Perspectives on Politics

Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets

--Manuscript Draft--

Manuscript Number:	POP-D-24-00703R3
Full Title:	Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets
Article Type:	Reflection
Corresponding Author:	Adrián Del Río University of Oslo: Universitetet i Oslo NORWAY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Oslo: Universitetet i Oslo
Corresponding Author's Secondary Institution:	
First Author:	Adrián Del Río
First Author Secondary Information:	
Order of Authors:	Adrián Del Río
	Woseeok Kim
	Carl Henrik Knutsen
	Anja Neundorf
	Agustina Paglayan
	Eugenia Nazrullaeva
Order of Authors Secondary Information:	
Abstract:	Assembling datasets is crucial for advancing social science research, but researchers who construct datasets often face difficult decisions with little guidance. Once public, these datasets are sometimes used without proper consideration of their creators' choices and how these affect the validity of inferences. To support both data creators and data users, we discuss the strengths, limitations, and implications of various data collection methodologies and strategies, showing how seemingly trivial methodological differences can significantly impact conclusions. The lessons we distill build on the process of constructing three cross-national datasets on education systems. Despite their common focus, these datasets differ in the dimensions they measure, definitions of key concepts, coding thresholds and other assumptions, types of coders, and sources. From these lessons, we develop and propose general guidelines for dataset creators and users aimed at enhancing transparency, replicability, and valid inferences in the social sciences.
Response to Reviewers:	Dear editors and reviewers, thank you very much for your input. We are greatful that our article is conditionally accepted for publication. We have added the small editing changes, replication files and updated bios. Your comments helped us improve the article greatly and think of new issues that we will handle in future publications (e.g., more rules and clarifications in the codebooks). We hope this article will have a pedagogical value for a non-expert audience and would-be scholars. In addition, our intention is to demonstrate the strenghts and weaknesses of our datasets for the sake of transparency in data creation, informing those who will use them in the near future

Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets

Adrián del Río,[1] Wooseok Kim,[2] Carl Henrik Knutsen,[1] Anja Neundorf,[2] Agustina Paglayan,[3] & Eugenia Nazrullaeva[4]

Abstract: Assembling datasets is crucial for advancing social science research, but researchers who construct datasets often face difficult decisions with little guidance. Once public, these datasets are sometimes used without proper consideration of their creators' choices and how these affect the validity of inferences. To support both data creators and data users, we discuss the strengths, limitations, and implications of various data collection methodologies and strategies, showing how seemingly trivial methodological differences can significantly impact conclusions. The lessons we distill build on the process of constructing three cross-national datasets on education systems. Despite their common focus, these datasets differ in the dimensions they measure, definitions of key concepts, coding thresholds and other assumptions, types of coders, and sources. From these lessons, we develop and propose general guidelines for dataset creators and users aimed at enhancing transparency, replicability, and valid inferences in the social sciences.

[1] University of Oslo
[2] University of Glasgow
[3] University of California, San Diego
[4] University of Konstanz

Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets

Adrián del Río,[1] Wooseok Kim,[2] Carl Henrik Knutsen,[1] Anja Neundorf,[2] Agustina Paglayan,[3] & Eugenia Nazrullaeva[4]

Abstract: Assembling datasets is crucial for advancing social science research, but researchers who construct datasets often face difficult decisions with little guidance. Once public, these datasets are sometimes used without proper consideration of their creators' choices and how these affect the validity of inferences. To support both data creators and data users, we discuss the strengths, limitations, and implications of various data collection methodologies and strategies, showing how seemingly trivial methodological differences can significantly impact conclusions. The lessons we distill build on the process of constructing three cross-national datasets on education systems. Despite their common focus, these datasets differ in the dimensions they measure, definitions of key concepts, coding thresholds and other assumptions, types of coders, and sources. From these lessons, we develop and propose general guidelines for dataset creators and users aimed at enhancing transparency, replicability, and valid inferences in the social sciences.

[1] University of Oslo
[2] University of Glasgow
[3] University of California, San Diego
[4] University of Konstanz

1. Introduction

Political scientists are interested in complex concepts: democracy, war, economic development, protest, or nationalism. To study them, researchers sometimes create original datasets that measure these concepts across multiple units (e.g., countries, provinces, municipalities). Constructing original datasets usually requires considerable resources, but the payoffs for the discipline as a whole can be large, as these datasets eventually become public and enable not only their creators but also other researchers to study a wide range of questions.

Making appropriate descriptive and causal inferences based on datasets created by other academics, however, is not straightforward. It requires understanding how the dataset was constructed—how variables capture multidimensional concepts, how each dimension is operationalized, what information and sources were used, or what coding assumptions were made. Consider, for example, the concept of democracy. There is widespread agreement among democracy researchers that democracy entails, at the very least, two dimensions: competitive elections *and* mass enfranchisement. Despite this agreement, cross-national datasets that code democracy frequently disregard the "mass enfranchisement" dimension (Munck and Verkuilen, 2002). Moreover, among those that do measure enfranchisement, some measure it continuously, while others use varying thresholds above which a country is considered democratic–e.g., a *majority of adult males* must be able to vote in Boix et al.'s (2013) classification, whereas *universal male* suffrage or, simply, *universal* suffrage is required in Skaaning et al. (2015).

This example demonstrates a broader point: dataset creators have ample freedom to choose which dimension(s) of a complex concept to measure and how exactly to measure it. As a result, different datasets may offer variables that, despite using the same terms and referring to the same basic concept (e.g., "democracy"),[5] measure different dimensions of that concept, have different validity and reliability characteristics, and are collected in different ways. One downstream consequence is that a high cross-measure correlation is not necessarily ensured. Even in the absence of measurement error, two variables that *appear* to tap into the same concept may exhibit divergences because of seemingly trivial—but, at closer inspection, important—differences in how they were constructed.

To further illustrate this matter, consider the different patterns that emerge in Figure 1 depending on which measure is used to capture four important phenomena: intra-state conflict, democracy, education centralization, and repression. The figure shows the global means of different measures for each of these concepts, using only country-year observations with data for all measures of a given concept. Still, the measures exhibit differences that may have meaningful implications for

[5] Sometimes dataset creators may even use the same terms to refer to entirely different concepts. For example, many scholars (Coppedge et al. 2020) and, especially, citizens across the world operate with entirely different notions of what "democracy" means (such as "regimes that produce economic development"; e.g., Knutsen and Wegmann 2016).

inference. For example, depending on which data source we use, the share of countries with intra-state conflict in recent years oscillates between 5% (HM) and 33% (UCDC); democratic erosion took place during the 1960s according to one measure (BMR) but not the other (RoW, which builds on V-Dem data); the degree of education centralization varied more from 1945 to 1995 according to one measure (EPSM) than the other (V-Indoc); and the level of repression around the world increased (PVI), remained similar (CIRI), or declined (PTS) in 2011 relative to 1981.[6] Our point is not that measures of the same (or at least similar) concept should never diverge—justifiable differences in conceptual specifications, operationalization, aggregation, or other features can lead to different measurement outputs—but that both data creators and users must be mindful and transparent about the measurement process and its potential implications for inference.[7]

Figure 1. Comparisons of Different Measures of the Same Concept

Please add Figure 1 here

Note: CoW=Correlates of War (Sarkees and Wayman, 2010); HM=Haber and Menaldo (2011); UCDP=Uppsala Conflict Data Program (Pettersson, 2022); BMR=Boix et al. (2013); RoW=Regimes of the World (Coppedge et al., 2023); EPSM=Education Policies and Systems across Modern History (del Rio et al., 2024); V-Indoc=Varieties of Indoctrination (Neundorf et al., 2023); CIRI=Cingranelli-Richards Human Rights Dataset (Cingranelli and Richards, 2010); PTS=Political Terror Scale-Amnesty International (Gibney et al., 2022); PVI=Physical Violence Index (Coppedge et al., 2023). CIRI and PTS are ordinal variables that have been rescaled to a unit interval using min-max scaling to facilitate comparisons.

In this paper, we discuss the advantages, limitations, and trade-offs involved in creating original cross-national datasets for research purposes and distill lessons and guidelines for both dataset creators *and* users.[8] To do so, we draw on the collective knowledge developed by the creators of three different but interrelated longitudinal, cross-national datasets on education systems: the Education Policies and Systems across Modern History (EPSM) dataset (del Rio et al., 2024), the Varieties of Indoctrination (V-Indoc) dataset (Neundorf et al., 2023), and the Historical Education Quality (HEQ) dataset (Paglayan, n.d.). While all three datasets contain seemingly similar measures of education, they differ in many respects. This goes for easily visible differences such as coding *de jure* (formal-legal) versus *de facto* (operation in practice) features of education systems or relying on country experts versus in-house coders, as well as more subtle but consequential differences such as how they deal with uncertainty or what thresholds are used to establish coding categories. Our goal is to codify good practices and share tacit knowledge

[6] The two datasets use different thresholds of how they define intra-state conflict based on the number of deaths. HM uses a threshold of 1,000 deaths while UCDC uses 25.

[7] This example pertains to descriptive inference, but the more general point on the relevance of choice of measure holds also for causal inference. In Appendix B, we provide a short application assessing the causal effect of democratization on education centralization, using the education centralization measures from EPSM and V-Indoc.

[8] We underscore that this paper focuses on the creation and use of research datasets by researchers. The creation and use of other types of datasets, notably including official statistics created by governments on everything from GDP to COVID-19 deaths, is also fraught with different pitfalls and is the subject of a separate literature (see, e.g., Jerven, 2013; Martinez, 2022; Knutsen and Kolvani, 2024).

developed through the experience of various data collection efforts made by different teams of researchers. We remark that not all of these practices or insights were obvious to us before embarking on the different data collection efforts. In Appendix A we give a more detailed overview of unanticipated challenges and how we changed strategies or adopted measures to mitigate them in the hope that future dataset creators may learn from our experiences.

By opening the black box of dataset creation, we hope to stimulate greater transparency at this stage of the research process. While the discipline has moved toward a norm of transparency in data *analysis*, a similar norm has yet to be developed regarding the process of dataset *creation*. Developing such a norm is crucial because, as will become clear in this paper, the choices made during the process of constructing new datasets can have far-reaching implications both for descriptive[9] and causal inferences.[10] To move the needle in this direction, we pay considerable attention to both the advantages and disadvantages associated with various data collection decisions. This is not because we believe that the latter are more prevalent in the datasets we examine relative to other datasets, but rather out of a conscious effort to normalize the process of making various measurement challenges and trade-offs as clear and transparent as possible. The peer-review process and other features of academia may incentivize dataset creators to hide or minimize the disadvantages or limitations of their datasets, which does a disservice to readers, users of these datasets, and the research community more broadly. We hope that by reflecting deeply and being open about the limitations of our datasets, we can raise awareness about the inherent limitations in assembling and using *any* dataset.

Our main contribution is to develop a set of guidelines for dataset creators and users, which we summarize in Table 2 of the final section following a detailed reflection on how to collect data and its challenges. These guidelines are aimed at enhancing transparency, replicability, and valid inferences in the social sciences. Furthermore, our paper contributes to ongoing methodological debates in political science. First, Gemenes (2012, 595) argues that using secondary sources (e.g., party newspapers, leaders' speeches, etc.) instead of primary sources (party election manifestos) to code the ideological positions of parties can increase non-classical measurement error.[11] We reach a similar conclusion in the context of coding *de jure* education policies. Moreover, we illustrate a common trade-off that researchers face when choosing whether to rely exclusively on primary sources (reducing measurement errors) or whether to also use secondary sources (reducing coding costs and increasing coverage). Second, we contribute to the ongoing debate about the advantages and disadvantages of relying on factual data sources versus expert assessments (e.g., Little and Meng, 2024; Knutsen et al., 2024). For instance, we highlight how different data types

[9] For a recent discussion on democracy measurement and time trends in global democracy, see Little and Meng (2024) and Knutsen et al. (2024).
[10] E.g., Casper and Tufis (2003).
[11] See also Dinas and Gimenes (2010).

may have varying benefits and drawbacks depending on what concept or concept dimensions one aims to measure, as well as whether one aims to capture *de jure* or *de facto* aspects of the concept.

