What do I look for in a survey dataset as a (crime survey) data user?

Survey Practice Forum: Challenges and Opportunities for Social Survey
Data Collection in Scotland

UNIVERSITY *of* STIRLING

Ben Matthews | University of Stirling

Overview

- Why use surveys to measure crime?
- Constructing a victimization time trend
- Sample size and the crime drop
- Sampling frames and victimization

Why surveys?

Why crime surveys?

- How much crime is there?

- Actually quite hard to know

- We can just rely on data from the criminal justice system: we need triangulation across data sources due to the 'dark figure' of crime

- Identifying the gap between survey reports of victimization and police recorded crime was the key (and radical) innovation of crime surveys

Buil-Gil, D., Brunton-Smith, I., Pina-Sánchez, J., & Cernat, A. (2022). Comparing measurements of violent crime in local communities: a case study in Islington, London. Police Practice and Research, 23(4), 489–506. https://doi.org/10.1080/156 14263.2022.2047047

Surveys don't just happen

- Crime data triangulation is not a big problem "in countries where alternative forms of data beyond official criminal justice statistics are obtainable, such as victim surveys"
- But is "formidable particularly outside consolidated western democracies where the range of available data is often limited."

Surveys as knowledge infrastructure

- Knowing that data will be archived and available years from now (and that data from years ago is archived) is key to exploring change over time

- Established social surveys are part of the 'knowledge infrastructure' in a way that 'big data' or administrative data sources aren't always

BIG DATA,
LITTLE DATA,
NO DATA
SCHOLARSHIP IN THE NETWORKED WORLD

Christine L. Borgman

Constructing a time trend

Getting the most from established surveys

- Established surveys let you:
  - Examine change over time
  - Analyse small population groups after pooling sweeps together to increase sample size
  - Ideally, both
- But there are practical challenges in using these data
- And open questions about how best to analyse them

Figure 3.2: The proportion of adults experiencing crime fell by 10.4 percentage points since 2008/09 and 1.9 percentage points since 2019/20.

Proportion of adults experiencing any SCJS crime, 2008/09 to 2021/22.
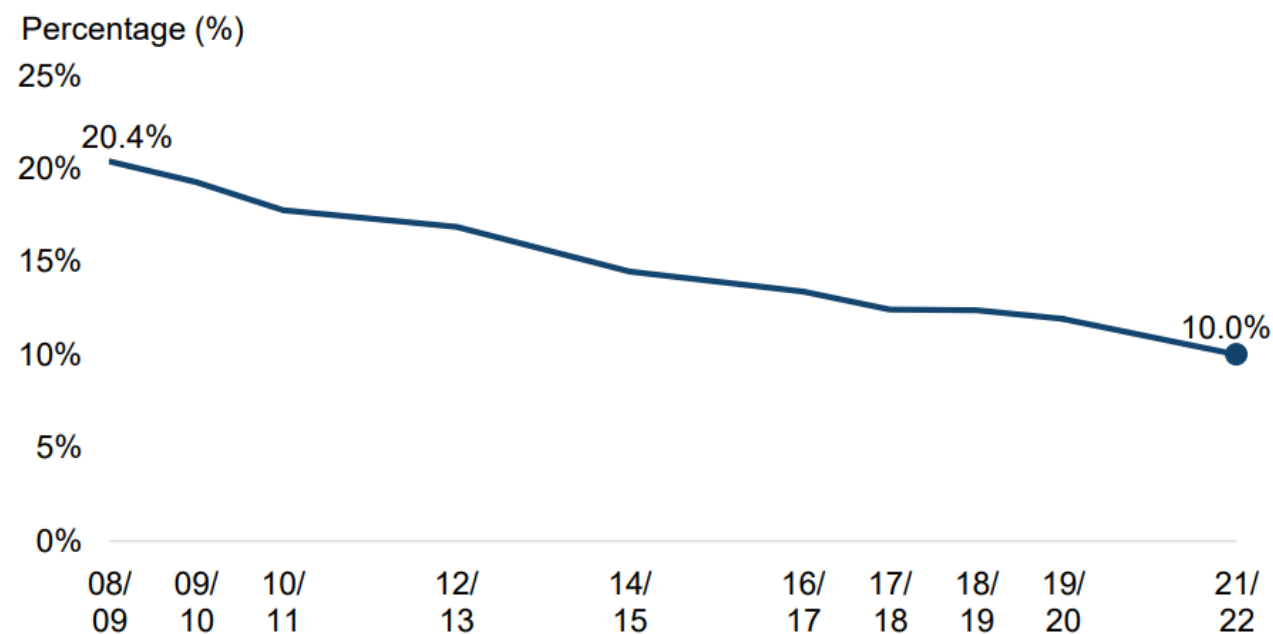
Base: SCJS 2008/09 (16,000); 2009/10 (16,040); 2010/11 (13,010); 2012/13 (12,050); 2014/15 (11,470); 2016/17 (5,570); 2017/18 (5,480); 2018/19 (5,540); 2019/20 (5,570); 2021/22 (5,520). Variable: PREVSURVEYCRIME.

