scientific data

OPEN

DATA DESCRIPTOR

Generational Ideology Tables of the Spanish Population

Cristina Aybar, Virgilio Pérez & Jose M. Pavía

Studying ideological preferences is essential for understanding the social and political changes a society undergoes over time. Generational ideology tables provide a structured framework to analyze ideological shifts within and across generations, offering valuable insights into societal evolution and political behavior. This paper introduces the Spanish Cohort Ideology Database (SCID), detailing the data sources and methodology used for its construction. By leveraging more than five million individual observations (which expand to over 100 million when employing double 5 × 5 moving windows) from over 1,800 surveys conducted since 1977 by the Spanish official center for sociological research (CIS), we construct 1,554 period and cohort ideology tables, including breakdowns by gender, education level, gender and education level, and region. SCID comprises age-year and age-generation tables with mean values, sample sizes, and variances, enabling the analysis of the dispersion/polarization of ideological self-placement. This work facilitates the analysis of social and political change processes from a cross-section and longitudinal perspective, creating a unique database that could also be developed in other countries, thereby enabling international comparative studies.

Background & Summary

The study of a population's ideological preferences is essential for understanding the social and political changes a society undergoes over time. These preferences not only reflect individual political leanings but are also deeply influenced by generational, demographic, and socioeconomic factors. According to Mannheim impressionable years' theory[1], the political identity of each person is built during their young age, and it lasts almost until the end of their life. Mannheim's thinks that the older we are, the more resistant to change we are. Similarly, according to social identity theory, the stronger our ideological alignment with a particular set of beliefs, the less likely we are to abandon them and the more likely we are to identify ourselves with the group (political party) that represents those beliefs[2].

As individuals adopt a specific ideological stance, they tend to gravitate toward political parties that align with their views, reinforcing their sense of belonging and loyalty to the party[3]. Furthermore, in a transitive way, political parties can actively foster social identification with the party and encourage the use of partisan categories among their rank and file, which can, in turn, create a more partisan and engaged group of supporters. This alignment can result in stronger support for the party's policies, candidates, and actions, as the party becomes an extension of their own identity[2].

By analogy with generational life tables, which make it possible to study longitudinal survival patterns and cohort-specific mortality trends over time, generational ideology tables enable studying, among other issues, how ideology of different cohorts evolve over time and how particular events impact on population, even testing Mannheim's and social identity theory's hypothesis.

In particular, age-year and age-generation ideology tables can be used to (i) tracking cohort-specific ideological trends; (ii) understanding and assessing the impact of historical events; (iii) comparing ideological trajectories across generations; (iv) evaluating the longevity of ideological alignments; (v) forecasting political and social trends; (vi) understanding interactions between demographics and ideology; and, (vii) guiding policy and political communication strategies. More specifically, period and cohort ideology tables allow researchers, political strategists, and policymakers to:

(i) Follow the evolution of political or ideological preferences within a particular generational cohort over their lifespan, helping to identify whether certain ideologies gain or lose influence as the cohort ages or as historical events shape their views.
(ii) Illustrate how key events (e.g., economic crises, wars, social movements, technological shifts) influence

GIPEyOP, Universitat de Valencia, Av Tarongers, s/n, Valencia, Spain. ✉e-mail: pavia@uv.es

ideological preferences within specific cohorts.

(iii) Identify whether certain ideologies are "generationally rooted" or influenced by age (life cycle effects) and period (historical context) factors.

(iv) Show how long a cohort remains aligned with a particular ideology, shedding light on ideological persistence or shifts over decades.

(v) Predict future ideological landscapes by examining the behavior of younger cohorts over time, helping political parties and policymakers anticipate long-term electoral trends.

(vi) Incorporate other demographic factors like education, income, gender, and ethnicity to explore how ideological preferences intersect with social identities over time.

(vii) Tailor strategies for specific generational groups, recognizing their unique ideological concerns and shifts.

In short, age-year and age-generation ideology tables would provide a structured framework to study the temporal and cross-section dynamics of ideological shifts within and across generations, contributing to a deeper understanding of societal evolution and political behavior, including shedding light to address long-controversial questions about the nature of partisan stability[2].

In Spain, the Centro de Investigaciones Sociológicas (CIS, the Spanish official center for sociological studies) has systematically collected data for decades to support the scientific study of Spanish society. The result is a unique and largely underutilized database that provides invaluable insights into the long-term political and sociological attitudes of Spaniards. Since the restoration of democracy in Spain, it also includes data on the ideological self-placement of the Spanish voting-age population (18 years and older) starting in 1977, enabling the study of their ideological trends.

Ideological self-placement, typically measured using political orientation scales ranging from far-left to far-right, is a highly useful tool for studying the evolution of political behavior. However, its comparison across time series presents methodological challenges, as survey designs and the scales used may vary over time. This is the case for the surveys conducted by CIS, which, over a period of nearly 60 years, has employed up to 28 different scales—though, starting in 1982, the overwhelming majority of the surveys have used the single scale 1–10[4]. This article addresses these limitations by systematizing and harmonizing CIS microdata to produce a set of cohort-age and year-age ideological tables, enabling the study of the evolution and distribution of Spaniards' ideological preferences over the past 50 years.

In this research, we have developed an approach based on combining all the results from CIS surveys, which include over five million basic observations from more than 1,800 studies with available microdata (80% of which have a sample size greater than 1,000, with some studies reaching $N \approx 30{,}000$), conducted since 1977, and an average of 156 variables observed in each study.

From this corpus of data, we have created the Spanish Cohort Ideology Database (SCID) after selecting the following variables: (i) YEAR (year of the survey); (ii) AGE (age of the respondent at the time of the survey); (iii) IDEOLOGY (ideological self-placement of the respondent); (iv) GENDER; (v) CCAA (autonomous community or region of residence of the respondent); and (vi) EDUCATION (highest level of education attained by the respondent). The variables AGE and YEAR serve as the foundation for constructing the age-year and age-cohort ideology tables that integrate SCID. The variable IDEOLOGY, which spans various ranges, has been harmonized to a 1–10 scale to ensure comparability across the different scales used in the surveys. These three main variables were used to construct tables that summarize the evolution of Spaniards' ideological preferences over time and across different generational cohorts.

