

DOING RESEARCH WITH LARGE DATABASES: OPPORTUNITIES AND CHALLENGES OF THE SWISS RECRUITS' STUDY

Henriette Haas

Source code of Swiss recruits' study available to the scientific community

Composing a valid database and creating programs to interpret the phenomena means serious work for years before publications become possible. The author of the syntax is therefore happy to share her source code and make her full database available to qualified colleagues on the basis of a users' agreement to protect the rights to attribution and work integrity.

Covering over 70% of the age cohort of Swiss males at age 20, this cross-sectional study belongs among the largest studies on self-report delinquency ever conducted. Given the size of the sample and the fact that the questionnaire (Killias 1997, Haas 1997) covers almost the entire social biography of the males, this study is a powerful tool. Applying the source code on the database, one can address and resolve many controversial subjects and explore the depths of relatively rare phenomena in social science.

Methodological requirements

It was Martin Killias who has initiated the recruits' study in 1995. In 1996, the questionnaire was then constructed as an interdisciplinary instrument on the basis of Killias' research and methodology in (sociological) criminology and the author's contributions in forensic psychology (Killias 1997, Haas 1997). The main themes are physical violence and threats, sexual transgressions, statutory offenses and different forms of victimization covered each with an array of questions about different types of acts that were intentionally perpetrated during the 12-month period before the recruits' training. The questionnaire also considers the frequencies of such behavior and includes a set of follow-up questions for each topic (see Killias 1991). Follow-up questions ask who was implicated in the incidents (e.g. as a victim), what were the consequences, who knew about the incident. The comprehensive approach to the reported events opens possibilities to observe a very detailed picture of delinquency. However it also creates some challenges. The complex information dispersed in dozens of questions must be re-united into one single variable to define the topic, which is about to be analyzed.

To analyze such data is like composing a giant puzzle in order to discover the general picture and to interpret entire scenes in it. In this case, the puzzle consists of 19 million pieces (900 questions presented to 21,347 recruits). This was the goal of the author's work. The psychological concepts were operationalized in her source code and are described in her book (Haas 2001) together with the resulting analyses.

In their classic text collection on measuring delinquency, Hindelang, Hirschi & Weiss (1981, p.213) have pointed out that the differential validity of self-reported information can vary considerably from one social group to the other. So in 1998, when the author began to analyze the raw data, nothing was known about the quality of the answers given by the most severe offenders within the sample. She then created and programmed sincerity checks specifically adapted to delinquents. Some 25 questionnaires were excluded because of a total lack of credibility. Another eleven questionnaires were doubtful, but were left in the sample. As some recruits crossed several answers when only one was required (e.g. the frequency), there was a need to determine, which one of two or more ambiguous answers should be taken into the database (details in Haas 2001, p.409ss). Eight questionnaires had to be excluded because of

scanning errors. After these tests, the SAS database “tnouv.ssd01” with 21,314 valid observations was composed from the raw data files.

As control theory has pointed out (Gottfredson & Hirschi 1990) the one most specific characteristic of offenders is their impulsive nature, which – of course – will also manifest itself in the way they fill out the questionnaire of a survey. The recruits’ data confirm the existence of this phenomenon. Severely delinquent males not only showed a tendency to leave out items, they also tended to lose patience when filling out two-dimensional questions with many different items and frequencies. Sometimes, at the end of a long series of questions about one particular theme, they began to cross all the maxima (probably meaning to say “yes I have done this and that many times”). If the follow-up questions were also filled out with practically every item on the list, the questionnaire was excluded as lacking credibility. However, those who had impulsively crossed all items at the end of a theme, but were precise in the subsequent questions, were left in the sample. The basic idea behind this decision was to accept some exaggerated responses, because they cause less of a distortion than a severe sampling bias which would have resulted from excluding all impulsive delinquents from further analyses.

Considering the systematic influence of missing data

Regoeczi and Riedel (2003, p.155ss) have criticized a common shortcoming in methodology: *“Homicide cases suffer from substantial levels of missing data, a problem largely ignored by criminological researchers. [...] The question that remains is whether the data are missing at random (MAR). [...] If the MAR assumption does not hold, the missing data are nonignorable. [...] A fundamental approach to the problem of missing data is to learn the relationship between missing and nonmissing data and to use that information to impute what the missing values are likely to be.”* In the recruits’ data we are confronted with the same problem. Scholars who treat the recruits’ data will want to consider missing values as a significant influence. For multivariate analyses when the algorithm automatically excludes all observations with one or more missing, those values need to be imputed by the means (or the medians or the cluster centers) corresponding to the respective level of the criterion variable (see Lessler & Kalsbeek 1992).

Measuring delinquency with scales

In order to grasp a certain type of behavior or of personal characteristics, it is necessary to elaborate scales with several levels. The absence of thoughts on distinguishing different levels of severity of a problem can produce rather absurd statistics. Some works assert that 90% of the women in Western countries are victims of spousal “violence” by putting a verbal argument on the same level as 20 assaults with weapons.