2. Background: Three datasets, same topics, different methods

We begin by providing a brief overview of the three datasets—EPSM, V-Indoc, and HEQ—that form the basis of the lessons we draw in later sections for dataset creators and users (for detailed dataset descriptions, see Appendix A). These cross-national longitudinal datasets offer rich information about the content of education, teacher training and recruitment policies, and the distribution of authority over the education system. However, while EPSM and HEQ focus primarily on *de jure* policies, V-Indoc focuses mainly on what the content of education and teacher recruitment look like in practice. The information used to construct each dataset also varies: EPSM relies on a combination of primary and secondary sources for 145 countries from 1789-2020; V-Indoc relies on country-expert assessments across 160 countries from 1945-2021; and HEQ, still under construction, relies exclusively on primary sources such as education laws, regulations, decrees, and national curriculum plans, and to date covers five countries over the past two centuries. Table 1 provides an overview of the datasets, including their key characteristics and main advantages or disadvantages.

These datasets illustrate a common trade-off between coverage in terms of country-years and potential sources of measurement errors and, hence, the precision of the data. For example, primary sources offer accurate data for datasets focused on capturing information about *de jure* policies, but gathering all the relevant primary sources can be extremely time-consuming and may not be possible for some countries or periods. Thus, researchers seeking to enhance the accuracy of their measures may need to decrease the coverage of their sample. This trade-off is evident when comparing the coverage of EPSM, which combines primary and secondary sources, and HEQ, which relies entirely on primary sources. While data assembly took around 19-22 hours per country for EPSM, it took between 3-6 months in the case of HEQ.

Related, while using secondary sources can enable dataset creators to expand the geographic and temporal scope of their dataset,[12] one downside of secondary sources is that the information they contain could be incomplete or inaccurate. In fact, in the process of assembling HEQ, the team discovered that the conventional wisdom inherited from influential studies about the history of education in some countries was not corroborated by actual historical records. These mistakes stemmed from a tendency in the secondary literature to assume (incorrectly) that a *de facto* education practice was grounded in a *de jure* policy or a tendency to focus on the most famous

[12] By relying on secondary sources in English and other languages (often combined with asking for interpretation and inputs from country-specific experts), and thus drawing on information from existing summaries of education policy changes over time, one upside was that the EPSM team could code countries whose local language they did not speak and thus make data collection for a larger number of countries feasible.

education laws and neglect lesser-known laws and regulations that nonetheless formed part of the *de jure* educational landscape. Indeed, early comparisons between EPSM and HEQ revealed some measurement errors (which were later corrected) stemming precisely from EPSM's reliance on secondary sources and expert knowledge when education laws were not accessible. These sources could be influential but sometimes inaccurate.

While relying on legal texts helps us measure *de jure* education policies, laws tell us little about whether these policies were, in fact, implemented. For information about on-the-ground education practices, we need a different data collection approach. Here again, a trade-off between breadth and accuracy arises. For example, one could obtain information about what children are actually taught in school based on classroom observations[13] or by surveying scholarly experts who have in-depth knowledge of a country's education system. The former will likely produce more accurate results, but conducting classroom observations is far more costly than surveying experts. Moreover, classroom observations allow us to gather data on current and future education practice, but we cannot rely on them if our goal is to collect data about the past.

Experts assessments are one approach for collecting data about past practices. By drawing on their in-depth contextual knowledge and evaluative judgment of a topic, country-specific experts—sometimes recruited locally from the country of interest—can offer guided insight into difficult-to-measure aspects of education systems, such as

Table 1. Advantages and disadvantages of different data collection methods

Dataset	*Overview*	*General characteristics and advantages*	*Disadvantages*
EPSM	Data source: Legal texts and secondary sources available. In-house trained RAs. Cases are distributed based on language expertise and cross-checked by a second person. Coverage: 145 countries, 1789-2020 (country-year N=22,862). Indicators: 21 indicators (4 on compulsory education, 7 on ideological content teaching, 7 on school autonomy, 3 on teacher training). Costs: Approx. 1000 USD per country	+ Large temporal and cross-national coverage. + Includes uncertainty measures per group of indicators. + Includes ample information detailing coding decisions and references to help users obtain qualitative information of the case. + Data sources available. + Measures of *de jure* education policies. + Do not rely exclusively on language expertise. + Relatively quick.	-Rely on primary and secondary sources available online or through library exchange, which can be limited for some countries and historical periods. - Even if cross-checked, secondary sources can be inaccurate. - The (first version of the) dataset excludes small countries (number of inhabitants below 1 million). - Only categorical variables or ordinal scales. - Requires resource-demanding

[13] What is *taught* in schools need not coincide with what is *learned* by students. Student outcomes can be measured, for example, by standardized tests of student knowledge and skills (e.g, PISA, TIMMS, etc.), surveys of political and economic attitudes (e.g., Cantoni et al. 2017), or other instruments. The datasets discussed in this paper were created with the intention of measuring education policies and practices, not student outcomes.

			measures and extensive communication to ensure cross-coder comparability and high reliability.
V-Indoc	Data source: Expert-coded questionnaire. Multiple coders per data point, providing judgments based on their expertise. Coverage: 160 countries, 1945-2021 (country-year N=10,923). Indicators: 27 indicators (21 on education) and 13 indices (aggregated indicators). Costs: Approx. 2,000 USD per country	+ Large cross-national coverage. + Includes uncertainty measures for all estimates. + Each indicator has an ordinal and continuous version. + Measures (mostly) *de facto* instead of *de jure* education practices. + Quick and relatively easy to update.	- Restricted in terms of time coverage, as expert knowledge of historical periods is limited. - Possible biased judgments by experts. - Expensive.
HEQ	Data source: Legal texts used by expert historians and a quality-assurance manager to answer a common questionnaire. Coverage: 5 countries, beginning with the first year when each country's national government starts to regulate the curriculum or teacher training and recruitment, up to 2015. Indicators: 39 indicators (5 on curriculum; 34 on teacher training and recruitment). Costs: Approx. 7,200 USD per country	+ Provides comprehensive measures of *de jure* education policies. + Relies on an exhaustive set of primary sources to substantiate each data point. + High accuracy and completeness of the information for each country-year. + Largest possible time coverage for *de jure.* policies beginning with the first year when the central government in each country began to regulate the curriculum or teacher training and recruitment.	- Limited cross-national coverage. - Expensive and time-consuming data collection. - Requires high levels of country and language expertise. - Focuses on primary education only.

politicized teacher firing or indoctrination (Marquardt and Pemstein, 2018). However, there are also disadvantages to using expert surveys. First, expertise may also be time-bounded; indeed, the reason why the temporal coverage of V-Indoc is limited to 1945 onwards is because pilot studies revealed that experts did not feel confident coding their country of expertise further back in time. Second, experts may draw on cognitive heuristics when responding to questions (Weidmann, 2022), and some responses may reflect coder bias (e.g., Little and Meng, 2024; nonetheless, this feature may also influence non-expert coding; see, e.g., Knutsen et al., 2024).[14]

Overall, the inherent tensions between breadth and accuracy (given resource constraints), and the choices made by each research team concerning which goal to prioritize result in EPSM and V-Indoc having accomplished a substantially broader coverage than HEQ in a much shorter time, but at the potential cost of accuracy. As we discuss in Section 4.4, such features of the data and the implications of the discussed trade-offs should also be considered by data users conducting

[14] In Appendix C we assess whether coding divergences between V-Indoc and the rest of the datasets are driven by the V-Indoc's number of coders employed and coders' self-reported uncertainty.

different types of studies; for instance, accuracy may be a relatively larger problem for single-country case studies, whereas smaller and selective samples may be a relatively larger problem for cross-country studies.

Another important consideration for data creators is monetary costs. While the assembly of any cross-national dataset is likely to demand considerable resources, the data collection approach chosen has consequences for costs. EPSM conducted all data collection in-house, hiring, and training research assistants, who gathered and coded primary and secondary sources and later discussed a final coding decision with the EPSM team, which resulted in an average cost of approximately $1,000 per country. V-Indoc relied on one postdoc and multiple research assistants to identify and recruit country experts, recruited and compensated close to five experts per country on average, and paid for the use of the V-Dem Institute's data collection and measurement infrastructure for an average cost of $2,000 per country. HEQ hired education historians from each country as consultants to gather all primary sources and to conduct an initial round of coding based on these sources; then, for all countries, a research assistant cross-checked the initial coding against the primary sources, which led to a back-and-forth with consultants before arriving at the final coding, for an average cost of $7,200 per country.

3. Advice for dataset creators

When collecting and assembling datasets, researchers invariably face challenges pertaining to validity, reliability, transparency, and reproducibility and need to make decisions to mitigate such issues. This section illustrates these challenges by drawing on experiences from and comparing across our three education datasets. On a related note, we discuss the practices and tools that helped us mitigate these issues and try to generalize different insights through a set of guidelines for future dataset makers, which we further detail and concretize in Appendix D1 and summarize in Table 2 in the concluding section.

3.1 Codebooks and specification/clarification

To enhance transparency, dataset creators must overcome the challenge that specific questions and question categories may have multiple plausible interpretations. A related challenge is that the meaning of specific terms (e.g., "public school") may vary across countries and over time. Thus, dataset creators should pay particular attention to specifying their codebooks so that one minimizes the number of plausible interpretations per key term, concept, question, or category, ideally ensuring that they can only be interpreted in one way. While this might sound straightforward, our experiences with codebook construction suggest it is often hard to achieve in practice. Accomplishing unambiguous interpretation requires anticipating all possible interpretations of answer categories and possibly breaking complex questions up into two or more to mitigate multi-dimensionality.

Maintaining consistent definitions of evolving terms or even concepts, and avoiding ambiguity in interpretations, and multi-dimensionality are, as indicated, often surprisingly difficult. Indeed, these issues may even be hard to detect. Yet, several (fairly straightforward) strategies can help address such issues. Dataset creators should provide clear definitions of key concepts, clarifications, and even hypothetical or brief empirical examples to illustrate the coding procedure. This not only helps ensure transparency to data users wondering exactly how questions and categories should be interpreted (or align with their specific research questions and contexts), but also improves inter-coder reliability and reproducibility and, crucially, ensures that the dataset provides information that is comparable across space and time.[15] Dataset creators should also invest considerable time when formulating questions and allow many people, including outsiders who may interpret questions very differently, to review the codebook. For example, the V-Indoc team took two years to develop the codebook (expert questionnaire) and relied on detailed feedback and advice from subject experts at multiple stages of the questionnaire development process. These experts also helped map abstract concepts onto specific questions, which is often a key challenge when developing codebooks.

A complementary strategy is to pilot the codebook in a subset of (preferably quite different) cases to detect potential issues with how questions and categories work and adjust the codebook accordingly. All three teams—EPSM, V-Indoc, and HEQ—followed this procedure and gained valuable lessons from piloting. For example, the EPSM team piloted an initial questionnaire on a dozen countries to assess the feasibility of collecting data in different geographic and institutional contexts. The resulting experiences—as well as subsequent experiences, after the main coding had started, as we detail in our Appendix D on hard lessons learned from our data collection experiences—were instrumental for altering or developing new answer categories, specifying coding rules-of-thumb for interpreting and scoring tricky cases, developing practices for references and for including justifications of coding decisions, and detailed coding instructions and developing materials for training research assistants. Similarly, the V-Indoc team met with the coders participating in eight pilot cases to discuss their coding experiences. These conversations enabled the team to identify questions that needed to be simplified to avoid being multidimensional. For HEQ, piloting in two countries helped identify questions where the team had not anticipated the full set of possible answers, as well as additional strategies for documenting sources (via pictures) to ensure reliability.

As noted, specifying key terms and question categories and writing detailed question clarifications also enable users to understand better how the data have been produced and, thus, the dataset's

[15] Providing a glossary of terms implies that the definition of the term "X" (e.g, public school) applied *for the purpose of data collection* in country A and year T might differ with how people living in country A in year T used the term "X." Applying a common, consistent definition ensures comparability across time and space, while at the same time being conscious about (and identifying when) terms possibly being used with different contents in source materials for different contexts.

contents. Additionally, these strategies enhance intercoder reliability and, therefore, replicability (if the second coder aims to replicate the data construction effort) and dataset consistency (if there is more than one coder for the dataset). Absent such strategies, different coders will likely rely on different heuristics when making coding decisions. In cases where multiple coders contribute to a dataset, this is likely to produce different patterns of missingness (e.g., because coders treat uncertain cases dissimilarly) and different uses of particular categories (e.g., because some coders have higher thresholds for assigning high scores than others). Insofar as coders are assigned cases based, e.g., on their regional, language, or historical period expertise, there may thus be systematic differences across subsets of observations that could correlate with other factors of theoretical interest (such as income level, state capacity, or democracy, which vary systematically across regions and periods). If so, this might contribute to biased inferences in studies using the data for operationalizing independent or outcome variables.