In addition to tables for the entire population, SCID also includes breakdowns by sociodemographic characteristics of the respondents, enabling the exploration of demographic and socioeconomic variations in ideological preferences. Specifically, SCID contains separate tables by gender, in two categories: 'Male' and 'Female'; by education level, in five categories: 'No education', 'Primary education', 'Secondary education', 'Vocational training', and 'Higher education'; by gender and education level, in ten categories (derived from systematically combining the two gender categories with the five education categories); and by region of residence (the 17 autonomous communities and 2 autonomous cities of Spain).

When segmenting variables—such as education or region—are considered, sample sizes shrink, increasing the variability of estimates. Therefore, to improve the quality and robustness of the estimates, we follow standard practice and leverage the expected similarity of adjacent observations to significantly reduce variance at the cost of slightly increasing bias[5]. This is a widely recognized approach that is employed in all branches of statistics and data science to effectively reduce overall error (measured as the sum of variance and squared bias). In particular, we have used the technique of double moving windows (multivariate rolling windows), which increases the sample size in the estimation of a value by integrating data from adjacent periods and ages. Using this methodology, additional tables were generated with different AGE-YEAR window configurations, such as 1 × 3, 3 × 1, 3 × 3, 1 × 5, 5 × 1, and 5 × 5.

We have generated three types of tables with each window configuration: (i) a table of average values that shows the average ideological evolution for each combination of AGE and YEAR; (ii) a table with the counts of observations used to calculate the means; and (iii) a variance table, which allows for the analysis of the dispersion and polarization of ideological self-placement. This approach was replicated across breakdowns by gender, education level, gender and education level, and autonomous community, generating a broad and detailed set of 777 age-year (period) tables. The information contained in these set of tables was also reorganized by AGE and COHORT (year of birth), creating a new set of 777 additional age-generation (cohort) tables. The cohort tables are the diagonals of the period tables, resulting in the final set of 1,554 tables that currently integrate SCID. Additionally, to make it easier to combine the data from different tables, all the data provided in the tables has

		# OBSERVATIONS (ROWS)				# VARIABLES (COLUMNS)		
TYPE STUDY	# STUDIES	MIN	MAX	MEAN	# OBS	MIN	MAX	MEAN
BAR	464	850	17,650	2,783.64	1,291,606	50	496	150.93
ELE	483	199	29,201	2,731.69	1,319,404	30	437	113.30
OTE	559	154	27,433	3,044.69	1,701,982	13	1397	195.08
POL	194	499	22,265	2,216.75	430,049	20	465	133.52
ICC	154	1,000	3,313	1,982.86	305,360	164	262	195.48
TOTAL	1854	154	29,201	2,722.98	5,048,401	13	1397	156.31

Table 1. Descriptive summary of the CIS studies included in the analysis: number of studies, volume of processed microdata, and average number of observations per study type.

been organized in a unique dataset in long-table format. Together, these tables allow for the exploration of ideological preferences from multiple perspectives.

This article aims to provide a methodological and empirical resource for the study of ideological trends in Spain, which can be applied to any other democracy. It offers synthesized and harmonized data that allows the exploration of how age, time, and generational factors interact in shaping ideological preferences. Ultimately, this work seeks to contribute to the analysis of social and political change processes in Spain from a longitudinal and generational perspective.

Methods

This study is based on a thorough process of data acquisition, cleaning, and harmonization from the CIS database (https://www.cis.es/en/estudios/catalogo-estudios), which includes 1,854 studies with available microdata at the time of writing. These studies encompass monthly barometers (BAR), electoral studies (ELE), political studies (POL), consumer confidence index (ICC) surveys, and studies on various other topics (OTE). Table 1 presents summary statistics that highlight the scope of the database, such as the total number of studies, the volume of individual processed microdata, and the average number of observations and variables per study type. Meanwhile, Fig. 1 depicts the temporal distribution of these studies, highlighting the number of responses collected in each survey.

The surveys conducted by the CIS are a benchmark for opinion studies in Spain[6–9], regularly and continuously collecting information on various topics, including: (i) basic sociodemographic variables; (ii) evaluation of the economic situation; (iii) political variables; (iv) personal attitudes; (v) education level; (vi) marital status and living situation; and (vii) labor and socioeconomic variables, along with a set of specific variables related to a current issue at the time of the survey. The number of CIS studies with available microdata continues to grow, as, in addition to new information being collected each year, the CIS is gradually expanding the database by incorporating microdata from surveys conducted in the 1980s and 1990s. As a result, the CIS database stands as the most comprehensive and extensive survey repository in Spain and one of the most valuable among those found in Western democracies[10].

Managing a dataset of such dimensions is no small task. It requires systematic work and process automation. In the specific case of the data provided by the CIS, we found it appropriate to create a thematic map, using text mining techniques[11] to extract key words[12,13] and natural language processing to establish meanings[14], with a twofold objective. First, to identify questions of interest within each study, and second, to link questions and topics across studies/surveys. The outcome of this process has been a dictionary that makes it easier to navigate between studies and questions, thus enabling tailored analyses. This tool, the result of extensive work in identifying topics and variables, opens up a wide range of analytical possibilities, such as the construction of cohort-generational ideological tables, which we address in this paper, significantly reducing the costs of data retrieval and processing.

The steps followed to obtain the relevant raw database and metadata are as follows:

(i) Download the microdata folder (ZIP file) of each study from the CIS website (https://www.cis.es/en/estudios/catalogo-estudios). Each folder contains several files, including the questionnaire and study code (PDF format), an ASCII file, an SPSS syntax file (or, more recently, an SAV file), and a description file of the interviewer's cards. Each download must be done manually and requires completing an individual form for each study with the downloader's identifying information and purpose.
(ii) Convert ASCII and syntax files into a SAV file, if necessary.
(iii) Convert the SAV file into a CSV file.
(iv) Identify relevant variables based on the questionnaire.
(v) Add two columns to each study indicating the year and month of its completion.
(vi) Transfer the variable identification row into a database file, creating a 'dictionary' that facilitates navigation between studies and questions, allowing tailored inquiries to be addressed.