Scottish Crime and Justice Survey

Access data

Access data

GN 33330
Scottish Crime and Justice Surveys, 1993-   ⌄

SN	Study description	Explore online	Select
9290	Scottish Crime and Justice Survey, 2021-2022		Login
8799	Scottish Crime and Justice Survey, 2019-2020		Login
8795	Scottish Crime and Justice Survey, 2018-2019	Nesstar	Login
8498	Scottish Crime and Justice Survey, 2017-2018	Nesstar	Login

sccjr-hackathon-2023 Public

Go to file | t | Add file ⌄ | <> Code ⌄

stuartnapier-sg added new file to combine data   56d4a30 · last year   🕑 31 Commits

📁 Handy resources	updates to variables in 4.2, added new data dictionary to ha...	last year
📄 .gitignore	updates to variables in 4.2, added new data dictionary to ha...	last year
📄 1. extract_datasets.R	Update 1. extract_datasets.R	last year
📄 2. check_datasets.R	modified that harmonise file slightly to save data in the data...	last year
📄 3. read_datasets.R	tidied up and mostly finalised data pipeline, including addin...	last year
📄 4.1 create_full_pooled_dataset.R	tidied up and mostly finalised data pipeline, including addin...	last year
📄 4.2 create_curated_pooled_dataset.R	updates to variables in 4.2, added new data dictionary to ha...	last year
📄 4.2.1 curated_dataset_create_binary_vars.r	tidied up and mostly finalised data pipeline, including addin...	last year
📄 5. create_full_sc_dataset.R	tidied up and mostly finalised data pipeline, including addin...	last year

```r
  mutate(ethnicity = case_when(
    ethnicity == "Minority Ethnic" ~ "Minority Ethnic",
    ethnicity == "White Other" ~ "White Other",
    is.na(ethnicity) ~ NA_character_,
    TRUE ~ "White Scottish/British"),
        ethnicity = fct_infreq(ethnicity))

scjs_combined <-
scjs_combined %>%
  mutate(simd_quint = factor(sjlabelled::as_character(simd_quint)),
         simd_top_15 = factor(sjlabelled::as_character(simd_top_15)))

scjs_combined |>
  count(survey_year, simd_top_15)


# export combined data ---------------------------------------------------

scjs_combined <-
scjs_combined |>
  select(-value, -year)


saveRDS(scjs_combined,
        here::here("01_data", "clean_data", "scjs_main_2008_09_21_22_explanatory_vars.rds"))
```

Pooled sample methodology

To overcome the issue of having too few respondents from minority ethnic backgrounds, a new approach was taken to bring together the results of multiple SCJS sweeps into one pooled sample. This would bolster the sample size both overall and for lesser represented groups too, meaning it would be possible to present findings for the survey by ethnicity for the first time. The methodology to implement this pooled sample is described below.