More generally, low intercoder reliability may cause additional problems for datasets coded by more than one person, it is important to consider additional strategies for ensuring consistent coding across individuals. For example, the EPSM dataset relies on five in-house coders who coded different subsets of countries. All coders were in frequent contact with each other and the research team, which meant that several other strategies could be applied to enhance intercoder reliability. Some important strategies were a) an intensive training scheme with repeated trial coding of the same cases to make sure that all coders understood the terms, tasks, and data sources similarly; b) developing and updating a joint Rules-of-Thumb (RoT) document for tricky cases (e.g., where coding decisions indicated by the codebook were ambiguous), detailing how particular types of cases were supposed to be interpreted and coded; c) active communication through a joint web platform and (sometimes) physical co-location when coding, allowing coders to find joint solutions to challenging cases; d) a second coder going through all original codings, with subsequent adjustments. These measures were intended to aid coders in having a similar understanding of terms and underlying concepts and applying similar (preferably made explicit in the RoT document) heuristics when approaching similar cases. Nonetheless, avoiding differential interpretations and uses of heuristics across coders is close-to-impossible to guard completely against, and may lead to increased uncertainty and even biases, as noted above. Dataset creators providing coder IDs and explaining coding decisions for each coded observation may be one strategy for allowing users to assess and possibly reduce such issues in their analyses.

The country-expert coded V-Indoc dataset relies on a Bayesian IRT measurement model to make estimates comparable across experts and countries. This model was developed for the wider V-Dem dataset to deal with several issues, such as experts having different understandings of questions and applying different thresholds when choosing between categories (for details, see Pemstein et al., 2020; Coppedge et al., 2020). The measurement model method incorporates several pieces of information (e.g., experts' coding of vignettes, bridge coding of selected countries and time periods, cross-coder divergences, coder's self-reported confidence, and estimates of

coder reliability), to adjust experts' scores before aggregating them to the country-year level, which enhances the comparability, reliability, and validity of the estimates while also generating uncertainty measures for each estimate. In this process, the measurement model transforms experts' original scores on an ordinal-level indicator to a (presumed underlying) interval-level scale. The latter transformation relies on non-trivial assumptions that are indicated in the codebook (alongside references to more detailed documentation) together with the measurement level of the variable contained in the dataset.

The latter point illustrates a more general one for codebook construction: Indicator entries should contain precise information about scaling in order to provide users with the requisite information to, e.g., avoid erroneous interpretations of scores and evaluate which kinds of analyses variables may be used for. We list this as one guideline for constructing codebooks, alongside several other pieces of advice indicated in this section, in Appendix Table D.1.1. Scaling information is provided in the codebooks of the three education datasets, although the information is sometimes insufficiently specified or otherwise problematic.[16]

3.2 Triangulating sources

A common practice among historians is to triangulate information from multiple sources, which helps to acquire a holistic picture of the object of study, assess the reliability of different sources, and enhance confidence in our conclusions when multiple sources point in the same direction. While triangulation is often used by researchers relying on qualitative evidence (e.g., by combining interviews with qualitative document analysis), the last two points are also relevant for the construction of quantitative datasets.

For example, the authors of EPSM first collected secondary data sources on the history of education and other relevant sources to identify key legislation and obtain background information about the case. Afterward, the authors collected all available legislation online or through library exchange. When data sources diverged, the team established a protocol and guidelines in their RoT document: If primary and secondary sources led to different coding decisions, primary sources were prioritized, and the level of confidence was also registered. If doubts prevailed after a second coder revised the case and checked intra-coder consistency, the team met and discussed potential sources of coding disagreement and strategies for additional source collection.[17] The goals of this

[16] For instance, the scale options listed for V-Indoc includes "dichotomous", which is strictly speaking not a scale option, and combines the ratio- and interval levels in another listed option. The EPSM codebook describes the measurement level of some indicators as "Multiple selection", whereas the correct measurement level is nominal. We thank a reviewer for alerting us to these and other issues with the published codebooks.The relevant entries will be corrected when updating the codebooks with future iterations of the datasets.

[17] To exemplify, one type of coding disagreement applying to former colonies, e.g. in Africa, stemmed from these colonies holding a dual education system. Since some of EPSM's items focus on the law that applies to the plurality of schools in a country, coders often required additional information on the number and types of schools built to make

procedure and the wider triangulation strategy were to improve the validity and reliability of the coding and to assess and express remaining uncertainty.

3.3 Data sources and type of coding tailored to concepts

No one way of gathering data—through automated text analysis, in-house coding, or expert surveys, to mention three examples—is superior to all others regardless of what type of concept one is trying to measure. Different data collection methods come with different strengths and weaknesses and are thus suitable for different purposes (Skaaning, 2018). The same goes for different data sources. If one wants to collect data on education laws, legal texts are a great source. Suppose one wants to collect data on how education is practiced in the classroom. In that case, legal texts may not represent these practices well, and other sources may be better suited (e.g., classroom observation, secondary sources on education systems, expert surveys, and surveys administered among local non-experts). Generally, data collection practices and data sources should be tailored to the concept one is trying to measure. Our three education datasets illustrate this point.

EPSM and HEQ set out to code (mainly) *de jure* characteristics of education systems, whereas V-Indoc explicitly aims to mainly code *de facto* characteristics that reflect how education is practiced. Coding how complex systems—be it education systems, state bureaucracies, or political regimes—actually work requires considerable in-depth knowledge. Acquiring such case-specific knowledge may be extremely time-consuming and thus infeasible for any single researcher or RA coding numerous cases. Structured (country-topic) expert surveys are thus often one effective and appropriate method for collecting and codifying extensive cross-country information in a comparable manner when questions require in-depth case knowledge; answering such questions is presumably less time-consuming for experts on a particular country since they can draw on prior knowledge (or already know which references to consult). The ambition to code how education systems work in practice is thus a key rationale behind V-Indoc employing country experts for their coding.[18]

Yet, building datasets based on answers from hundreds of experts comes at a cost. Even presenting specific questions, defining key concepts in detail, and ensuring entirely consistent coding is difficult (though, as discussed, using measurement modeling approaches may help). Limited

a coding decision in these colonies. Such information was first sought through written material, and second, if information was inconclusive, through contacting local country-specific experts.

[18] We used three main channels to recruit potential country experts. First, with the help of research assistants, we consulted the ratings of top universities in each country and collected emails of all faculty members (research and teaching focused), postdoctoral scholars, and graduate students whose research expertise is in the field of education. Second, we used Google Scholar to find academic journals, books and book chapters, policy reports, as well as regional conferences on education, and collected emails of the authors/participants. Third, we contacted education-related NGOs and policy experts outside of academia, asking them to circulate our call among their network.

communication between experts and the survey team, as well as between experts, means that implicit, individual coding heuristics may remain (instead of becoming collective and explicit via joint discussions), and divergent interpretations of concepts are hard to catch and clarify. Thus, for data types and concepts that do not require the same amount of in-depth contextual knowledge, it may be preferable to use the same group of (in-house) coders to ensure consistent coding, especially when terms may carry multiple meanings (e.g., "primary education level" or "ideological training"). Put differently, the relative benefits of in-house coding compared to expert coding increase when the level of country expertise and contextual knowledge required is smaller and conceptual ambiguity is larger. Sometimes, the sources that must be used also require specific expertise that is not country- but source-specific (e.g., some type of database or a particular type of legal text). Also, in this case, it makes more sense to train a few coders (e.g., RAs) than ask experts for each country to code.

It is possible to devise strategies that harness benefits from different approaches. The HEQ dataset, for example, relied on country-specific expert historians' local knowledge of the legal educational landscape. This knowledge increased the accuracy and completeness of information about *de jure* policies, but it also relied on an in-house quality assurance manager to ensure comparability across countries and consistency in responses within the same country over time. Even for data coded entirely in-house, when particular cases have complexities, data creators can develop protocols for finding and checking with people with knowledge of local context to obtain information.

4. Advice for dataset users

This section turns to potential pitfalls and advice for users of datasets, focusing on how different dataset characteristics—also for similarly sounding variables that differ in subtle ways across datasets—may affect inferences. Specifically, we highlight three issues, where measures across different datasets might capture the same concept, but dataset creators: (1) focus on different dimensions of that concept; (2) emphasize *de jure* or *de facto* dimensions of this concept; or (3) apply different thresholds when creating categories for the measures. To facilitate comparisons, we harmonize indicators across the three datasets so that they follow a common scale.[19] We refer to Table 2 for a summary and Appendix D for further specific guidelines.

Before we turn to these specific issues, we want to stress a general point about the importance for dataset users to be aware of the features of an existing dataset–including the goals of the dataset creators with collecting the data in the first place–to understand what that dataset can be used for, and what kinds of analyses should be avoided. Suppose that a researcher wants to identify when governments first began to mandate the inclusion of civic education in the curriculum. Both EPSM and HEQ could be used to gain some insight into this question because they contain data on curriculum policies. However, because they (intentionally) focus on *national* policies, researchers

[19] See Appendix E for more details on how the indicators are harmonized, including their original scales.

would need to complement what they can learn from these datasets with information about subnational policies obtained from other sources. Alternatively, suppose that a researcher wants to draw general conclusions about education systems globally. In that case, they should opt for datasets like EPSM and V-Indoc, which have good geographic coverage across all regions, and avoid relying on HEQ, which focuses on Europe and Latin America. Finally, suppose a researcher wants to conduct a single-country case study or focused comparisons of education policies pertaining to indoctrination. In that case, they should opt for datasets like HEQ that provide more fine-grained information about each country in the dataset, instead of relying on broader-coverage datasets like EPSM and V-Indoc, which data is suited for analyzing aggregate trends.

4.1. Same name, different content

As previewed in the introduction, several dataset creators occasionally attempt to capture the same concept but differ in the dimension of that concept that they (want to) measure. The introductory example we used was a multidimensional concept of democracy. Some datasets measure only the presence of contested elections (Cheibub et al., 2010), while others incorporate suffrage rights (Boix et al., 2013) or try to measure additional dimensions of democracy such as respect for freedom of speech or other civil rights (Coppedge et al., 2023).

Key concepts in the education literature experience a similar issue, which this section illustrates for the concept of *education centralization*. Following the emerging literature on education and state-building (e.g., Paglayan, 2022a, 2022b), education centralization refers to the concentration of authority over education policy decisions in the hands of the national government (Ansell and Lindvall, 2020; Paglayan, 2021; Neundorf et al., 2024; del Río et al., 2024). High levels of centralization denote that the national government has total control over education, while low levels reflect that education decisions are made at the regional, local, or school level.

While education centralization, as a concept, subsumes all kinds of education policy decisions, most available measures focus on the distribution of authority across government levels for a few policy areas. For example, most studies about education decentralization in Latin America during the 1990s refer specifically to decentralization in the responsibility to fund schools and/or manage their day-to-day operations (Murillo, 1999; Grindle, 2004; Kaufman and Nelson, 2004). In another example, Ansell and Lindvall's (2020) binary measure of education centralization is based on who has authority over the appointment, promotion, and payment of teachers. Some case studies instead focus on the presence of national examinations and grading standards (Zhao, 2012; Clarke et al., 2003), or national school inspection systems (Cermeño et al., 2022).

EPSM, V-Indoc, and HEQ all offer indices that measure education centralization but emphasize different dimensions (see Appendix B for a summary of how these indices are constructed). EPSM focuses on (*de jure)* the existence of a national curriculum and includes an additional dimension of national government control over school funding and management (at different levels of

education). V-Indoc focuses on national government control over education content by establishing national curricula and approving textbooks. HEQ also measures whether a centralized curriculum exists and whether the national government approves textbooks.[20] Figure 2 depicts trends in education centralization across the three datasets in the five countries for which our data overlap, first for the comprehensive indices (Figure 2a) and second for the centralization of the curriculum indicators, which are a part of all the indices and have been harmonized to make their scales comparable (Figure 2b). The darker the cells, the more centralized the education system is.