The above-detailed process presented various challenges, such as variations in the number of questions across surveys, substantial differences in the formulation of questions and coding of variables, and inconsistencies in their positions within the datasets. After a thorough analysis and identification of common questions, the following variables of interest were selected: (i) YEAR (year the survey was conducted); (ii) AGE (respondent's age at the time of the survey); (iii) IDEOLOGY (ideological self-placement); (iv) GENDER (sex); (v) CCAA (region

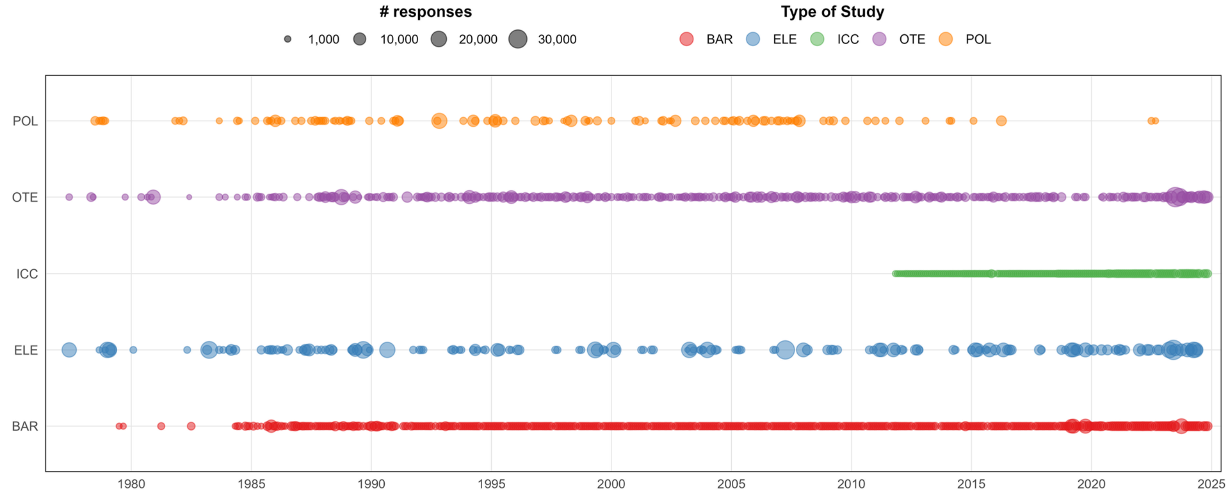

Fig. 1 Temporal distribution of the number of studies and responses processed, categorized by study type, for the CIS surveys included in the analysis.

of residence); and (vi) EDUCATION (highest level of education achieved). The COHORT variable is derived by subtracting the AGE variable from the YEAR variable. Once the variables of interest were selected, their possible values were analyzed to standardize response options and detect errors. This process helped address inconsistencies, ranging from issues in variable identification to corrupted or miscoded files, with the situation being reported to the CIS when necessary.

As an alternative to following this process, interested readers may use a new tool—called 'FICHERO INTEGRADO DE DATOS (FID)'—which has been made available and is being developed by CIS. This tool facilitates the extraction of time series data for selected variables from their database. However, compared to our approach, the use of the FID service presents some relevant limitations. First, extractions can only be performed from their available collections since 1985, with only the BAR and ICC collections fully accessible (since 1985), and only a part of the OTE and POL studies currently available. In particular, no studies from the ELE group, which includes all pre-electoral and post-electoral surveys conducted by CIS, are available. Second, extractions must be performed independently from each of the ten collections in which CIS has classified its available FID studies. In contrast, using our dictionary, one can select all studies containing a specific variable (not only from the collections available in FID) or all studies conducted at a specific point in time.

The variable measuring the ideological self-placement of the respondent on the left-right scale (IDEOLOGY) not only offers a wide range of possible values, both numerical and textual, but also presents an additional complication. It has been formulated in CIS surveys using 28 different ways over time, employing different scales (1–5, 1–6, 1–7, 1–10, and 0–10) and using various labels[4]. Therefore, a rescaling process was necessary to unify all studies into a single scale. The 1–10 scale was chosen due to its widespread use, as more than 90% of the CIS studies that ask about ideological self-placement employ this scale. Indeed, starting in mid-1982, 94.84% of the studies and 96.09% of responses used the 1–10 scale. Prior to that, between 1977 and mid-1982, CIS almost exclusively employed the 1–7 scale. In fact, all but one survey conducted during that period used the 1–7 scale. To standardize this variable, we have used the following equation[4]:

$$x_k = x_{k-1} + \frac{9}{n - 1}$$

where $x_1 = 1$, k takes values from 1 to 10, with n being the number of options in the original scale.

After discarding observations with missing values for the IDEOLOGY variable, the complete raw dataset, combining all the surveys, contained a total of 3,617,541 observations, ranging from 1977 to 2024. Subsequently, the primary variables used to construct the tables (AGE and YEAR) were reviewed, discarding observations with AGE in intervals, as well as missing values (NA, text, and values below 18). After this additional cleaning, the dataset was reduced to a total of 3,540,099 observations.

The large number of surveys considered in this study highlighted significant challenges in processing the classification variables (GENDER, EDUCATION, and CCAA), particularly EDUCATION, which contained up to 907 possible levels. To address this complexity, these levels were grouped into five categories: 'No schooling completed', 'Primary education', 'Secondary education', 'Vocational training' and 'Higher education'. Table 2 provides a summary of the absolute and relative basic frequencies for the categories of the GENDER, EDUCATION, and CCAA variables, offering a clear view of their distribution across the depurated dataset.