Firstly, each survey sweep results in a certain number of interviews being completed, and this varies from sweep to sweep. At the beginning of the SCJS, the sample size (n) target was to complete approximately 16,000 interviews, however in the more recent years this target has been roughly 6,000. Table Y shows the exact number of interviews completed across the 9 sweeps.

Selection for participation in the SCJS is random, however it is not completely random for the entire country as a whole. The design framework of the survey is to be representative at the level of the 13 Police Division areas in Scotland. This deviation from a true random sample is expressed in what is known as the 'design effect' (Deff). The exact value of the design factor used in each survey sweep is provided by the contractors who run the fieldwork process of the SCJS.

Using the design factor, we can calculate the 'effective sample size' of each survey. The effective sample size (neff) is the overall sample size divided by the design effect, and is an expression of the number of responses needed under a pure random sample to obtain the same results as the actual sample framework.

$$neff_i = \frac{n_i}{Deff_i}$$

Having done this, the proportion of the overall effective sample across the 9 sweeps can be calculated for each individual sweep, denoted as 'pool_neff' in Table B.2. This will be used later on when discussing weighting, and is an important step in ensuring that potential bias from uneven sample sizes across sweeps is reduced.

The only remaining step is to implement the pooled design effect on the weightings for the data. Survey data is weighted in order to ensure that the data better reflects what it is trying to measure. Each observation is assigned a weighted value to indicate how important it is to the analysis. The procedures for the implementation of the weighting methodology were developed by the Scottish Government with the Methodology Advisory Service at the Office for National Statistics (ONS), full detail on this can be found in the SCJS technical report. Multiplying the pooled effect sample by the weight results in the pooled weight for the survey which is then used throughout the analysis.

$$Pool\_scale\_weight\_INDIV_j = Pool\_neff_i \times WGTGINDIV\_scale_j$$

Table B.2: Information on SCJS sweeps to implement pooled methodology.

Survey year	2008-09	2009-10	2010-11	2012-13	2014-15	2016-17	2017-18	2018-19	2019-20
Sample size (n)	16003	16036	13010	12045	11472	5567	5475	5537	5568
Design Effect (Deff)	2.25	2.25	2.25	1.69	1.44	1.80	1.49	1.37	1.46
Effective sample size	7112	7127	5782	7127	7981	3100	3678	4045	3803
Pooled effective sample	0.143	0.143	0.116	0.143	0.160	0.062	0.074	0.081	0.076

Pooled sample methodology

To overcome the issue of having too few respondents from minority ethnic backgrounds, a new approach was taken to bring together the results of multiple SCJS sweeps into one pooled sample. This would bolster the sample size both overall and for lesser represented groups too, meaning it would be possible to present findings for the survey by ethnicity for the first time.

Limitations of pooled sample

The main limitation is that any sense of change over time is lost by combining survey years. The pooled sample is able to give a sense of the overall picture across all surveys, it is not able to tell whether there have been improvements or changes from one survey to another. Given some of the data from the pooled sample is over a decade old, it remains unclear whether the picture presented in the pooled sample is still representative of how it is today. It would be possible to perform further analysis yet and split the pooled sample data into two or more time periods, allowing for a 'before and after' analysis, but the analysis as presented in this report focusses only on one overall picture.