We can draw several take-away points from the figure. First, differences in the dimensions included in education centralization can have important consequences for scores (and thus, e.g., trends) in combined indices. This difference is indicated by comparing the EPSM and HEQ indices in Figure 2a, especially for Chile and Argentina. The diverging index scores suggest that a national government's control over education content, which is covered by both EPSM and HEQ and displays similar trends across datasets (c.f. Figure 2b), does not entail that it also controls other aspects of education systems, such as funding and management (only included in the EPSM index). For example, Pinochet's Chile (1970-90) maintained a centralized national curriculum but engaged in the decentralization of education funding and management to municipalities (Cox, 2005; Ministerio de Educación, 1980a, 1980b, 1980c). Indicatively, the major differences between the EPSM and HEQ indices in Figure 2a—which might, at first sight, be interpreted as low reliability for one or both of the measures—mostly disappear when focusing more specifically on curriculum centralization in Figure 2b.

Our first recommendation to dataset users is thus to be aware of the number and type of dimensions covered by the measures that they use. This is especially important when relying on indices, which are a common practice in empirical research and require researchers to be deeply familiar with the complex decisions and indicators involved in the creation of those indices.

Figure 2. Education/Curriculum Centralization
(EPSM, V-Indoc, and HEQ)

[20] In addition, HEQ measures centralization in teacher training and certification policies, although in what follows we focus on its curriculum and textbooks measures only.

Please add Figure 2 here

Note: See Appendix B for a description of the centralization indices across the three datasets. The EPSM and V-Indoc datasets have missing values in Germany 1945-49 as both datasets follow V-Dem's coding of country-years.

4.2 De jure and de facto

Another interesting pattern from Figure 2 appears when comparing V-Indoc and HEQ. Both measure education centralization based on the curriculum and textbooks but use different data collection methods. This contributes to explaining why these datasets sometimes arrive at different conclusions about the degree of education centralization. Consider the case of Argentina. Between the transition to democracy in 1983 and 1993, Argentina appears to have a more centralized curriculum according to HEQ than V-Indoc. The difference is likely to be driven, at least in part, by the fact that V-Indoc experts presumably take into account not only *de jure* but also *de facto* centralization, whereas HEQ focuses exclusively on *de jure* policies.[21] Indeed, while Argentina's 1884 law of primary education established a national curriculum for all public schools, its enforcement was imperfect, and, in practice, subnational governments had leeway to deviate from it, especially after 1983. This informal practice is captured by V-Indoc. HEQ, by contrast, with its focus on *de jure* policies, only recognizes subnational intervention in the curriculum starting in

[21] The two V-Indoc items on centralization of curriculum and textbooks are constructed to capture both de facto and de jure dimensions. The questionnaire instructions mention: "We are interested in changes over time at the aggregate country level. Please make sure your answers reflect *educational reforms* or *changes in teaching practices over time* [emphasis added]." (p. 4).

1994, when a new law formally recognized the ability of provinces to have some say over the curriculum.[22]

De jure vs. *de facto* distinctions are important in the social sciences. Researchers are, for example, often interested in understanding the extent to which changes in legislation or formal institutions produce changes in policies, practices, or power relations (Acemoglu and Robinson, 2006; Ansell and Lindvall, 2020), or whether legislation mostly institutionalizes already existing practices (Przeworski, 2004; Paglayan, 2019). Works on state capacity highlight how and why changes to legislation may not always translate into effective implementation (e.g., Fukuyama, 2004). Nevertheless, the information that researchers need to study such issues empirically is often unavailable. For researchers interested in understanding education systems, combining datasets such as V-Indoc (mostly *de facto*), EPSM (mostly *de jure*), and HEQ (purely *de jure*) can help accomplish this goal, as we illustrate in this section.

Several factors can affect gaps between *de jure* policies and *de facto* practices, including the state's fiscal and administrative capacity, the existence of school inspections, political regime type, conflict, or a country's territorial size (Lopez, 2020; Cermeño et al., 2022; Paglayan, 2024). We do not aim to explain what causes those gaps here, which is an important question that we leave for future research. Instead, we use our education data to identify and describe such gaps.

Figure 3. Trends in politicized teacher recruitment
(V-Indoc and HEQ)

[22] The 1994 Federal Law of Education gives the National Ministry of Education in Argentina the duty to establish a set of nationwide curricular prescriptions for each subject (Common Core Curriculum) but leaves considerable flexibility for provinces and municipalities to add other topics, skills, or materials to this common core (Ministerio de Educación, 1994).

Please add Figure 3 here

Note: the harmonized indicator for politicized teacher recruitment is coded as 1 if there are any political or moral requirements to becoming a teacher, and 0 otherwise.

Figure 3 draws on measures from V-Indoc and HEQ to demonstrate the relevance of the *de jure* vs. *de facto* distinction on one specific dimension of education systems: the politicization of teacher recruitment practices.[23] HEQ, which focuses on *de jure* policies, includes measures on whether applicants to teacher education programs must show proof of moral competency (yes/no) or belong to a particular religion (yes/no), and whether public primary school teachers are required to swear allegiance to the state and/or the constitution (yes/no) or to a particular party or a ruler (yes/no). Using this information, we create a dichotomized indicator of politicization in teacher recruitment that takes a value of 0 when neither of these requirements is present and a value of 1 when at least one is present. In V-Indoc, the indicator of political teacher hiring measures whether the teacher hiring criteria are *de facto* based on teachers' political views, and/or political behavior, and/or moral character.[24] The possible answer categories are: rarely/never, sometimes, often, and almost exclusively. To ease comparisons, we dichotomize the V-Indoc indicator: 0 means hiring

[23] While EPSM contains a question on ideology in teacher training, it allows for multiple answer categories that do not have exact matches with categories employed in HEQ and V-Indoc.

[24] The V-Indoc expert coders were explicitly instructed to answer this question based on *"actual practice (de facto, not legislation pertaining to the recruitment procedures for teachers)."*

decisions are rarely or never based on politicized criteria, while 1 combines the three politicized categories (i.e., sometimes; often; almost exclusively).[25]

When plotting the two measures for five overlapping countries and years in Figure 3, we observe both *de jure* and *de facto* politicization in teacher recruitment in Germany across the entire period, *de jure* but not *de facto* politicization in Italy, and some years of convergence and divergence between the measures in Argentina, Chile, and Spain. Instead of relying only on one dataset measuring either *de jure* or *de facto* aspects, contrasting otherwise fairly similar measures from two datasets may give nuanced and important descriptions of the historical developments of education systems.[26]

Let us elaborate on the added informational value of measuring both *de jure* and *de facto* aspects by returning to the case of Argentina. In the 1940s and 1950s, the Peronist administration introduced a requirement for current and new teachers to swear allegiance to the Peronist doctrine as a condition for employment in public schools. The regime purged the profession of numerous teachers—many from a middle-class background—who opposed the regime and refused to swear allegiance to it. This period is aptly captured by both HEQ's *de jure* and V-Indoc's *de facto* measures of politicization in teacher recruitment. After Peron went into exile in 1955, subsequent national governments removed the legal requirement for teachers to swear allegiance to a specific party or regime, and no new legal requirements focused on regulating teachers' political leanings were introduced, as reflected by the HEQ measure. However, in practice, the politicization of teacher recruitment remained in place for decades. First, members of the Peronist party took control of many teacher hiring commissions at the subnational level and used that power to favor the appointment of Peronist teachers. Second, during the 1970s, the dictatorship of Rafael Videla persecuted teachers not only of Peronist affiliation but also those suspected of opposing the regime. In other words, as captured by the V-Indoc measure, the politicization of the teaching profession remained in place beyond what the law stipulated, also during two periods after 1955.

4.3 Same concepts but different thresholds

[25] We note that differences between the HEQ and V-Indoc may also stem from HEQ focusing on primary school teachers, whereas V-Indoc asks about hiring practices for "the majority of teachers" in primary and secondary schools.
[26] We surmise that this lesson might apply also for other concepts such as "democracy", where both *de jure* and *de facto* measures exist, but where measurement debates have often centered on which type of measure is "better" (e.g. in terms of reducing particular measurement errors; see, e.g., Little and Meng, 2024; Knutsen et al., 2024) rather than how to fruitfully combine insights gained from different measures.

Sometimes, similar measures from different datasets may capture the exact same concept yet use different thresholds to establish coding categories. For instance, measures capturing similar minimalist (electoral) and dichotomous democracy concepts may lead to widely different empirical distributions of regimes if one has a very high bar for considering elections sufficiently "free and fair" and another applies a lower bar (Kasuya and Mori, 2021). More generally, differences in such thresholds can stem from researchers operating with different (often implicit) theoretical assumptions or even from different data collection strategies. This section illustrates how different thresholds affect inferences by discussing how HEQ and EPSM identify religious education in the curriculum.[27]

Briefly, HEQ aims to codify whether religious education is part of the official curriculum. To do so, it identifies whether religion is included as a compulsory, standalone course. The subject need not be about religion exclusively. A subject called "moral and religious education," for example, would satisfy the HEQ criterion for coding a country as mandating religious education. Similarly, EPSM requires that religious education form part of a standalone subject in the mandatory curriculum for a country to be classified as having religious education. However, an additional requirement in the case of EPSM is that religion must form part of the current regime's political system and/or consist of an official school of thought that has the status of an "official" ideology in the regime, which would be the case, for example, if religion is mentioned in the constitution. This additional criterion restriction reflects EPSM's aim to capture instruments for indoctrination and regime legitimation (and religion is only one of several relevant "ideology categories" for which compulsory, standalone civics courses are coded). As a result of this coding decision, countries with a compulsory, standalone religious course where religion is irrelevant to regime ideology will be coded as having religious education in HEQ but not in EPSM. In other words, the added criterion in EPSM means that there is a higher threshold for coding "religious education" in this dataset than in HEQ.

These different thresholds imply that HEQ is more likely to identify religious instruction than EPSM, and this is indeed what we observe in Figure 4. One case where the different thresholds help explain the divergence in how religious instruction is coded across HEQ and EPSM is post-Pinochet Chile. In 1996, Decree No. 40 introduced religion to the curriculum as a compulsory subject, leading the HEQ dataset to identify this change in the curriculum. However, because religion did not form part of the democratic regime's ideology after 1996, EPSM does not register this addition of religion into the curriculum.

[27] V-Indoc also contains information on the presence of religious content in education. But instead of considering standalone courses, it considers the history curriculum. While there might be relevant differences in thresholds for coding the presence of religious education when comparing V-Indoc's measure against those from the other datasets (e.g., V-Indoc requires that religion must be a dominant regime ideology to be coded), we leave it out of the discussion here.

Figure 4. Religious Instruction in Primary Schools (EPSM and HEQ)

Please add Figure 4 here

Note: HEQ and EPSM focus on standalone compulsory courses to detect religious values in primary education, while V-Indoc examines its presence in history courses. The y-axis reflects a harmonized scale for the three indicators between 0 and 1. For V-Indoc, the values reflect the proportion of coders (out of the total number of coders) who consider religion to be one of the top two ideologies/dominant models in the history curriculum.

This example of seemingly similar measures carrying different informational content indicates that dataset users should pay careful attention to coding rules and thresholds. This requires spending time reading also the fine print in codebooks and other dataset documentation before selecting which measure is most appropriate to use for a particular purpose.

5. Lessons

This paper has highlighted how various and specific choices on data collection and measurement influence how (even seemingly similar) indicators and indices are scored. We have done so by comparing and contrasting measures from three novel historical datasets on education systems and policies. In addition to detailing various choices faced by dataset creators and their consequences for measurement, we have discussed key challenges and issues that dataset collectors need to be attentive to, as well as strategies for mitigating them. Likewise, we addressed several, and often hard-to-detect, issues that dataset users need to be aware of, specifically highlighting how even measures that may initially seem identical could carry quite different informational content.

We hope our discussions contribute to ongoing debates on measurement, e.g., on the appropriateness of relying on expert-coded versus "objective" data, by unveiling limitations in assembling and using datasets of different kinds for different purposes. By demonstrating the importance of even (seemingly) minor assumptions and under-communicated data collection choices, we also hope that our reflections can promote a shift towards more transparency on data collection process choices and limitations with the resulting datasets. This would, in turn, contribute to enhancing the reliability and replicability of future research.

To help researchers in this endeavor, Appendix D provides a checklist, both for academic data producers and users, based on the lessons we have discussed in this study. We summarize the checklist as a set of guidelines in Table 2. One important caveat is that these guidelines reflect our experiences and considerations pertaining to the coding of country-level, historical (education) datasets, and they should not be viewed as *the* best practices that everyone should follow regardless of the type of data or other considerations (following some of the guidelines for dataset creators does, for example, require substantial resources for coding). Appendix D1 and D2 provide more detailed suggestions from each guideline, examples of how to implement them, and discussions of their potential benefits.