With all the selected data, a unique dataset in long-table format was constructed, incorporating all observations from almost 1,900 studies. From this dataset, and combining the variables AGE, YEAR, and IDEOLOGY, three matrices were generated for each AGE-YEAR combination: one recording the number of respondents, another containing the average value of the IDEOLOGY variable, and a third showing the variance of this

	Levels of the variable	# of basic sample sizes	
GENDER	Man Woman	1,809,718 1,728,057	(0.5115) (0.4885)
EDUCATION	No schooling completed Primary education Secondary education Vocational training Higher education EACH LEVEL SHOULD BE PRESENTED IN A SEPARATE ROW. THE LEVELS ARE: (i) No schooling completed; (ii) Primary education; (iii) Secondary education; (iv) Vocational training; and, (v) Higher education	230,076 709,721 1,017,967 593,466 778,329	(0.0691) (0.2132) (0.3057) (0.1782) (0.2338)
CCAA (region)	Andalusia (01) Aragon (02) Asturias (03) Balearic Islands (04) Canary Islands (05) Cantabria (06) Castile and Leon (07) Castile–La Mancha (08) Catalonia (09) Valencian region (10) Extremadura (11) Galicia (12) Madrid (13) Region of Murcia (14) Navarre (15) Basque Country (16) La Rioja (17) Ceuta (18) Melilla (19)	499,999 113,320 88,313 73,752 120,730 55,138 218,367 149,217 500,094 325,088 96,331 277,197 375,832 93,154 55,194 203,018 40,949 8,059 7,934	(0.1514) (0.0343) (0.0267) (0.0223) (0.0366) (0.0167) (0.0661) (0.0452) (0.1515) (0.0985) (0.0292) (0.0840) (0.1138) (0.0282) (0.0167) (0.0615) (0.0124) (0.0024) (0.0024)

Table 2. Basic absolute and relative sample sizes for the classification variables: GENDER, EDUCATION, and CCAA (region).

variable. These matrices provide an initial representation of the distribution of ideological preferences over time and across age groups.

Figure 2 illustrates the number of observations for each AGE-YEAR pair. For the entire dataset, the average sample size per AGE-YEAR pair is 920.22 observations. However, the sample size matrix reveals a significant reduction for ages over 80, where the average number of observations drops to 120.35, and even further during the first half of the study period (up to the year 2000), with an average of 51.63 observations.

To increase the number of observations on which each estimate is based on for each AGE-YEAR pair, the double moving window technique (multivariate rolling windows) was applied, based on the assumption that ideological preferences evolve gradually across these two dimensions[15–17]. This technique assumes that changes in ideological preferences are relatively gradual over time and across adjacent ages/cohorts. In other words, it is assumed that, on average, the ideology of a person of age A in year Y is closer to that of a person of the same age in, say, years $Y-1$ and $Y+1$ or to that of a person of age, say, $A-1$ and $A+1$ in year Y than to that of a person of the same age in, say, year $Y+10$ or to that of a person of age, say, $A+10$ in the same year. For example, to calculate the average value of the IDEOLOGY variable for an age A in a year Y, data from adjacent years (in the case of the 1×3 window, $Y-1$ and $Y+1$) or nearby ages (in the case of the 3×1 window, $A-1$ and $A+1$) are included.

Six additional window configurations were defined: 1×3, 3×1, 3×3, 1×5, 5×1, and 5×5, where the first number indicates the range of ages considered, and the second specifies the range of years. Note that a 1×1 window configuration is equivalent to not using rolling windows. For example, in a 1×3 window, responses from a fixed age in a given year were combined with those from the preceding and following years. Hence, for instance, for the 1×3 and 3×1 configurations, the procedure triples the sample size for each age-year combination while preserving consistency in average values. Similarly, the 5×5 configuration multiplies the sample size by a factor of 25.

This approach not only increases the granularity of the data but also enhances statistical reliability by reducing the influence of outliers or small sample sizes with almost no impact on bias due to the specific window configurations employed. These configurations were chosen to strike a balance between reducing variance and limiting bias, while being symmetrical in terms of age, time, or both age and time. Smaller windows (e.g., 1×3 or 3×1) provide localized smoothing while preserving temporal or age-specific granularity, whereas larger windows (e.g., 5×5) further reduce variance but at the cost of potential bias due to broader averaging. Figure 3 provides a graphical representation (heatmap) of the sample sizes for each of the six additional proposed configurations.

Once the tables incorporating all available information were created, additional tables were generated, disaggregated by gender (using the response options for the variable GENDER: 'Man' and 'Woman'), by educational attainment (using the categories: 'No schooling completed', 'Primary education', 'Secondary education', 'Vocational training' and 'Higher education'), and by autonomous community (covering all 19 autonomous regions in Spain).

A total of 777 period (age-year) tables were constructed, distributed as follows: 259 containing average values, 259 recording variances, and 259 displaying the number of observations (effective sample sizes). This total results from combining the seven window configurations (1×1, 1×3, 3×1, 3×3, 1×5, 5×1, and 5×5) with 37 possible categories (overall totals and breakdowns by gender, educational attainment, gender and educational attainment, and autonomous community). Additionally, and to make it easier to navigate in the data from a cohort-based perspective, 777 generational (age-cohort) tables were constructed by reorganizing the data from

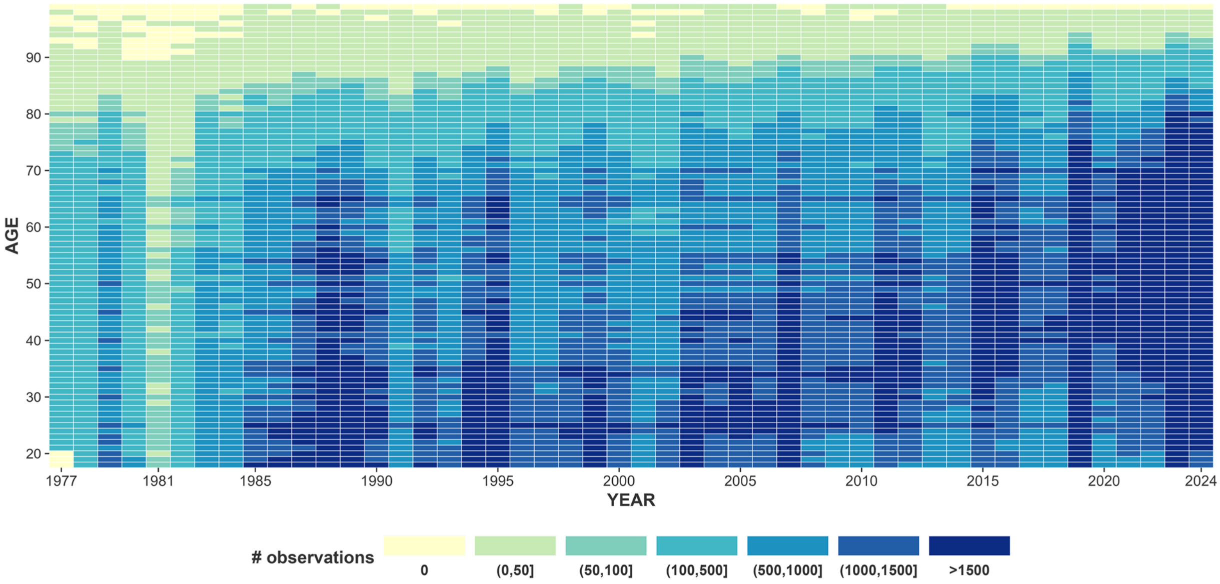

Fig. 2 Heatmap of the sample size for each AGE-YEAR combination.