Computing with ordinal and nominal scales follows mathematical rules that are not trivial. Some characteristics may be distributed on a bell curve, some on a Poisson curve whereas others may behave in an entirely chaotic way. Therefore, it is not recommendable to assume evenly spaced distances between ordinal levels of such a scale. The scale of delinquency in general (Haas 2001, p.225ss & p.426ss) provides a good example in this respect. It differentiates between three levels: (1) non-delinquent males, (2) average offenders and (3) persistent offenders. Correlating other characteristics with it illustrate that offending is not a continuum of behavior. The differences between the persistent offenders (who are responsible for the large majority of crimes, as was discovered by Wolfgang, Figlio & Sellin 1972) and the average offenders are much greater than those between the non-delinquent males and average offenders (further reflections see Cusson 1998, p.84).

Scales also need to be gauged according to statistical norms that have been established within comparable samples of the population (Hindelang, Hirschi & Weiss 1981, p.75, 87, 88). They mention that the problem of gauging scales has often been neglected by traditional sociology, whereas psychology – as a discipline – has invested great efforts to create gauged instruments since its very beginnings in the 20th century.

The available source code contains some 20 different composite variables and scales in order to grasp different concepts such as violence, possession and use of weapons, delinquency in general, transgressions against sexual integrity, educational control and bonds within the family (modeled according to the Glueck factors), and psychological disorders. The details of their construction can be found in Haas 2001 and subsequent articles. Some instruments are operationalizations of a diagnostic concept. Other instruments like the scales of “violence” and the scale of “delinquency in general” are innovations, thus prototypes. The diagnostic concepts of *conduct disorder* and *antisocial personality disorder* that were modeled closely to the criteria of the DSM-IV (1994) have proven their reliability and validity in countless studies. Their operationalizations were therefore used to gauge new concepts.

In practice, how does one go about to gauge a scale? First you need to test its stability. When minor changes of a definition are made, they should not cause an unpredictable variation of the overall number of individuals concerned by each level. Stability can only be established when none of the variables used to construct the scale behaves in a totally chaotic way. In a second step the external validity needs to be controlled by comparison with empirical data on a given phenomenon. Then again, there is the question of theoretical validity. Without good reasons new scales should not contradict previous concepts for measuring similar phenomena, especially when they have empirically proven their validity. Just like in psychopathology or in penal law, the criteria for each category or each level need to be quite encompassing. For example, the highest level of violence of the author’s scale (2001, p.188), defining the category of the most seriously violent offenders (n=341) states the following conditions to be met during the 12 months preceding the recruits’ training: “*more than 20 violent acts that have led to a bodily injury, any act committed with a weapon that has led to a bodily injury, 3 or more acts of violence with a weapon (knife, firearm, blunt object, poison) even if nobody was seriously injured, any act of theft while confronting a victim even in the absence of any other criterion*”. Killias (2001, p.176; 2002, p.185) also refers to this definition.

Working with some else’s source code

When treating large amount of data, one is confronted with a dilemma. The scales that are handy to use in order to assemble the dozens of variables on a topic into one can only be produced with long computer programs. The main working program with all the scales and composite variables is over 3,500 lines of SAS[®] syntax. It is a well-known but unresolved problem of computer syntax, that only its creator can read it fairly easily; for outsiders it can be quite illegible. Computer programs are individual creations, therefore autonomous use of another author’s long computer programs or of outputs that are not one’s own is second hand knowledge. The difficulties in applying the scales of the present source code stems from the fact that many auxiliary variables needed to be created for the construction of each scale. The names of these variables had to be short (maximum 8 letters) and they are not very obvious. This presents a risk insofar as someone who is not entirely familiar with the program could calculate correlations among variables that are both based on an identical ingredient. The advantage of the auxiliary variables is though, that they allow modifications of a scale when they should be necessary for a specific purpose.

Unfortunately reading outputs produced by a colleague has also become increasingly difficult, even for those who are familiar with the book, because of the necessity to modify composite variables according to specific goals. Incorrect interpretation can then lead to fallacies. Here is an example when such a modification could be necessary: One composite variable named “region” relates to the ethnic origin. It has three levels: the first covers males with both natural born Swiss parents (n=16,580). The second (n=3,495) covers those with at least one immigrant parent, who however does not come from a region in crisis. The third level concerns recruits with at least one parent coming from a region in social or political crises (Balkans, Near-East and Maghreb) (n=522). This variable is handy to grasp the overall influence of immigration. However, one should be careful not correlate it with variables that concern only one parent, e.g. the father, because it mixes the origins of both parents. If one wanted to analyze the correlation between immigration and excessive drinking or spousal violence by the father, then one should use another variable for immigration (Haas 2001, p.24), which treats each parent separately (e.g. 392 fathers coming from regions in crises).

