonymised' by the removal of all author name(s) and institution names on the title page, and any identifying electronic document properties will also be removed. Authors do not need to remove their name(s) from the main text or references but any reference to their work or themselves should be expressed in the third person.

Abstract. The abstract (no subheads or paragraphs) should be no more than 250 words (not part of the main word count).

Keywords. These should be included just below the author list (minimum 3 maximum 10).

Abbreviations. Words to be abbreviated should be spelt out in full the first time they appear in the text with the abbreviations in brackets. Thereafter the abbreviation should be used.

References. Please use the APA 6th edition and refer to examples in the full Guidelines.

Authors not complying with these reference guidelines will be asked to make all the necessary alterations themselves, if their paper is accepted for publication.

Notes. As a general rule, supplementary notes should be avoided, but if thought to be essential, they should *not* appear on the page as Footnotes, but instead, be included as Endnotes.

Supplementary material. Supplementary material may be uploaded with the submission, and if the paper is published, will be visible to the reader via a link on the RHS of the screen.

Submissions containing graphs, tables, illustrations or mathematics. All graphs, tables and illustrations should be embedded in the submitted text, and have clear, self-explanatory titles and captions. Mathematical expressions should be created in Word 2003 (or a compatible package) with equation editor 3.0, unless the author has good reason to use other software, in which case please contact crandall@slls.org.uk. All biological measures should be reported in SI units, as appropriate, followed, in the text, by traditional units in parentheses.

Author citation*.* If the paper is accepted for publication, a version of the paper with all authors cited in full on the title page will be used. Only individuals who have contributed substantially to the production of the paper should be included.

Copy editing. All accepted manuscripts are subject to copy editing, with reference back to author with suggested edits and queries for response.

Proofs. The corresponding author will be asked to view a layout proof of the article on the website and respond with final amendments within three days.

(Full Author Guidelines at: http://www.llcsjournal.org/index.php/llcs/about/submissions#authorGuidelines**)**

Open Journal System

The ***LLCS*** journal is produced using the Open Journal System, which is part of the Public Knowledge Project. OJS is open source software made freely available to journals worldwide, for the purpose of making open access publishing a viable option for more journals. Already, more than 8,000 journals around the world use this software.

Copyright Notice

Authors who publish with Longitudinal and Life Course Studies agree to the following terms:

- Authors retain copyright and grant the Journal right of first publication with the work, simultaneously licensed under a Creative Commons Attribution License that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.
- Following first publication in this Journal, Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal, provided always that no charge is made for its use.
- Authors are permitted and encouraged to post their work online (e.g. in institutional repositories or on their own website) prior to and 'as approved for publication', as this can lead to productive exchanges and additional citations, as well as complying with the UK Research Councils' 'Green Model' for open access research publication.

INTRODUCTION

PAPERS

STUDY PROFILE

COMMENT AND DEBATE

LLCS Journal can be accessed online at: www.llcsjournal.org

Published by
The Society for Longitudinal and Life Course Studies
Institute of Education, 20 Bedford Way, London, WC1H 0AL
info@slls.org.uk http://www.slls.org.uk