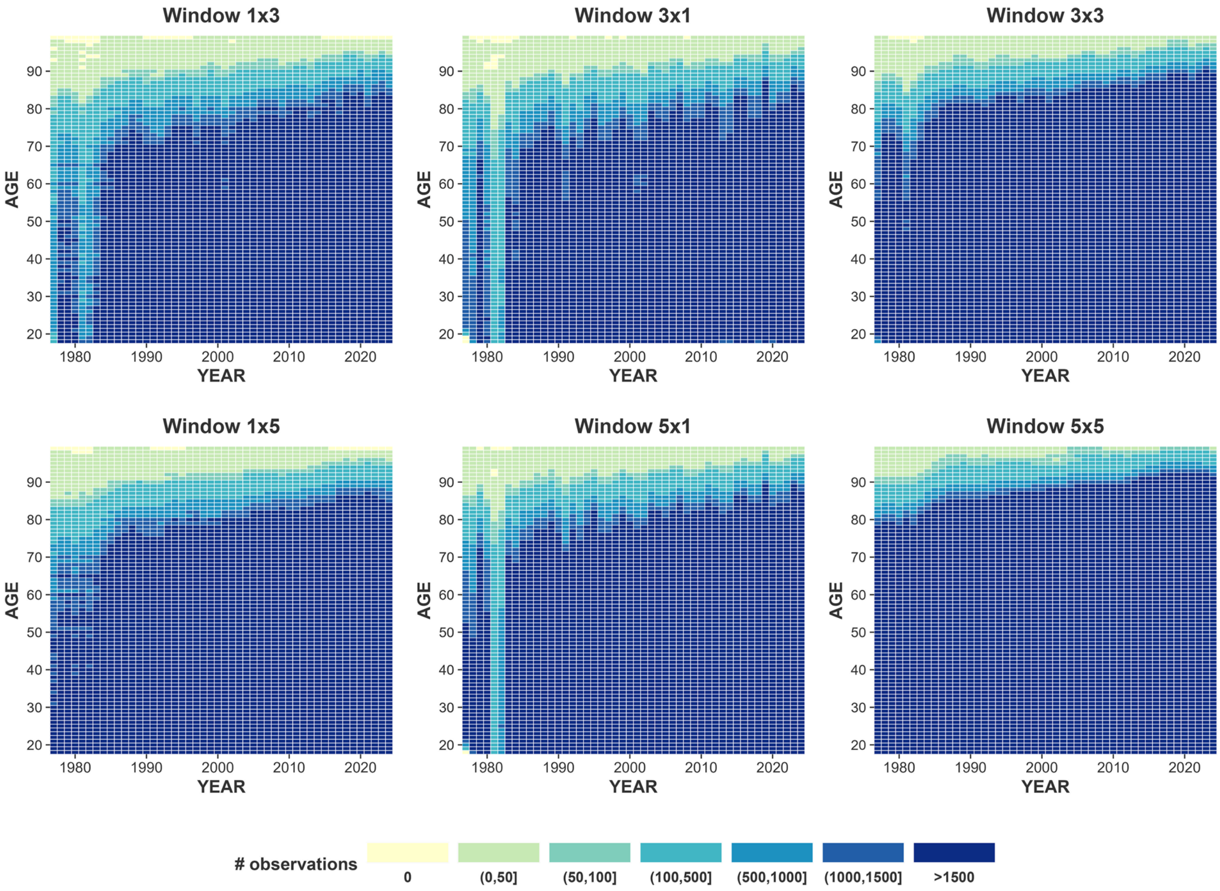

Fig. 3 Heatmap of sample sizes for each AGE-YEAR combination across the six rolling window configurations considered: 1 × 3, 3 × 1, 3 × 3, 1 × 5, 5 × 1, and 5 × 5.

the 777 period tables based on the respondent's year of birth. Furthermore, all the data provided in the tables has been organized in a unique dataset in long-table format. This methodological approach enabled the creation of a harmonized set of period and cohort ideology tables that capture ideological trends in Spain from

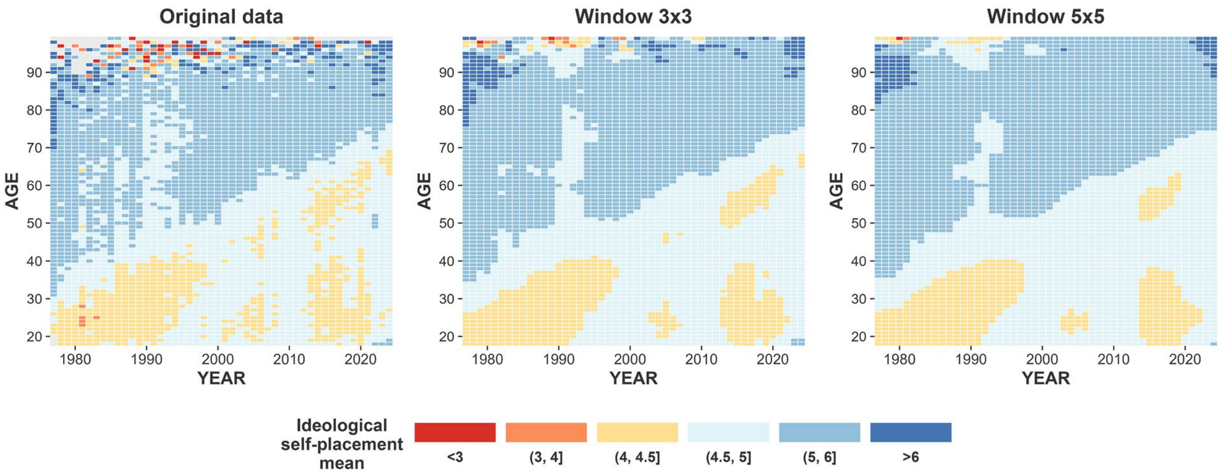

Fig. 4 Heatmap of the average ideological self-placement for each AGE-YEAR pair. Basic estimates and smooth estimates generated with 3 × 3 and 5 × 5 rolling windows.

a cross-section and longitudinal perspective, integrating temporal evolution with generational and regional variations.