For data creators, we invite researchers to apply some of the measures described in this paper (and used for some or all of our three example datasets) to enhance reproducibility and transparency. This entails being explicit about all coding decisions and documenting the data sources underpinning such decisions, especially in tricky cases. If data producers have doubts about coding decisions, they should not be afraid of exposing the limitation but rather explain the source of uncertainty and rationale behind the coding decision (and even plausible, alternative decisions) in the dataset documentation. Besides a detailed codebook, a rule-of-thumb document could be useful in cases where clear rules are inapplicable or ambiguity in coding decisions remains. Such documentation not only makes coding assumptions explicit to users but also enhances coding consistency by making different coders use the same explicit heuristic instead of several implicit ones.

For data users, a careful reading of articles introducing the dataset, codebook, and other documentation is a must, as it can reveal key assumptions underlying the dataset and ensure proper interpretation and inference. Datasets are typically based on non-trivial assumptions about the relevant properties that characterize a phenomenon, often linked to research goals. Against this backdrop, dataset users should ensure that they select and cross-check those datasets that match the theoretical assumptions and purposes of their own research.

Table 2. Guidelines for Data Creators and Data Users

For data creators

1. As part of the codebook, precisely define potentially ambiguous terms, key concepts, the dimensions of key concepts that are measured, and measurement scales for each variable.[28]
2. Specify questions to be coded as much as possible and add clarifications to the main questions.
3. Include at least one item for each concept dimension and if a dimension is complex, try to break it up into two or more questions.
4. Ask experts on the topic for feedback on the codebook.
5. Conduct pilot studies, selecting diverse countries.
6. Make sure that all coders understand the concepts, tasks, and data sources similarly. Create a rule-of-thumb document to provide a set of instructions about how the data collection should proceed and what to do when data sources are unclear.
7. Active communication is key if more than one coder is involved in the data collection process.
8. If possible, have an external coder cross-checking cases.
9. Assess the extent to which the dataset has been coded consistently and make transparent the strengths and limitations of the dataset.
10. If possible, use multiple data sources to inform your coding decisions.
11. Think critically about and discuss the strengths and weaknesses of different types of data sources, before devising strategies for how to search for and use sources.
12. Include references in the dataset and facilitate access to the data sources.
13. Make your dataset publicly available (including online data exploration) and create ways to obtain

[28] Appendix D1.1 includes a detailed checklist for best practice on creating codebooks.

feedback from data users.

For data users

14. Carefully read the dataset's documentation to reveal key assumptions underlying the dataset (e.g., threshold assumptions and underlying dimensions of the operationalizations).
15. Prioritize datasets that match the theoretical assumptions and purposes of your research over popular measures or cross-national and temporal scope.
16. When engaging in convergent validation exercises, pay careful attention to conceptual differences underlying measures that may, at first glance, seem to measure similar concepts.
17. If possible, identifying the sources of (dis)agreement in similar measures across datasets could expose different assumptions made by dataset creators and provide nuanced insights that could aid both descriptive and causal inference.

Depending on the data user's research design and goals, our study also highlights how one fruitfully can combine variables (also from different datasets) to measure different dimensions of the same concept. Nevertheless, given the caveats noted above, data users should make sure *only* to use variables that represent appropriate operationalizations of the author's concept of interest. Our study has highlighted how even variables that seem to be similar and which may even have identical names (e.g., "education centralization index"), can tap into very different (dimensions of) concepts, leading to low correlation. When this is the case, "robustness tests" that blindly substitute one variable for another may lead to very different results. Thus, providing a detailed appendix where researchers test whether the results are robust to alternative popular measures that, on the surface, seem similar may lead researchers astray. Instead, we hope that our advice on gathering detailed information and carefully evaluating the relevance of measures could motivate theory-driven discussions of the relevance of particular tests, robustness, and generalization rather than (only) data-driven discussions.

Lastly, an implication of our discussions is that medium or low correlations between measures (especially between different datasets) pertaining to the same concept are not necessarily indicative of low reliability in any of the measures assessed. Instead of prematurely concluding that divergences stem from measurement error, data users should closely inspect codebooks, documentation, and descriptions of the different measures, as it is possible that the measures differ because they capture different dimensions of a concept or even different concepts being referred to with the same term. Dataset producers, too, should pay careful attention to be clear and upfront about what concept(s) they measure and how, and they should make comprehensive documentation readily available for users.

References

Acemoglu D and Robinson JA (2006) *Economic origins of dictatorship and democracy*. Cambridge: Cambridge University Press

Ansell BW and Lindvall J (2020) *Inward conquest: the political origins of modern public services*. Cambridge: Cambridge University Press.

Boix C, Miller MK and Rosato S (2013) A complete data set of political regimes, 1800-2007. *Comparative Political Studies* 45(12), 1523-1554.

Cantoni, D., Chen, Y., Yang, D. Y., Yuchtman, N., and Zhang, Y. J. (2017). Curriculum and Ideology. *Journal of Political Economy* 125(1), 338-392 .

Casper G and Tufis C (2003) Correlation versus interchangeability: the limited robustness of empirical findings on democracy using highly correlated data sets. *Political Analysis* 11(2), 196-203.

Cermeño AL, Enflo K and Lindvall J (2022) Railroads and reform: how trains strengthened the nation state. *British Journal of Political Science* 52(2), 715-735.

Cheibub JA, Gandhi J and Vreeland JR (2010) Democracy and dictatorship revisited. *Public Choice* 143(1/2), 67-101.

Cingranelli D and Richards D (2010) The Cingranelli-Richards human rights dataset. CIRI Human Rights Data Project.

Clarke S, Timperley H and Hattie J (2003) *Unlocking formative assessment: practical strategies for enhancing students' learning in the primary and intermediate*. Auckland: Hodder Moa Beckett.

Coppedge M et al. (2020) *Varieties of democracy: measuring two centuries of political change*. Cambridge: Cambridge University Press.

Coppedge M et al. (2023) V-Dem codebook v13. Varieties of Democracy Project.

Cox CD (2005) *Las políticas educacionales en el cambio del siglo*. Santiago: Editorial Universitaria.

Del Río A, Knutsen CH and Lutscher PM (2024) Education policies and systems across modern history: a global dataset. *Comparative Political Studies* 58(5), 851-889.

Dinas E and Gemenis K (2010) Measuring parties' ideological positions with manifesto data: a critical evaluation of the competing methods. *Party Politics* 16(4), 427-450.

Fukuyama F (2004) *State-building: Governance and world order in the 21st century.* Ithaca: Cornell University Press.

Gemenis K (2012) Proxy documents as a source of measurement error in the Comparative Manifestos Project. *Electoral Studies* 31(3), 594-604.

Gibney M et al. (2022) The political terror scale 1976-2021. The Political Terror Scale.

Grindle MS (2004) Good enough governance: poverty reduction and reform in developing countries. *Governance* 17(4), 525-548.

Haber S and Menaldo V (2011) Do natural resources fuel authoritarianism? A reappraisal of the resource curse. *American Political Science Review* 105(1), 1-26.

Kaufman R and Nelson JM (2004) *Crucial needs, weak incentives: social sector reform, democratization, and globalization in Latin America*. Baltimore: Johns Hopkins University Press

Knutsen CH and Wegmann S (2016) Is Democracy about Redistribution? *Democratization* 23(1), 164-192

Knutsen CH and Kolvani P (2024) Fighting the disease or manipulating the data? Democracy, state capacity, and the COVID-19 pandemic. *World Politics* 76(3), 543-593.

Knutsen CH et al. (2024) Conceptual and measurement issues in assessing democratic backsliding. *PS: Political Science & Politics* 57(2), 162-177.

Jerven M (2013) *Poor numbers. how we are misled by African development statistics and what to do about it*. Ithaca: Cornell University Press.

Kasuya Y and Mori K (2021) Re-examining thresholds of continuous democracy measures. *Contemporary Politics* 28(4), 365-385.

Little AT and Meng A (2024) Measuring democratic backsliding. *PS: Political Science & Politics* 57(2), 149-161.

Lopez D (2020) State formation, infrastructural power, and the centralization of mass education in Europe and the Americas, 1800 to 1970. Presented at the 2020 Annual Meeting of the American Political Science Association.

Marquardt KL and Pemstein D (2018) IRT models for expert-coded panel data. *Political Analysis* 26(4), 431-456.

Martinez L (2022) How much should we trust the dictator's GDP growth estimates? *Journal of Political Economy* 130(10), 2731-2769.

Ministerio de Educación (1980a) Decree DFL N° 13.063. Biblioteca Nacional del Congreso de Chile.

Ministerio de Educación (1980b) Decree DFL N° 3.166/80. Biblioteca Nacional del Congreso de Chile.

Ministerio de Educación (1980c) Decree DFL N° 5.077/80. Biblioteca Nacional del Congreso de Chile.

Ministerio de Educación (1996) Decree 40. Biblioteca Nacional del Congreso de Chile.

Ministerio de Educación (1994) Ley No. 24.195. Biblioteca Nacional del Congreso de Argentina.

Munck GL and Verkuilen J (2002) Conceptualizing and measuring democracy: evaluating alternative indices. *Comparative Political Studies* 35(1), 5-34.

Murillo MV (1999) Recovering political dynamics: teachers' unions and the decentralization of education in Argentina and Mexico. *Journal of Interamerican Studies and World Affairs* 41(1), 31-57.

Neundorf A et al. (2023) Varieties of political indoctrination in education and the media (V-Indoc) codebook. DEMED project.

Neundorf A, Nazrullaeva E, Northmore-Ball K, Tertytchnaya K and Kim W (2024) "Varieties of indoctrination the politicization of education and the media around the world. *Perspectives on Politics*, online first.

Paglayan AS (2019) Public sector unions and the size of government. *American Journal of Political Science* 63(1), 21-36.

Paglayan AS (2021) The non-democratic roots of mass education: evidence from 200 years. *American Political Science Review* 115(1), 179-198.

Paglayan AS (2022a) Education or indoctrination? The violent origins of public school systems in an era of state-building. *American Political Science Review* 116(4), 1242-1257.

Paglayan AS (2022b) The historical political economy of education. In Jenkins J and Rubin K (eds), *The Oxford Handbook of Historical Political Economy.* Oxford: Oxford University Press.

Paglayan AS (2024) *Raised to obey: the rise and spread of mass education*. Princeton: Princeton University Press.

Paglayan AS (n.d) Historical education quality (HEQ) dataset. Work in progress.

Pemstein D et al. (2020) The V-Dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data. V-Dem Institute Working Paper 21.

Pettersson T (2022) UCDP dyadic dataset codebook v22.1. Uppsala Conflict Data Program.

Przeworski A (2004) Institutions matter? *Government and Opposition* 39(4), 527-540.

Sarkees MR and Wayman F (2010) *Resort to war: 1816-2007*. Washington: CQ Press.

Skaaning SE (2018) Different types of data and the validity of democracy measures. *Politics and Governance* 6(1), 105-116.

Skaaning SE, Gerring J and Bartusevičius H (2015) A lexical index of electoral democracy. *Comparative Political Studies* 48(12), 1491-1525.

Weidmann NB (2022) Recent events and the coding of cross-national indicators. *Comparative Political Studies* 57(6), 921-937.

Zhao Y (2012) *World class learners: educating creative and entrepreneurial students*. Thousand Oaks: Corwin Press.

Acknowledgements

We are thankful for the editors and reviewers' input that contributed to improve this manuscript. Special thanks to Marcus Österman and Sophie Mainz for suggesting to write additional detailed guidelines for data creators as well as to Jane Gingrich, Svend-Erik Skaaning, and Susanne Garritzmann for their critical feedback on an early version of this manuscript. We also appreciate the input of the participants of the 2024 EPSA conference and the Practices of 2024 Comparative-Historical Analysis Conference at Aarhus University. The Historical Education Quality Database has been funded by numerous centers and programs at Stanford University, including the King Center on Global Development, The Europe Center, SEED, and the Vice Provost for Graduate Education. The Varieties of Indoctrination (V-Indoc) was generously funded by a European Research Council Consolidator Grant "Democracy under Threat: How Education Can Save It" (DEMED) (grant number 865305). The Education Policies and Systems across Modern History

(EPSM) has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 863486).