Scottish Government
Riaghaltas na h-Alba

Ethnicity in the Justice System

CRIME AND JUSTICE

social research
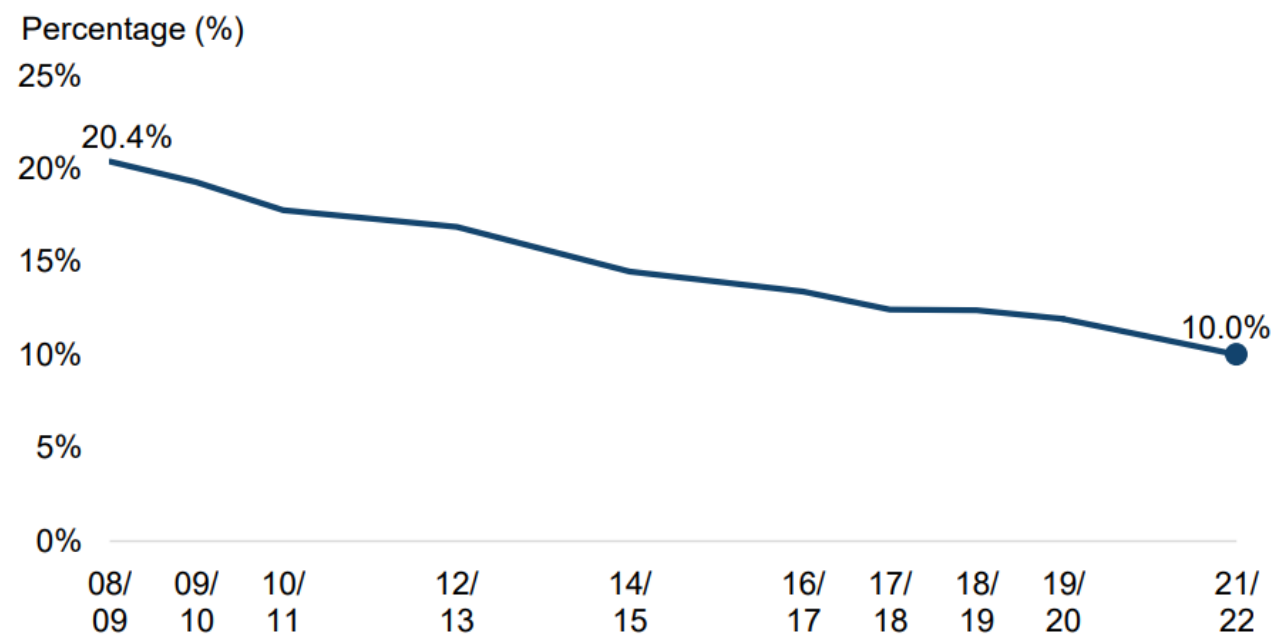
Figure 3.2: The proportion of adults experiencing crime fell by 10.4 percentage points since 2008/09 and 1.9 percentage points since 2019/20.

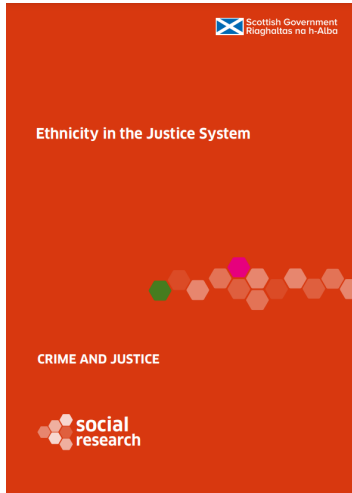Proportion of adults experiencing any SCJS crime, 2008/09 to 2021/22.

Base: SCJS 2008/09 (16,000); 2009/10 (16,040); 2010/11 (13,010); 2012/13 (12,050); 2014/15 (11,470); 2016/17 (5,570); 2017/18 (5,480); 2018/19 (5,540); 2019/20 (5,570); 2021/22 (5,520). Variable: PREVSURVEYCRIME.
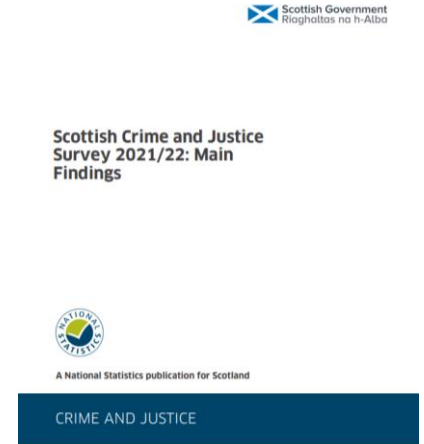
'Complete pooling'

