

# Data Codebook: Open Science Practices in Criminology and Legal Psychology

Michael Beck

2025-09-08

## Table of contents

<b>1</b>	<b>Codebook</b>	<b>1</b>
1.1	Summary Statistics . . . . .	1
1.2	Sample . . . . .	3
1.3	Variable Description . . . . .	4
1.4	Missing Values . . . . .	4

## 1 Codebook

A comprehensive dataset of academic articles from the category of Criminology and Legal Psychology, analyzed for open science practices, statistical methods, and journal characteristics. This dataset includes metadata from Crossref, journal information from Clarivate, open access status from Unpaywall, and derived variables indicating statistical analysis usage and open science practices (open data, open materials, preregistration) that were generated using machine learning. The dataset spans multiple disciplines and provides insights into the adoption of open science practices in academic publishing.

### 1.1 Summary Statistics

**Dataset Path:** data/sample\_analysis.qs2

**Dataset dimensions:** 3997 observations x 26 variables

**Data collection period:** 2013 - 2023

Table 1: Data summary

Name	df
Number of rows	3997
Number of columns	26

Column type frequency:

character	8
Date	1
factor	9
numeric	8
<hr/>	
Group variables	None

**Variable type: character**

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
doi	0	1.0	16	34	0	3997	0
title	0	1.0	5	653	0	3997	0
journal_issn	0	1.0	9	9	0	81	0
abstract	2405	0.4	59	5781	0	1592	0
journal_eissn	0	1.0	9	9	0	78	0
journal_name_std	0	1.0	4	79	0	78	0
journal_publisher	0	1.0	3	14	0	17	0
journal_name	0	1.0	4	79	0	78	0

**Variable type: Date**

skim_variable	n_missing	complete_rate	min	max	median	n_unique
published_date	0	1	2013-02-01	2023-12-30	2019-02-01	511

**Variable type: factor**

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
is_statistical	0	1.00	FALSE	2	No: 2234, Yes: 1763
is_open_access	0	1.00	FALSE	2	No: 2456, Yes: 1541
is_open_materials	2234	0.44	FALSE	2	No: 1688, Yes: 75
is_prereg	2234	0.44	FALSE	2	No: 1718, Yes: 45
is_open_data	2234	0.44	FALSE	2	No: 1725, Yes: 38
txt_only_abstract	0	1.00	FALSE	2	No: 3453, Yes: 544
txt_source	0	1.00	FALSE	3	TXT: 3254, PDF: 407, HTM: 336
journal_category	0	1.00	FALSE	3	PSY: 1986, LAW: 1618, CRI: 393

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
journal_jif_quartile	0	1.00	FALSE	3	Q1: 3811, Q2: 174, Q3: 12

### Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
txt_count	0	1	8702.57	6745.20	263.00	3353.00	8105.00	12355.00	65046.00	
txt_flesch	0	1	25.96	12.34	-47.94	18.91	26.78	34.35	64.48	
published_year	0	1	2018.49	3.01	2013.00	2016.00	2019.00	2021.00	2023.00	
published_month	0	1	6.61	3.59	1.00	3.00	7.00	10.00	12.00	
published_day	0	1	2.90	5.57	1.00	1.00	1.00	1.00	31.00	
journal_total_citations	0	1	13986.37	18229.13	85.00	998.00	2921.00	17454.00	57108.00	
journal_x2023_jif	0	1	4.70	3.65	0.80	1.80	3.20	6.10	18.20	
journal_x2023_jci	0	1	2.35	0.91	1.45	1.72	2.08	2.92	7.27	

## 1.2 Sample

The sample consists of academic articles from the fields of Criminology and Legal Psychology. The articles were filtered using keywords to indicate whether the articles employed specific statistical methods or adhered to open science practices.

Table 6: Filtering steps and case drops for the generation of the sample used in the analyses.