Figure 4 displays three heatmaps illustrating the evolution of average values for the IDEOLOGY variable based on age (Y-axis) and year (X-axis) for the entire population, using various rolling window configurations. The heatmap on the left shows the original values without any additional aggregation (1 × 1 window). The central heatmap is based on a 3 × 3 window, smoothing variations by incorporating data from adjacent ages and years. Finally, the heatmap on the right uses a 5 × 5 window, offering a more smoothed depiction of ideological trends. These configurations highlight how progressive aggregation through rolling windows reduces noise and facilitates the identification of overarching patterns in ideological evolution, particularly for cohorts (Southwest-to-northeast diagonals in the figure) with smaller sample sizes.

Data Records

Each of the 1,554 tables, as well as the aggregated long-table dataset, generated in this study is available in two formats: RDS and XLSX. They are stored in a Mendeley Data repository, accessible at https://doi.org/10.17632/769tznfbsx. The tables are organized into a hierarchical folder structure designed to facilitate their location and loading.

At the first level of the folder-hierarchy, there are two main folders: “period_tables” and “cohort_tables”. The first contains the period (age-year) tables, and the second contains the cohort (age-generation) tables. Both folders are subdivided into two subfolders: one for files in the RDS format and another for files in the XLSX format.

At the final level of the folder-hierarchy, the tables are grouped into 37 subfolders corresponding to the levels of the classification factors: 2 for gender, 5 for educational attainment, 10 for each gender-education combination, 19 for regions, and one additional folder for the overall population. The names of these subfolders combine the word “tables” with a suffix identifying the level of the factor. For example, the subfolder “tables_Higher education” contains the tables corresponding to respondents with higher education. For the regional folders, numeric identifiers (“01”, “02”,…, “19”) were used, following the order presented in Table 2.

Within each subfolder at this final level, there are a total of 21 files. Each file is labeled by combining the type of information it contains (“mean” for averages, “size” for sample sizes, and “var” for variances) with the configuration of windows used. For instance, a file named “mean_5 × 5.xlsx” contains the average ideology values obtained using a double 5 × 5 window, corresponding to the group identified by the folder name. The same file names are used across all subfolders to facilitate loading via programming. This organization allows researchers to quickly identify the tables that best meet their specific needs and load them with ease.

Alternatively, practitioners can find the same information in the files “SCID.xlsx” and “SCID.rds”, which contain all the data in a single dataset in long-table format. These files are located in the parent folder and include seven columns labeled: ‘YEAR’, ‘AGE’, ‘group’, ‘window’, ‘size’, ‘mean’, and ‘var.’ The variable COHORT has been omitted to save space, as it can be derived from YEAR and AGE. The values in the ‘size’, ‘mean’, and ‘var’ columns are the same as those available in the 1,554 tables, while the remaining columns provide context for these values. ‘YEAR’ and ‘AGE’ indicate the year and age corresponding to the value, ‘group’ specifies the subgroup population, and ‘window’ denotes the double window used for the calculations.

All the described files, which make up the SCID database[18], are stored in a Mendeley Data repository, accessible at https://doi.org/10.17632/769tznfbsx.

Technical Validation

The validation of SCID was conducted through an internal consistency analysis using distinct statistical approaches. First, we assess the implicit assumptions underlying double rolling windows. Second, we analyze the implications of scale changes within the context of our research. Finally, we evaluate the statistical consistency of the estimates obtained across different window configurations.