'Partial pooling'

'No pooling'

Treats people from every sweep as if they are the same

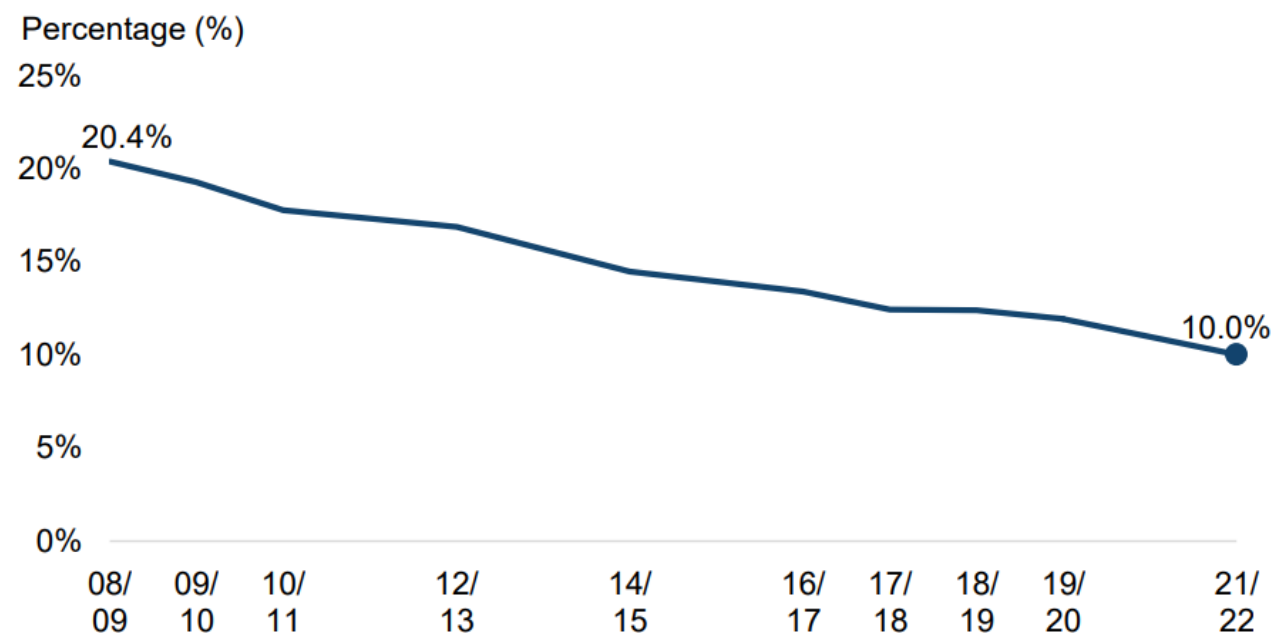Treats people from every sweep as if they are completely different

Sample size and the crime drop

Figure 3.2: The proportion of adults experiencing crime fell by 10.4 percentage points since 2008/09 and 1.9 percentage points since 2019/20.

Proportion of adults experiencing any SCJS crime, 2008/09 to 2021/22.

Base: SCJS 2008/09 (16,000); 2009/10 (16,040); 2010/11 (13,010); 2012/13 (12,050); 2014/15 (11,470); 2016/17 (5,570); 2017/18 (5,480); 2018/19 (5,540); 2019/20 (5,570); 2021/22 (5,520). Variable: PREVSURVEYCRIME.