Online Appendix

Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets

Adrián del Río,[1] Wooseok Kim,[2] Carl Henrik Knutsen,[1] Anja Neundorf,[2] Agustina Paglayan,[3] & Eugenia Nazrullaeva[4]

Appendix A contains an expanded description of the three datasets on education practices and policies. Here we also briefly discuss challenges that were encountered during the data collection.

In Appendix B, we describe the operationalization of the Education Centralization Indices and show how different measurements can produce different substantive implications. We use the education centralization indices in the EPSM and V-Indoc datasets to illustrate this point.

Appendix C assesses whether coding divergences between V-Indoc and the rest of the datasets are driven by the V-Indoc's number of coders employed and coders' self-reported uncertainty.

Appendix D expands our guidelines in the article's Table 2 with detailed suggestions for data creators and users, examples of how to implement them, and discussions of their potential benefits. The appendix, for instance, contains a set of more detailed guidelines for constructing codebooks.

Appendix E provides specific details, including original scales, on how indicators from the different datasets are harmonized before they are compared.

Contents

[1] University of Oslo
[2] University of Glasgow
[3] University of California, San Diego
[4] University of Konstanz

Appendix A. Expanded description of the three datasets on education practices and policies

The appendix provides a description of the three datasets' methodologies, sources, and content. This description is not exhaustive, and we recommend that readers consult the authors' papers introducing the dataset to obtain fine-grained information. We also provide some reflection on the challenges we encountered while collecting each of the datasets. We discuss lessons learned and our suggested guidelines on how to construct academic datasets in Appendix D in more detail.

A.1. Education Policies and Systems Across Modern History (EPSM)

The Education Policies and Systems across Modern History (EPSM) dataset (Del Rio, Knutsen, and Lutscher 2024) is an in-house coded dataset, with the core team of coders comprising the first author and four research assistants. The first version of EPSM covers 145 countries with current populations exceeding 1 million inhabitants, and the time series for each polity follows the V-Dem time series. This implies that the longest time series in the dataset extends back to 1789 and that several countries are also coded during their colonial period and not only as independent states.

EPSM contains four groups of variables about education policies and systems—mainly for primary and secondary education—with a particular focus on political control, of different kinds over the education system. The four groups of variables register (1) characteristics of compulsory education, (2) the existence and character of courses with ideological content, (3) school autonomy with regards to, e.g., funding and operation, and (4) existence and control over teacher training. In total, this yields 21 (typically multi-category) questions. Most of these questions pertain to *de jure* instead of *de facto* characteristics of education policies and systems, but the authors draw on an extensive number of sources to code the different variables beyond legal texts, including reports from governments and international organizations (e.g., UNESCO-International Bureau of Education) as well as secondary source material in the form of scholarly articles published in various social science disciplines, history books, and books on the country's education system.

Regarding the coding process, the authors undertook several measures to enhance measurement validity and reliability. These included an intensive training scheme, with feedback and multiple rounds of trial coding for the RAs, and regular communication and feedback within the team. RAs coded countries where they have language expertise. This expertise was used to read data sources in Spanish, English, Portuguese, Russian, Italian, French, Norwegian, German, Danish, and Swedish. When language limitations occurred, we relied on a combination of automatic translations and consultations with country experts. These country experts, whom we also contacted when there were very divergent interpretations of a case among coders within the original team, helped us identify relevant data sources and correctly interpret legislation in difficult cases.

In addition to extensive clarifications in the codebook and double-checking of all coding by a second coder, the team developed a Rules-of-thumb document to—after team discussion of

difficult cases—codify various coding decisions that may be recurring and clarify what heuristics should be used in different types of cases. The goal of this extensive document was to enhance within-team intercoder reliability since all coders would rely on the same explicit heuristics (instead of different implicit ones for each individual). The document presumably also contributed to reducing coding time (by providing templates for the resolution of tricky coding decisions) and to documenting the rationale behind coding decisions for end users. Among other tools used to enhance reliability and transparency, the EPSM comes with lists of sources as well as rough coder-provided estimates of uncertainty and coding notes at the country-year-question level. Altogether, EPSM took more than 3,150 hours to code, typically 19-22 hours per country, excluding time consulting experts, team coordination, and training.

A.1.1. Challenges during data collection

During the data collection, we had to adjust and change strategies to avoid some problems that might have undermined the dataset's quality. These changes of plans concerned some issues related to (1) the need for expertise – i.e., training of and communication between coders; (2) issues with within-country heterogeneity; and (3) difficulties with ensuring comparability across education systems.

Expertise

It was more difficult than anticipated to codify data in a reliable, comparable, and effective manner. More specifically, follow-up and quality checks of pilot coding showed that we had underappreciated how hard it would be for the RAs, for instance, to assess the quality of data sources, interpret certain (ambiguous) terms, distinguish de jure vs de facto aspects of education systems. To address this challenge, we invested significant time in training the RAs and were in constant communication with them to monitor their decisions to avoid mistakes being carried over from one case to the next.

Another issue we faced was the turnover of RAs. RAs who stayed for a short period tended to produce less reliable and valid coding and, e.g., relied more on secondary sources. Multiple checking was needed, delaying the data collection process. In this case, it is crucial to obtain a highly qualified supervisor who has in-depth experience with all aspects of the data collection and can monitor and engage in dialogues with junior coders, which was key to centralized questions, coding, and data sources.

Country as a unit of analysis

Our unit of analysis entailed making hard decisions on what to do when we observe multiple education systems in a given country. For example, we found that the content of education (and duration) was different between rural/urban schools, girls/male schools, or other groupings or the curriculum content was different across regions in a given country (e.g., India, US). We initially

tried to handle this problem by clarifying which unit to consider (e.g., a plurality of schools) in such circumstances and expanded these criteria as we went along and detected ambiguous cases. However, such a strategy still leads to a loss of information and truncation of relevant variation in the dataset.

Comparability of education systems

The duration of primary and secondary education changes over time and differs across countries. This makes it challenging to make individual-level inferences/claims about who is exposed to what education system and what its effects are. We, therefore, use the ISCED scale to make the levels of education as comparable as possible, but still – if we are, e.g., testing specific hypotheses about exposure to indoctrination at particular ages – we might not safely assume that civic education courses registered under "secondary education" in our dataset took place, e.g., when the kid was 16, across countries and over time (even when our coding is identical).

A.2. Historical Education Quality Database (HEQ)

The Historical Education Quality Database (HEQ) is an ongoing effort led by Paglayan to construct a database that enables us to compare the quality of primary education provision across countries and over time. The database contains time-series country-level measures of student learning going back to 1870 and, notably for this paper, country-level data on the content of primary school curriculum policies and teacher training and recruitment policies.[5] The data collection process for HEQ relies exclusively on primary sources: all national laws, decrees, regulations, or guidelines affecting the curriculum or teacher training and recruitment in primary schools were assembled beginning with the first year when a country's national government began to regulate the curriculum, teacher training, or teacher recruitment, all the way up to 2015. For example, for Germany (Prussia), the data on curriculum policies began in 1763, and the data on teacher training policies began in 1748.

The primary sources that form the basis for HEQ are assembled by one or more expert historians or economic historians who specialize in the history of curriculum and/or teacher policies of a specific country. Experts use primary sources to provide answers to a standard questionnaire developed by Paglayan. The questionnaire on national curriculum policies asks, for example, what subjects are listed in the curriculum, how much time ought to be allocated per subject according to the curriculum, whether any topics or subjects are banned, and who has the authority to select school textbooks. The questionnaire on national teacher training and recruitment policies asks, for example, what type of education degree (if any) is required to becoming a primary school teacher, what are the criteria for admission to a teacher education program, what is the content of teacher

[5] The focus is on national policies, but for countries that lack national policies, the database provides information on the policies that apply to the most populous subnational unit.

education programs, what is the length of teacher training, and whether teacher certification is renewal or for life. When no national curriculum or teacher policies exist, and all such policies are set at the subnational level, experts are instructed to focus on documenting the policies of the most populous subnational jurisdiction.

After consultants submit the primary sources and their initial responses to the questionnaire, a quality assurance manager ensures that the corresponding primary sources substantiate each of the responses to the questionnaire and that there are no internal inconsistencies across related questions. The data collection process takes 6 to 12 months per country, including the period of quality assurance. Given this process's in-depth and time-consuming nature, the database currently includes complete information for five countries: Argentina, Chile, Italy, Spain, and Germany.

A.2.1. Challenges during data collection

To date, the process of assembling the HEQ Database has encountered three main challenges: (a) ensuring comparability across countries; (b) recruiting expert historians; and (c) assuring the quality of experts' responses. The following paragraphs provide additional information about what these challenges entailed and how the HEQ Database has addressed them.

Ensuring comparability across countries

The key challenges here, which are common when assembling cross-national datasets on education systems, are (a) what to do when the national level is not the only level that regulates primary education (e.g., federal countries or countries where some education policy decisions fall under the purview of subnational authorities), and (b) what to do when the definition of "public school" or other key terms varies across countries.

There are numerous possible ways to deal with the first challenge, none of which is *a priori* better than the rest. Reflecting on the goals of data collection and the ways in which the team hopes the dataset will be used are important for making decisions about how to proceed. In the case of the HEQ Database, given that the goal was to document *national* policies, the decision was made to document subnational *de jure* policies only in those cases when the national authority had no say. For example, in cases where the curriculum is regulated by both a national authority *and* subnational authorities, the HEQ Database codes only those curriculum policies that are set by a national authority. When decision-making authorities are *fully* delegated to subnational authorities, the ideal approach would have been to document *de jure* policies in all subnational units, and then obtain a weighted average of what policy looks like at the national level. However, the level of resources allowed the HEQ team to document the education policies of only one subnational unit within the country. In these cases, the decision was made to document *de jure* policies of the *most populous* subnational unit.

With respect to the second challenge, piloting the questionnaire that experts were asked to complete was crucial for identifying terms whose definitions might vary across countries. Subsequently, the HEQ team provided clear definitions of any such terms (e.g., "public school," "primary school," "teacher certification") and instructed experts to complete the questionnaire using these definitions instead of the definitions used in their home country. In addition, the HEQ core team was quick to respond to experts' queries on how to adapt local definitions to the definitions used by the HEQ Database for comparability across countries.

Recruiting expert historians

Identifying individuals who had not only the expertise but also the time and willingness to work as paid consultants for HEQ was more challenging than expected. The HEQ team's approach was to submit a "Call for Experts" search among individuals who had a strong publication record on the history of education policies in reputable English or Spanish journals (the two languages that the core team was fluent in). A few responded by expressing interest and submitting an application. Some indicated their lack of availability but suggested other experts (e.g., colleagues, PhD students writing their dissertation on the history of education in a given country, etc.). Many did not respond at all, and some responded wondering whether the email they had received was spam. The size of this last group made it clear that tapping into shared professional networks was crucial to lend credibility to the HEQ initiative. The recruitment of experts proceeded more smoothly once the HEQ team was introduced to expert historians by a third professor who knew them both.

Assuring the quality of experts' responses

Many cross-national datasets that claim to code *de jure* policies (e.g., on suffrage, electoral rules, etc.) do not specify which specific laws or regulations underlie each coding decision, much less provide a copy of these laws/regulations. In an effort to assure the quality of the data contained in the HEQ Database and ensure transparency and replicability in the social sciences, the HEQ team made the decision early on to require expert consultants to (a) indicate the source (i.e., law/regulation) underlying *each* response, and (b) submit PDF copies of these sources. As part of its quality assurance process, the HEQ core team then checked whether each of the responses submitted by the expert was, in fact, consistent with the corresponding law/regulation. This was an onerous process that exceeded the core team's capacity. To facilitate the core team's work, the decision was made after the piloting stage to request that experts not only indicate the name of the relevant law/regulation and provide a copy of it, but also, that they indicate the specific article/page number within the law/regulation that validated their response. While this helped reduce the amount of time needed by the HEQ core team to validate experts' responses and improved the transparency of the dataset, the challenge was that it increased the amount of work on the part of experts, making it somewhat more challenging to identify experts who were willing to work as consultants for HEQ.