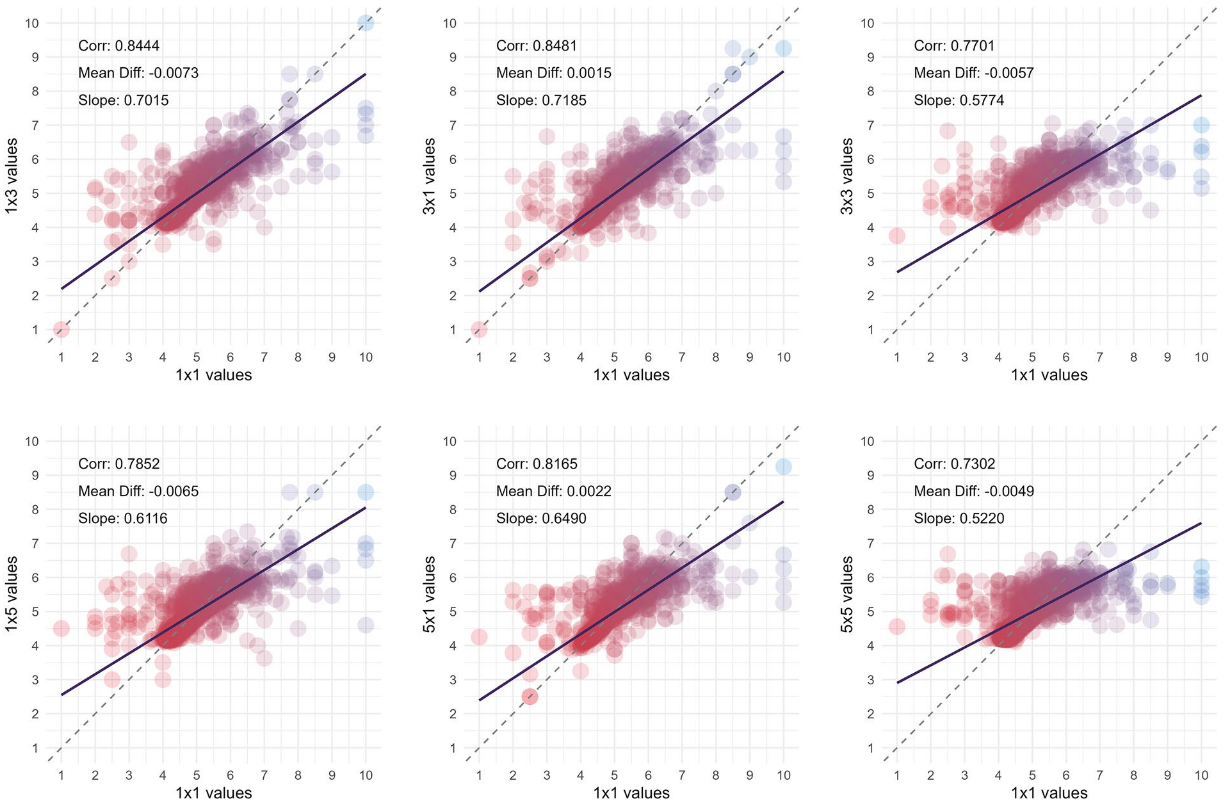

Fig. 5 Comparison of average IDEOLOGY values between the baseline configuration (1 × 1) and various two-rolling window configurations (1 × 3, 3 × 1, 3 × 3, 1 × 5, 5 × 1, and 5 × 5). The points in the scatter plots are colored on a scale ranging from red (low IDEOLOGY values) to blue (high IDEOLOGY values), highlighting consistent patterns across the entire range of values.

The use of rolling windows relies on the assumption of similarity among neighboring observations—where 'neighbors' are defined as individuals of nearly the same age in a given year or the same age in an adjacent year. To evaluate this assumption, we have compared the average IDEOLOGY value estimated using a 1 × 1 window for each AGE-YEAR pair (A, Y) to: (i) the corresponding values of pairs $(A-1, Y)$, $(A+1, Y)$, $(A-1, Y-1)$, $(A, Y-1)$, $(A+1, Y-1)$, $(A-1, Y+1)$, $(A, Y+1)$ and $(A+1, Y+1)$—i.e., all immediate neighbors following a king's move pattern in a Cartesian age-time system—and (ii) all other pairs. As expected, the absolute value average distance in the first set of comparisons is significantly smaller than in the second, 0.2637 versus 0.6296.

Another key aspect of SCID construction is the transformation of self-reported ideological positions across different scales to express all responses on the 1–10 scale, which is used in more than 90% of the studies. This transformation is particularly critical when comparing average values before and after mid-1982, as prior to that date, the CIS almost exclusively employed the 1–7 scale: 97.22% of studies and 98.94% of responses. Fortunately, after that date, studies using the 1–7 scale (as well as other scales employed by CIS) occasionally coincide with studies using the 1–10 scale. We leverage this to compare average values across years and ages between studies using the original 1–10 scale and those using the 1–7 scale transformed to a 1–10 scale. While the choice of scale matters[4], its impact on averages is relatively minor, as the average difference between the original 1–10 scale and the transformed 1–7 scale across years and ages is just 0.0798 points.

Finally, the average values obtained across the different window configurations are compared with the baseline configuration (1 × 1). Figure 5 displays six scatter plots illustrating these comparisons for each window configuration (1 × 3, 3 × 1, 3 × 3, 1 × 5, 5 × 1, and 5 × 5). Each plot includes the correlation coefficient, the average mean difference between configurations, and the slope of the linear regression line.

The results show a high correlation between the configurations, supporting the consistency of the average values regardless of window size. Additionally, the mean differences are minimal. The points in the scatter plots are colored on a scale ranging from red (low IDEOLOGY values) to blue (high IDEOLOGY values), highlighting consistent patterns across the entire range of values.

The data available in SCID, however, are not without limitations. First, it cannot be ruled out that individuals within the same generation, or across different generations, may interpret the ideological scale differently. Furthermore, it is even possible that the same individual could interpret the scale differently at two distinct points in time due to their personal evolution, the socioeconomic context, or the political environment. Second, this study focuses solely on the ideological scale, even though in certain regions of Spain (e.g., Catalonia and the Basque Country), this scale interacts with the national identity scale, shaping its meaning in complex ways. More generally, the use of a single-dimensional scale may oversimplify the complexities of ideological positions, which could be multidimensional in nature. Third, samples where respondents did not answer the ideology question were excluded from the computations. This could introduce biases in the final statistics if the missing (nonresponse) mechanism is not completely at random[19]. By acknowledging these limitations, future research can work towards addressing these issues to refine the robustness and applicability of SCID.

Usage Notes

SCID is designed for researchers studying ideological trends in Spain from cross-sectional, longitudinal, and generational perspectives. The tables are provided in accessible formats (RDS and XLSX) to facilitate integration with statistical analysis tools. To ensure proper use, attention should be paid to the selected window configurations, as these influence both sample size and the interpretation of results. For instance, broader configurations like 5×5 provide more robust estimates for wide age groups and time periods, while narrower configurations like 1×1 would theoretically allow for more precise analyses of specific age and time slices.

Code availability

Two R script files are provided for generating tables and storing them as individual files from the depurated dataset. These scripts are available in the "code" folder of the Mendeley Data repository, accessible at https://doi.org/10.17632/769tznfbsx, where the SCID is stored. The *fn_rw_tables.R* file contains two functions: *rw_tables()* and *save_results()*. The *rw_tables()* function processes a dataset in wide format, constructing a set of tables that summarize ideological self-placement across different age and year/cohort groups. It performs a cartesian join to create all possible combinations of the primary (e.g., AGE) and secondary (e.g., YEAR or COHORT) variables. The function then aggregates key statistics such as sample size, mean values, and variance using a rolling window approach to smooth fluctuations over time. The *save_results()* function automates the storage of these tables in both RDS (R serialized object format) and XLSX (Excel spreadsheet format). It ensures structured output by creating necessary subdirectories and following predefined naming conventions. The function loops through all generated tables, saving them in separate files, which facilitates easier retrieval and further analysis. It also provides verbose output to track the saving process. The *script.R* file provides the necessary code to generate and save the tables in both RDS and XLSX formats, following the naming conventions and folder structure outlined in this document. To ensure that readers can replicate the process and verify the functionality of the provided code, the script also includes lines that generate a dataset with 5,000 observations and the same variables as the depurated dataset used to create the tables described in this paper. The depurated data cannot be directly provided as it contains raw data from CIS surveys, which are proprietary and subject to CIS copyright rights. The data processing and table generation were conducted using R (version 4.4.1)[20] in combination with RStudio (2024.09.0 Build 375)[21]. For data manipulation and transformation operations, the packages dplyr (version 1.1.4)[22] and data.table (version 1.16.2)[23] were primarily used, chosen for their efficiency in handling large datasets and optimizing computational processes. The degree of accuracy/reliability of the results derived from the exploitation of the CIS raw data is the sole responsibility of the authors.