Scottish Crime and Justice Survey 2021/22: Main Findings

A National Statistics publication for Scotland

CRIME AND JUSTICE

Sample size and the crime drop

	2008/09	2021/22	Change
Response rate	70.9%	47.3%	-23.6 %pt
Core achieved sample	16,003	5,516	-65%
Victimization prevalence	20.4%	10.0%	-51%
Number of victims interviewed*	2,786	532	- 81%
Number of repeat victims interviewed*	1,083	164	- 85%

Bell et al. (2024, January). *Scottish Crime and Justice Survey 2021/22 Technical Report*. https://www.gov.scot/binaries/content/documents/govscot/publications/statistics/2023/11/scottish-crime-justice-survey-2021-22-main-findings/documents/technical-report/technical-report/govscot%3Adocument/Crime%2BSurvey%2B%2B2021-22%2B%2BTechnical%2BReport.pdf

* Own calculations

Sample size and the crime drop

- There is an irony here: adults in general *don't* have much of an experience of crime – victimization prevalence as estimated by the SCJS was ~10% in the most recent sweeps
- So general population victimization surveys are less able to explore *victims*' experiences of crime unless they increase sample size over time
- This is even more true for repeat victims

Scottish Women's Aid. (2021, December). *Response to the consultation on the Scottish Crime and Justice Survey (SCJS), December 2021*. https://womensaid.scot/wp-content/uploads/2022/04/SWA-response-to-Scottish-Crime-and-Justice-Survey-consultation.pdf

Small numbers means more volatility

Investigating changes in the rate of violent crime is methodologically challenging. The repeat nature of domestic violence makes it hard to measure (Ellingworth *et al.* 1995 ; Farrell and Pease 2007 ; Farrell and Pease 2010), especially the experiences of those relatively few people who suffer large numbers of repeat offences. Small numbers of individuals who experience high frequencies of victimization are particularly problematic because of the volatility this can introduce when assessing changes over time. The conventional approach has been to ignore or 'cap' the frequency of victimizations of this group in order to avoid this problem. However, this is unsatisfactory because it omits a significant aspect of victimization. Thus, there appears to be a tension in the production of official crime statistics between two goals: accurately assessing the year on year changes in the rate of crime and accurately reporting the extent and distribution of violent crime victimization.

The crime drop and survey design

- There is a tension between the aims of providing 'a valid and reliable measure of adults' experiences of crime' and 'examin[ing] trends over time in the number and nature of crimes in Scotland'

- These are not specific to SCJS but maybe more acute than with e.g. Crime Survey for England and Wales due to smaller sample size

Sampling frame

The 'average' person

- Nationally representative sample surveys are designed to reflect the experience of the average person or household
- "The SCJS is a random probability sample survey of private households" using the Royal Mail Postcode Address File (PAF)
- SCJS aims to "provide a valid and reliable measure of adults' experience of crime"…

Saunders, A., Caldwell, K., Magnani, L., Cook, B., Page, L., Rose, J., Martin, C., & Hockaday, C. (2021). *Scottish Crime and Justice Survey 2019/20: Technical Report*. Scottish Government. p.13

Sampling frames and victimization

- … but if we want a measure of crime for all adults, not just all adults in private households, we need to assume that "the subset of the adult population not captured in the SCJS experience the same level of victimisation as adults in the household resident population"[1]

- But this excludes homeless people, students, people in prison, nomadic/itinerant populations and those living in refuges [2]

[1] Scottish Women's Aid. (2021, December). *Response to the consultation on the Scottish Crime and Justice Survey (SCJS), December 2021*. https://womensaid.scot/wp-content/uploads/2022/04/SWA-response-to-Scottish-Crime-and-Justice-Survey-consultation.pdf

[2] Carr-Hill, R. (2013). Missing Millions and Measuring Development Progress. *World Development*, *46*, 30–44. https://doi.org/10.1016/j.worlddev.2012.12.017

Sampling frames and victimization

- If the outcome you're interested in (victimization) affects selection into the survey – like in our examples – this is not a 'problem' that can be solved' other than some specific scenarios
- Adults in private households and all adults are just different target populations – neither is necessarily right or wrong for a given application

Schuessler, J., & Selb, P. (2023). Graphical Causal Models for Survey Inference. *Sociological Methods & Research*, 00491241231176851. https://doi.org/10.1177/00491241231176851

Thank you

UNIVERSITY *of* STIRLING

ben.matthews@stir.ac.uk