A.3. Varieties of Indoctrination (V-Indoc)

The Varieties of Indoctrination (V-Indoc) data (Neundorf et al., 2023) is an expert-coded dataset that covers up to 160 countries and provides annual data from 1945 to 2021. V-Indoc aims to measure state-led indoctrination via education and the media. V-Indoc collected data on diverse aspects of public-school education, including centralization, the autonomy of teachers as well as politicized hiring and firing of teachers, the democratic (autocratic) content of the curriculum, the promotion of patriotism in schools and through educational content, and the extent to which the curriculum emphasizes the teaching of politics and ideology. Most of these questions explicitly refer to de facto practices in education and therefore require some subjective assessments by the expert coders.

V-Indoc follows the data generation process established by the Varieties of Democracy (V-Dem) project. The structured expert survey consists of 27 questions (21 on education), and responses were entered via the V-Dem online platform. The survey was typically coded in English, but experts could also choose Arabic, French, Portuguese, Russian, and Spanish translations, which is a standard V-Dem practice (see Coppedge et al. (2023) for more discussion of the methodology). Questions typically included a clarification text to explain the key concepts of each question. Responses were provided as ordinal options, e.g., on a scale from 0 to 3. Experts then provided annual ratings for each question for the country of their expertise using a customized coding grid and web platform. The experts were further asked to provide a certainty rating of their responses. Indicators were subsequently aggregated into 13 indices to measure abstract concepts capturing the politicization of education and the media.

To ensure the success of this data collection, numerous experts were recruited to code every country.[6] The assumption is that every single coder's ratings might be somewhat biased or uncertain. To address these concerns, we follow the V-Dem project (see, e.g., Knutsen et al., 2023, pp. 14-5) in designing particular items that could be less prone to general bias and using ordinal response scales with specific definitions for each category; different categories are aimed to serve as 'benchmarks' and facilitate coding. We have also designed and asked experts to code anchoring vignettes that have helped us standardize how experts code in general. Finally, we have used V-Dem's measurement model to correct for both variations in expert reliability and scale perceptions (Pemstein et al., 2020).

To maximize the number of country experts, the research team (including six research assistants) reached out to 24,000 education experts worldwide in 2021. More than 1,400 experts expressed interest in participating in the survey. We then conducted an expert vetting process and fielded the final survey from January to May 2022. The 760 vetted experts completed the survey for which they were compensated. The median number of coders per country-year is 5, with 1 coder as the

[6] We further used so-called "bridge coders", which coded multiple countries. This data is used to calibrate the responses across different countries.

minimum (e.g., for Angola, Burkina Faso, Bolivia, Gambia) and 20 coders as the maximum (Brazil and the United States). Overall, the development of the questionnaire, expert recruitment, data collection, and running of the measurement model took two years to complete.

A.3.1. Challenges during data collection

Key definitions

The V-Indoc project relied on providing detailed definitions for education-related terms, which can vary in meaning across countries, or where the meaning may change over time. Examples of these challenging terms include "formal public education", "primary / secondary education", or what to do about decentralized education systems. To ensure that expert coders remain as consistent as possible when codifying the data, definitions for these concepts were included as part of general instructions to expert coders, which were shown at the start of the expert survey. Of course, a major limitation is that we could not guarantee that all experts would strictly adhere to these guidelines when answering survey questions. We also included question-specific instructions where appropriate.

Designing the codebook

Designing the codebook – survey questions and answer categories – was another challenging process. We were guided by the V-Dem team, which had a lot of experience in designing expert surveys, and we tried to follow the best practices that they suggested. We could not simply ask experts whether a country's regime is involved in indoctrination. Therefore, we had to consider different (uni)dimensions where indoctrination might be evident and develop questions that focus on specific school subjects as proxies, such as history, social sciences, and language education.

In addition to unidimensional questions, we needed to design answer categories based on a four-point Likert scale (e.g., rarely, sometimes, often, extensively). Most variables in the V-Dem data are coded using a five-point scale. However, after multiple discussions, our team decided that adding another answer category would often be an artificial solution. We struggled to distinguish cases that fall in the middle categories with a five-point scale, and experts would likely face the same difficulty.

We had one question where we needed to provide a threshold value for questions that had just two (0 or 1) answer categories: what proportion of instructional weekly hours in the curriculum is dedicated to mathematics and natural sciences (V-Indoc Codebook 2024, p. 30). Our challenge was to infer the correct threshold that would work across countries and over time since 1945. We used the available data on the curriculum from Benavot (2004) to calculate a threshold value of 25%. However, it turned out to be too low, and all experts' answers were consistently biased upward.

Expertise

The quality of expertise is crucial in the case of V-Indoc data. For our project, we had to recruit a new pool of education experts. We used Qualtrics to distribute an online expression of interest form to experts. In this form, we asked experts to provide basic information: their email, institutional affiliation, list of publications, information about their website (if any), highest educational degree, current position, as well as the area(s) of their expertise in education (e.g., the main country of expertise and the second country of expertise, the time period(s) they focus on).

We used three main channels to recruit potential experts. First, with the help of research assistants, we consulted the ratings of top universities in each country and collected emails of all faculty members (research and teaching focused), postdoctoral scholars, and graduate students whose research expertise is in the field of education. Second, we used Google Scholar to find academic journals, books and book chapters, policy reports, and regional education conferences, and we collected emails from the authors/participants. Third, we contacted education-related NGOs and policy experts outside of academia, asking them to circulate our call among their network. From July 2021 to February 2022, we reached out to 24,000 education experts from around the world. More than 1,400 experts responded to our call and expressed interest in participating in the expert survey. The final V-Indoc survey was taken by 760 experts out of 1,400. One lesson here for us was that the response rate was quite low, around 6%. With the help of research assistants who possessed a background in comparative education, the list of experts was vetted according to modified V-Dem expert criteria.

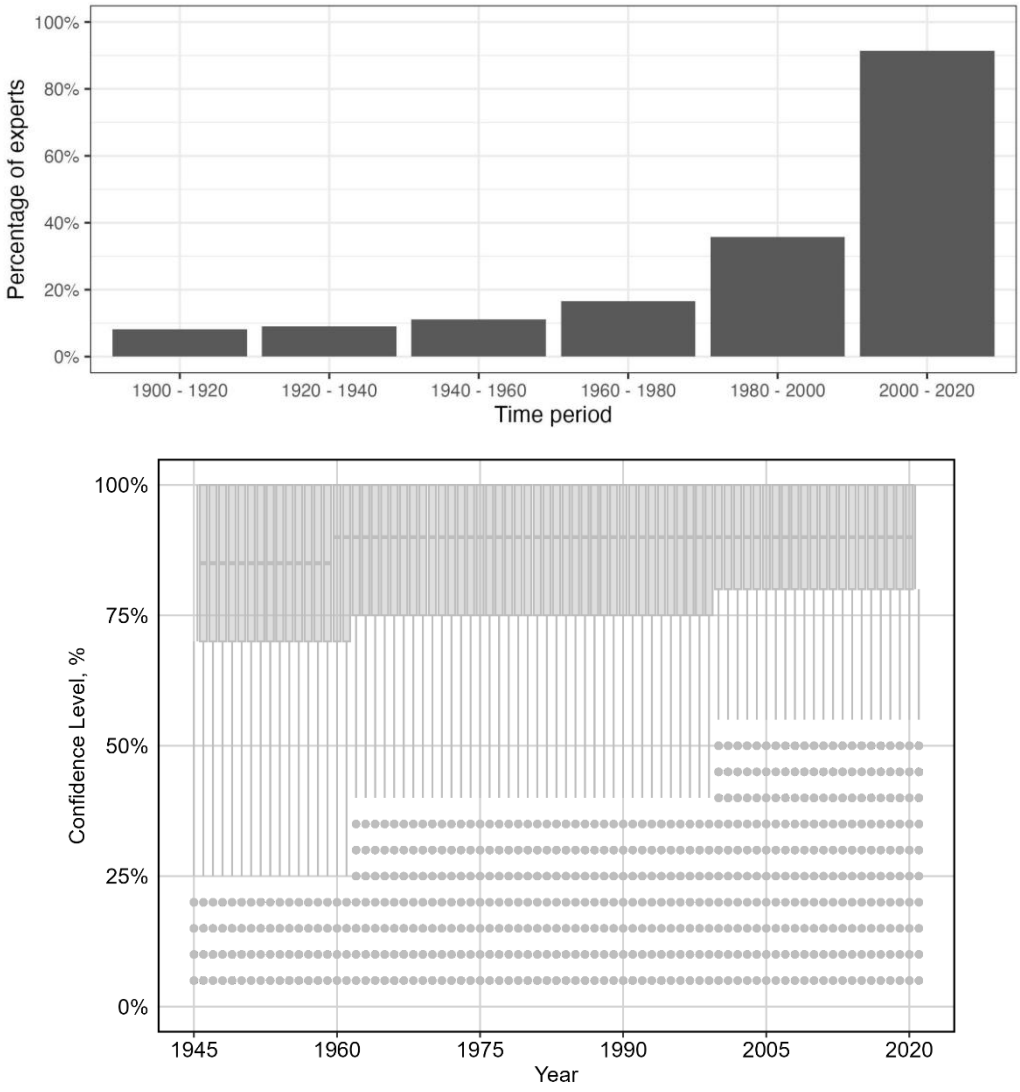
However, the information we could collect about the experts was quite limited, and we could not always guarantee their expertise in each case (e.g., based on their education and/or CV, without a detailed list of publications). Additionally, the levels of uncertainty that experts provide with each answer do not always reflect their expertise. We encountered instances where experts could give factually incorrect answers (based on validation) yet expressed high confidence in their responses.

Coding back in time

As part of our online expression of interest form, we asked experts about their expertise and how far back in time they believed they could answer questions about education systems in the countries they know well. Figure A1 below (top panel) reflects the experts' answers based on a recruitment sample of 1,400 individuals. It shows that most experts are confident in coding the most recent years since the 2000s, but their confidence decreases the further back in time the questions go. The original idea for V-Indoc was to start coding as early as 1900, but this evidence suggests that even coding as far back as 1945 might require a separate pool of experts who specialize in the history of education. Figure A1 (bottom panel) also shows the overall distribution of expert coders' confidence levels for all questions in the V-Indoc education data, based on 760 experts who participated in the final expert survey [Neundorf et al. 2023]. Although the median levels of confidence are quite high, there are some outliers where coders reported low confidence (in some

cases below 20%). Additionally, the interquartile range of confidence has changed over time, reflecting the same pattern observed at the recruitment stage. Coders tend to be more confident in coding data from more recent years, particularly post-1965 (when more data on education became available) and from the mid-1990s onward.

Figure A1. Experts' confidence levels to code education-related questions over time at recruitment *(top panel)* vs while coding *(bottom panel)*



Notes: The top panel plots the proportion (%) of experts who indicated at the recruitment stage whether they could answer questions about certain time periods (answers categories were presented as these time periods). The bottom panel plots the distribution of experts' answers about their confidence levels for all questions over time. The boxplot shows the median levels, as well as the interquartile range (bottom and top 25%), as well as some outliers.

Source: V-Indoc data at the recruitment (top panel) and final (bottom panel) stages [Neundorf et al. 2023].

Appendix B. How Data Collection Methods Affect Inferences: Evidence from the Effect of Democratization on Education Centralization

In this section, we estimate the average treatment on the treated (ATT) effect of democratization (as defined by Boix-Miller-Rosato) on two indices of education centralization, one from EPSM and the other from V-Indoc, to illustrate how differences in the way concepts are measured can produce different substantive implications.[7] EPSM's education centralization index is coded as 0 when the country does not have a department of education at the national level. In cases where such authority exists, the index is the interaction between two indicators (both re-scaled to a unit interval) that measure the extent to which (1) the curriculum is centralized under the national government and (2) the state operates and funds primary and secondary education. The V-Indoc education centralization index is constructed using indicators of the centralization of the (1) curriculum and (2) textbooks, which are based on ordinal expert survey responses that are converted to continuous variables and aggregated into an index using V-Dem's Bayesian Item Response Theory measurement model (Pemstein 2020). The two indices range from 0 to 1, and higher values reflect greater education centralization. To make them more comparable, we use standardized versions of these indices in the following analysis.