Received: 3 February 2025; Accepted: 12 March 2025;
Published online: 12 May 2025

References

1. Mannheim, K. in *Essays on the Sociology of Knowledge: Collected Works* (ed. Kecskemeti, P.) Ch. The Problem of Generations (Routledge, 1927).
2. Greene, S. Social identity theory and party identification. *Social Science Quarterly* **85**, 136–153, https://doi.org/10.1111/j.0038-4941.2004.08501010.x (2004).
3. Campbell, A., Philip, C., Warren, M., & Stokes, D. *The American Voter* (University of Chicago Press, 1960).
4. Aybar, C., Pérez, V. & Pavía, J. M. Scale matters: unravelling the impact of Likert scales on political self-placement. *Quality & Quantity* **58**, 3725–3746, https://doi.org/10.1007/s11135-023-01825-2 (2024).
5. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning*, 2nd ed https://doi.org/10.1007/978-0-387-84858-7 (Springer, 2009).
6. Martínez, V. Diseño de encuestas de opinión: barómetro CIS. *Qüestiió* **23**, 343–362 (1999).
7. Pavía, J. M. & García-Cárceles, B. Una aproximación empírica al error de diseño muestral en las encuestas electorales del CIS. *Metodología de Encuestas* **14**, 45–62 (2012).
8. Cea D'Ancona, M. A. Calidad, confianza y participación en encuestas. *Papers* **107**, e3074, https://doi.org/10.5565/rev/papers.3074 (2022).
9. Cazcarro, I., Serrano, A. & Sarasa, C. Valoración del gasto en servicios públicos por la ciudadanía española: sociodemografía, actitudes, valores y economía pública. *REIS: Revista Española de Investigaciones Sociológicas* **185**, 43–64, https://doi.org/10.5477/cis/reis.185.43-64 (2024).
10. Pavía, J. M. & Aybar, C. Field rules and bias in random surveys with quota samples. An Assessment of CIS Surveys. *SORT-Statistics and Operations Research Transactions* **42**, 183–206, https://doi.org/10.2436/20.8080.02.74 (2018).
11. Weiss, S.M., Indurkhya, N., Zhang, T., & Damerau, F.J. *Text Mining: Predictive Methods for Analyzing Unstructured Information* https://doi.org/10.1007/978-0-387-34555-0 (Springer, 2005).
12. Onan, A., Korukoğlu, S. & Bulut, H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications* **57**, 232–247, https://doi.org/10.1016/j.eswa.2016.03.045 (2016).
13. Rose, S., Engel, D., Cramer, N. & Cowley, W. Automatic keyword extraction from individual documents, in *Text Mining: Applications and Theory* (eds. Berry, M.W. & Kogan, J.) **Ch. 1** https://doi.org/10.1002/9780470689646.ch1 (Wiley, 2010).
14. Sun, S., Luo, C. & Chen, J. A review of natural language processing techniques for opinion mining systems. *Information Fusion* **36**, 10–25, https://doi.org/10.1016/j.inffus.2016.10.004 (2017).
15. Castelnuovo, E. Fitting U.S. trend inflation: A rolling-window approach, in *DSGE Models in Macroeconomics: Estimation, Evaluation, and New Developments* (eds. Balke, N., Canova, F., Milani, F. & Wynne, M. A.) **Ch. 5** (Emerald Publishing Limited, 2012).
16. Dimoudis, D., Vafeiadis, T., Nizamis, A., Ioannidis, D. & Tzovaras, D. Utilizing an adaptive window rolling median methodology for time series anomaly detection. *Procedia Computer Science* **217**, 584–593, https://doi.org/10.1016/j.procs.2022.12.254 (2023).

17. Wang, W. *et al*. Forecasting methanol-to-olefins product yields based on Relevance Vector Machine with hybrid kernel and rolling-windows. *Chemical Engineering Science* **301**, 120656, https://doi.org/10.1016/j.ces.2024.120656 (2025).
18. Aybar, C., Pérez, V. & Pavía, J. M. Spanish Cohort Ideology Database. *Mendeley Data* https://doi.org/10.17632/769tznfbsx (2025).
19. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys* https://doi.org/10.1002/9780470316696 (John Wiley & Sons, 1987).
20. R Core Team. R: A language and environment for statistical computing, version 4.4.1. *R Foundation for Statistical Computing* https://www.R-project.org/ (2024).
21. Posit, P. B. C. RStudio: Integrated Development Environment for R, version 2024.09.0 Build 375. https://posit.co/products/open-source/rstudio/ (2024).
22. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* https://CRAN.R-project.org/package=dplyr (2024).
23. Barret, T. *et al. data.table: Extension of 'data.frame'* https://CRAN.R-project.org/package=data.table (2024).

Acknowledgements

The authors wish to thank two anonymous reviewers and the editor for their valuable comments and suggestions. They also acknowledge the support of Generalitat Valenciana through project CIAICO/2023/031 (Conselleria de Educación, Cultura, Universidades y Empleo) and of Ministerio de Ciencia, Innovación y Universidades through project PID2021-128228NB-I00. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The usual disclaimers apply.

Author contributions

C.A.: Data curation, Funding acquisition, Investigation, Supervision, Software, Validation, Writing – original draft, Writing – review & editing. V.P.: Data curation, Software, Investigation, Visualization, Writing – original draft, Writing – review & editing. J.M.P.: Conceptualization, Funding acquisition, Supervision, Validation, Writing – review & editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.