May it be allowed to issue a word of caution for all those who consider the use of the recruits’ data without wanting to use or to create elaborate scales for defining their issues. Students may prefer to take just one particular aspect (for example one single variable of question 70h (“sexual penetration against the partner’s consent) and then call it “sexual offense” in their paper. Next semester, maybe even at the same institute, another student writes on sexual violence by taking different items and not astonishingly produces different results. If published, the contradictory figures resulting of shortcuts will diminish the scientific credibility of the institute. Such practice could also damage the reputation of the Recruits’ study itself in the long run.

The preceding remarks on the difficulties should in no way discourage researchers to use the possibilities of the programs. Their use is still time saving compared to a new construction of different scales from the scratch. The author is also happy to provide support in joint projects. On the CD with the database tnouv.ssd01 and the source code, partners in research will receive electronic copies of the book and of all papers in English or German. They explain concepts in detail and offer corresponding figures or percentages.

References

- American Psychiatric Association (APA) (Eds.) (1994). *Diagnostic and Statistical Manual of Mental Disorders DSM-IV*. (4th ed.), Washington DC.
- Cusson, M. (1998). *Criminologie actuelle* [Contemporary Criminology]. Paris: Presses universitaires de France.
- Escard, E., Haas, H. & Killias, M. (2003). Comportements suicidaires et violences: réflexions criminologiques à partir d’une étude portant sur 21’314 jeunes recrues suisses. [Suicidal Behavior and Violence: Criminological Reflections based on the study on 21,314 Swiss Recruits]. *Encéphale*. XXIX: 1-10.
- Gottfredson, M.R. & Hirschi, T. (1990). *A General Theory of Crime*. Stanford Cal: Stanford University Press.
- Haas, H., Farrington, D., Killias, M., & Sattar, G. (2004). The impact of different family configurations on delinquency: a detailed examination. *British Journal of Criminology*, 44: 1-13.
- Haas, H. (2003). *Psychologie de la déposition, victimologie et techniques d’entretien*. [Psychological Textbook on Witness’ Testimony, on Victims and the Forensic Interview]. Recherches juridiques lausannoises, Edition Schulthess.

- Haas, H. & Killias, M. (2003). The Versatility vs. Specialization Debate: Different Theories of Crime in the Light of a Swiss Birth Cohort. In: C. Britt & M. Gottfredson (Eds). *Control Theories of Crime and Delinquency* (Advances in Criminological Theory, Vol. 12), New Brunswick: Transaction Publ.
- Haas, H. (2001). *Agressions et victimisation: une enquête sur les délinquants violents et sexuels non détectés* [Violence and Victimization: A Study of Undetected Sexual and Violent Offenders]. Aarau, Switzerland; Sauerländer Verlag. ISBN 3-7941-4915-7.
- Haas, H. (1997). *Expériences de vie et comportements. Questionnaire complémentaire des Enquêtes fédérales auprès de la jeunesse et des recrues (ch-x)* [Life experience and behaviour. Complementary Questionnaire of the Swiss Federal Surveys of Adolescents and Recruits (ch-x)]. Berne, Switzerland.
- Hindelang, M., Hirschi, T. & Weis, J. (1981). *Measuring Delinquency*. London: Sage Publications.
- Killias, M. & Haas, H. (2002). The Role of Weapons in Violent Acts: Some Results of a Swiss National Cohort Study. *Journal of Interpersonal Violence* Vol. 17(1): 14-32.
- Killias, M. (2002). *Grundriss der Kriminologie*. [Textbook of Criminology] Bern: Staempfli Verlag.
- Killias, M. (2001). *Précis de Criminologie*. 2nd Ed. [Textbook of Criminology] Berne: Ed. Staempfli.
- Killias, M. (1997). *Expériences de vie et comportements. Questionnaire des Enquêtes fédérales auprès de la jeunesse et des recrues (ch-x)* [Life experience and behavior. Questionnaire of the Swiss Federal Surveys of Adolescents and Recruits (ch-x)]. Berne, Switzerland.
- Killias, M. (1991). *Précis de Criminologie*. 1st Ed. [Textbook of Criminology] Berne: Staempfli.
- Lamon, Ph. & Haas, H. (2003). Les incendiaires non détectés. [Undiscovered Arsonists]. *Revue internationale de criminologie et de police technique et scientifique*. 56(4): 451-466.
- Lessler J. & Kalsbeek W. (1992). *Nonsampling Error in Surveys*. New York: John Wiley & Sons.
- Regoeczi, W. & Riedel, M. (2003). The Application of Missing Data Estimation Models to the Problem of Unknown Victim/Offender Relationships in Homicide Cases. *Journal of Quantitative Criminology*, Vol. 19(2): 155-183.
- West D. & Farrington D. (1977). *The Delinquent Way of Life*. London: Heinemann.
- Wolfgang, M., Figlio, R., Sellin, T. (1972). *Delinquency in a Birth Cohort*. Chicago: University of Chicago Press

Address

Henriette Haas, PhD
 Lecturer of forensic psychology
 at the University of Zurich
 email: henriette.haas@access.unizh.ch
henriette.haas@unil.ch

Head of Division Analysis
 SAP / fedpol
 Bollingenstr. 56
 CH-3003 Berne
 Switzerland