Step	Filtering Step	Before	After	dropped	Filtering Logic
1	Remove duplicate DOIs	95042	55904	39138	~!duplicated(doi)
2	Remove missing publication dates	55904	55904	0	~!is.na(published_date)
3	drop published_date > date_from	55904	47309	8595	~published_date > date_from
4	drop published_date < date_to	47309	45922	1387	~published_date < date_to
5	Exclude using keywords for simple reviews	45922	45557	365	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))
6	Exclude using keywords for corrections and errata	45557	45144	413	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))
7	Exclude using keywords for front/back matter	45144	42320	2824	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))
8	Exclude using keywords for announcements	42320	42279	41	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))
9	Exclude using keywords for comprehensive reviews	42279	41908	371	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))
10	Exclude using keywords for other non-research content	41908	40860	1048	~!str_detect(title, regex(str_c(kws, collapse = " "), ignore_case = TRUE))

This was used to generate a sample. The initial sample

Table 7: Filtering steps and case drops for the generation of the analytical statistical inference sample

Step	Filtering Step	Before	After	dropped	Filtering Logic
1	Filter for publications that could be downloaded (from Ddwnloader metadata)	4265	4066	199	<code>~status == "Done"</code>
2	Filter by minimum full text length (> 1000 characters)	4066	3997	69	<code>~nchar(fulltext) &gt; 1000</code>

The sample itself consists of articles that were identified as using specific statistical methods or adhering to open science practices. Most analysis are conducted on the statistical subset of the data as reported in the analyses contexts.

### 1.3 Variable Description

Variable	Description	Source
<code>doi</code>	Digital Object Identifier for the article.	Crossref Metadata.
<code>title</code>	Article title.	Crossref Metadata.
<code>abstract</code>	Abstract text.	Crossref Metadata.
<code>published_date</code>	Publication date (published_print, if not available published_online).	Crossref Metadata, partially imputed.
<code>published_day</code>	Publication day (1-31).	Derived from published_date.
<code>published_month</code>	Publication month (1-12).	Derived from published_date.
<code>published_year</code>	Publication year (YYYY).	Derived from published_date.
<code>journal_category</code>	High-level journal subject category.	Clarivate Journal Metadata
<code>journal_eissn</code>	Journal electronic ISSN.	Crossref Metadata.
<code>journal_issn</code>	Journal print ISSN.	Crossref Metadata.
<code>journal_jif_quartile</code>	Journal JIF quartile (Q1-Q4).	Clarivate Journal Metadata
<code>journal_name</code>	Original journal name.	Crossref Metadata.
<code>journal_name_std</code>	Standardized journal name (cleaned/normalized).	Crossref Metadata.
<code>journal_publisher</code>	Journal publisher name.	Parsing of URLs received by the dx.doi.org redirection.
<code>journal_total_citations</code>	Total citations to the journal (year 2023).	Clarivate Journal Metadata
<code>journal_x2023_jci</code>	Journal Citation Indicator (2023).	Clarivate Journal Metadata
<code>journal_x2023_jif</code>	Journal Impact Factor (2023).	Clarivate Journal Metadata
<code>txt_count</code>	Total word count of processed text.	Derived during text processing
<code>txt_flesch</code>	Flesch Reading Ease Score.	Derived during text processing
<code>txt_only_abstract</code>	Text available only from abstract (not full text).	Derived during text processing
<code>txt_source</code>	Source of text used for analysis.	Derived during text processing
<code>is_statistical</code>	Dummy: study uses statistical analysis.	Derived from full text.
<code>is_open_access</code>	Dummy: article is Open Access.	Derived from Unpaywall API.
<code>is_open_data</code>	Dummy: open data available.	Derived from full text, only available for statistical inference publications.
<code>is_open_materials</code>	Dummy: open materials available.	Derived from full text, only available for statistical inference publications.
<code>is_prereg</code>	Dummy: study preregistered.	Derived from full text, only available for statistical inference publications.

### 1.4 Missing Values

All missing values are coded as `NA`. Missing values in the variables `is_open_data`, `is_open_materials`, `is_prereg` are produced due to the fact that these variables are only applicable to a subset of the data (studies that are classified as “statistical”). As a result, any non-statistical studies will have missing values for these variables. Missings in `abstract` result from non-availability in the crossref metadata for the specific publication.