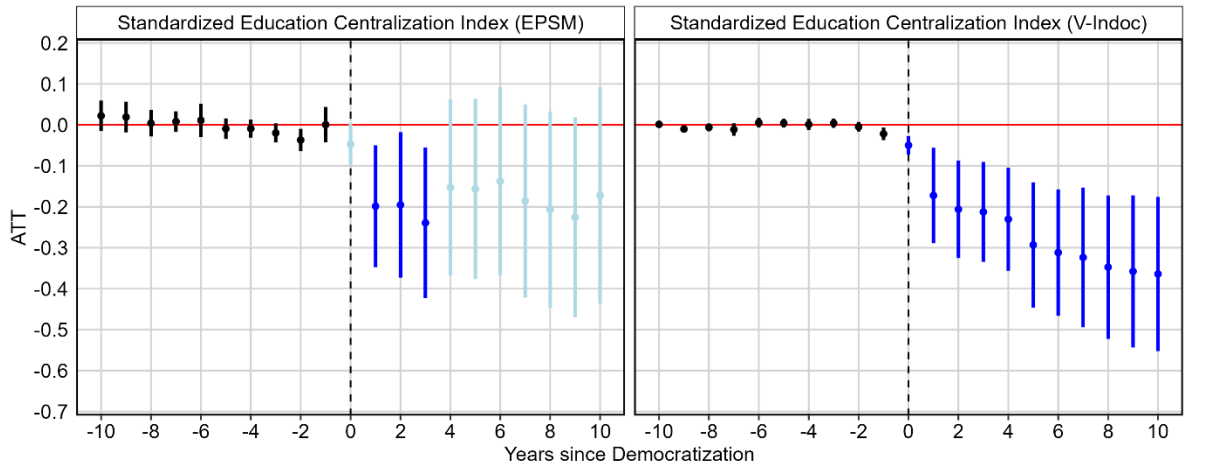
Our estimate of ATT effects is based on the sample of overlapping country-year observations in both datasets (9,551 observations from 144 countries over 1945-2020) and uses the counterfactual matrix completion estimator presented by Liu, Wang and Xu (2022), which directly imputes counterfactual outcomes for treated observations (i.e., democratizers) using information from non-treated observations (i.e., non-democratizers) while accounting for unobserved unit heterogeneity and time-varying confounders to obtain ATT estimates. We also derive the uncertainty associated with these estimates using 1,000 iterations of a non-parametric bootstrap.

Figure B1 plots the estimated ATT effect from this analysis along with corresponding 95% confidence intervals. Even though these indices measure a similar concept (i.e., education centralization), and the ATT estimates are derived from the same estimation method, sample, and measure of democracy, the inferences that we can draw from the two models have different substantive implications: education centralization exhibits a more persistent and more precisely estimated decrease in the aftermath of democratization when using V-Indoc's index (right panel) relative to the EPSM index (left panel). V-Indoc's index suggests that a transition to democracy is a critical juncture during which national governments decentralize educational authority, a pattern that is reinforced as democracies mature. When relying on EPSM's index, we also observe an initial pattern of decentralization after democratization, but decentralization does not intensify over time;

[7] The HEQ centralization index used in Figure 2 is based on the average of two ordinal indicators that range from 0 to 2 and respectively measure the centralization of the curriculum and textbooks, which in turn is re-scaled to a unit interval. The coverage of the HEQ dataset does not enable us to include it in the analysis presented in this Appendix.

on the contrary, the point estimates, while remaining consistently negative, lose statistical significance over time.

Figure B1. ATT Estimates: Democracy and Education Centralization



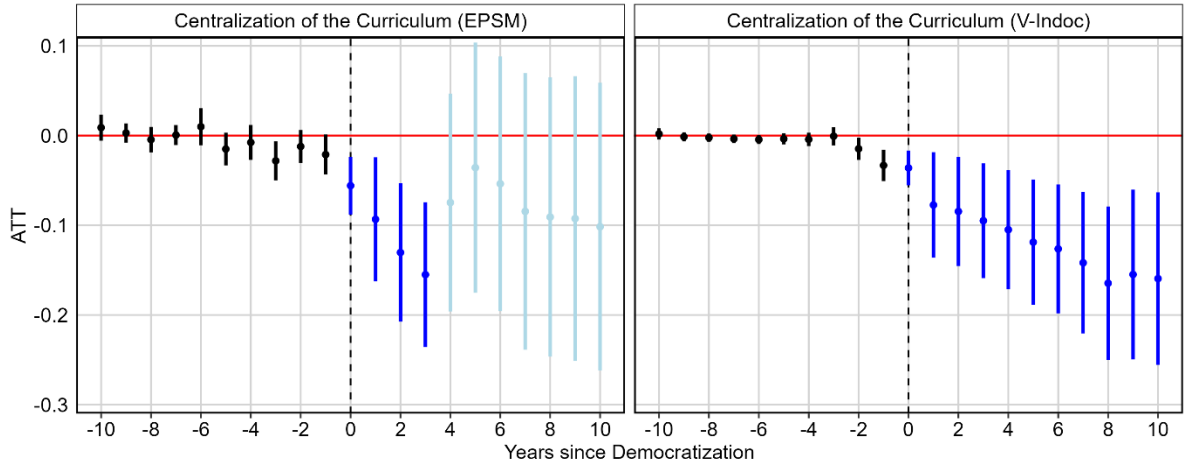
Note: darker (lighter) colors indicate that the ATT estimate is statistically significant (insignificant) at the 0.05 level.

One possibility is that the divergences in Figure B1 reflect the fact that the indices differ in what they measure, as discussed in the main paper. To consider this possibility, we replicate the analysis in Figure B1 restricting the analysis only to the respective *centralization of the curriculum* indicators used to construct each index. In the case of the V-Indoc indicator, the measurement model also generates supplementary variables that map continuous indicators back to their original ordinal scale. We use this ordinal version of the V-Indoc centralization of the curriculum indicator, and further harmonize the EPSM and V-Indoc indicators so that their values map onto the same ordinal scale, i.e., the official curriculum is set by (0) no national authorities, (1) both sub-national and national authorities, or (2) only national authorities. With this harmonization, the correlation between the two indicators is relatively high (r=0.60). Nonetheless, Figure B2 shows that the subsequent ATTs of democratization across these indicators generally align with the patterns observed in Figure B1.

Another possible explanation for the difference in results in Figure B1 is that V-Indoc aims to capture *de facto* education centralization whereas EPSM registers *de jure* centralization. In the paper, we discussed cases where trends in education centralization clearly diverged, depending on whether one considers the law or education practices. Previous work has highlighted how legislation and formal institutions (even constitutions) may sometimes survive democratization processes (Albertus and Menaldo 2018). It is, for example, possible that some countries that retain legislation that implies a more centralized education system on paper after democratization nonetheless devolve increasing curricular responsibility and/or other education choices to lower

levels of government - as was discussed in the case of Argentina. If several democratizers follow such a pattern, this might explain the (modest) differences in point estimates and larger confidence intervals for the EPSM regressions. This is only one potential (and admittedly speculative) interpretation of the differences in results. We note that the large confidence interval for the EPSM regressions might also imply that we would be making a type 2 error by concluding that there is no effect of democratization (the point estimates are consistently negative, and standard errors are relatively large for the longer lags). Nonetheless, this application serves to illustrate the broader point that choice of measure may impact the substantive conclusions drawn about the causes of education system features.

Figure B2. ATT Estimates: Democracy and Curriculum Centralization



Note: darker (lighter) colors indicate that the ATT estimate is statistically significant (insignificant) at the 0.05 level.

Appendix C. Uncertainty and the number coders

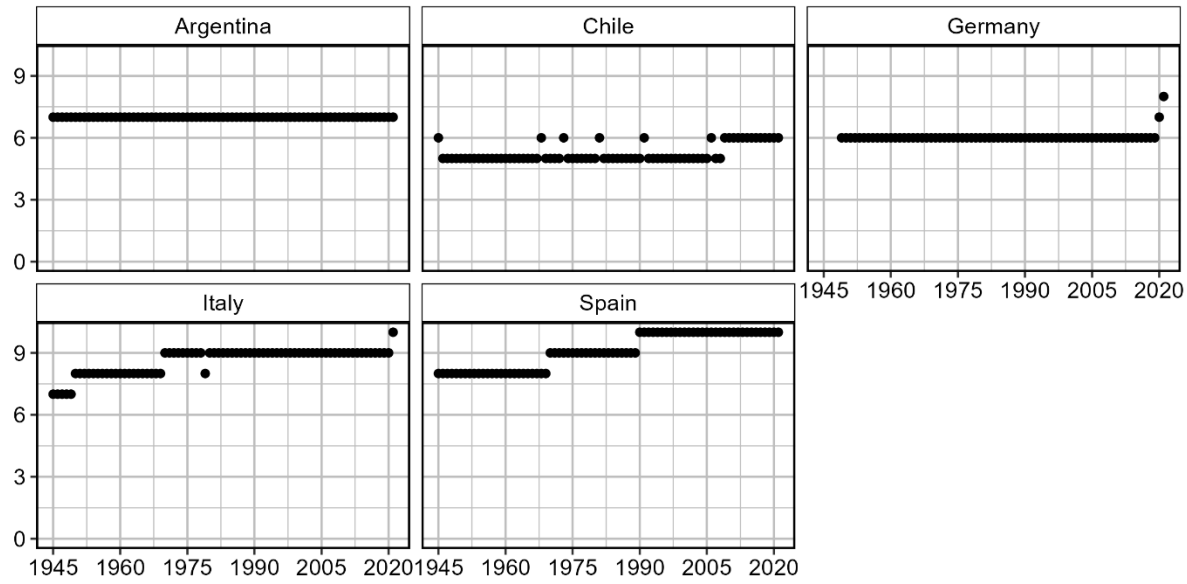
This section assesses whether coding divergences between V-Indoc and the rest of the datasets are driven by the V-Indoc's number of coders employed and coders' self-reported uncertainty. Figure C1 shows V-Indoc used between five and eleven coders, which ensures having a number of responses to implement the V-Dem's measurement model and correct for both variations in expert reliability and scale perceptions (Pemstein et al., 2020).

In addition, Figures C2a-c plot the levels of coders' uncertainty when coding education centralization, teachers' training, and the content of civic education. Overall, the figure shows that most coders are highly certain about their coding decisions. Nevertheless, few cases gather substantial variation in the levels of coders' confidence, especially in Argentina and Chile.

Furthermore, higher levels of uncertainty do not align with instances where the article found divergent patterns between V-Indoc and the rest of the datasets. For example, the three datasets coded similarly the level of education centralization in Argentina after the late 1980s, which is the starting period where coders are more uncertain about their coding decision –coders' confidence ranges from 55% confidence to 100%.

In sum, this section's findings show that we are skeptical that coding uncertainty and the number of coders affect the article's descriptive inferences.

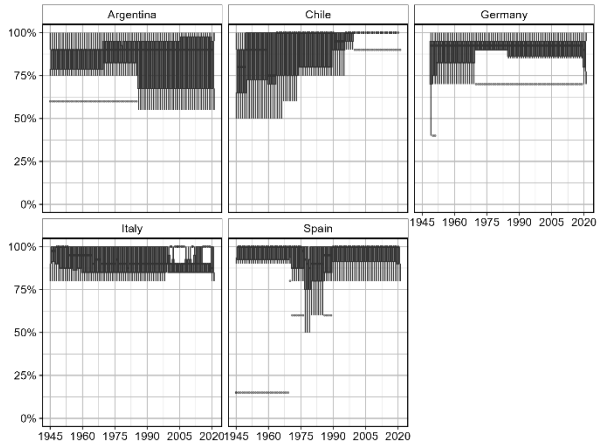
Figure C1. Number of coders in V-Indoc for five countries over time



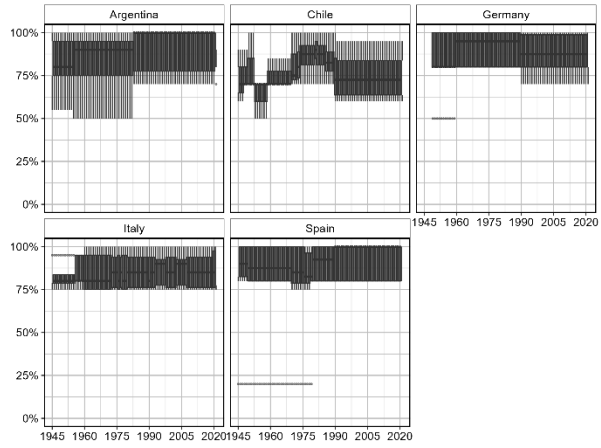
Note: For each of the countries in the comparison group, V-Indoc includes more than three coders. Calculations are based on the coder-level dataset.

Figure C2. Distribution of V-Indoc coders' confidence levels

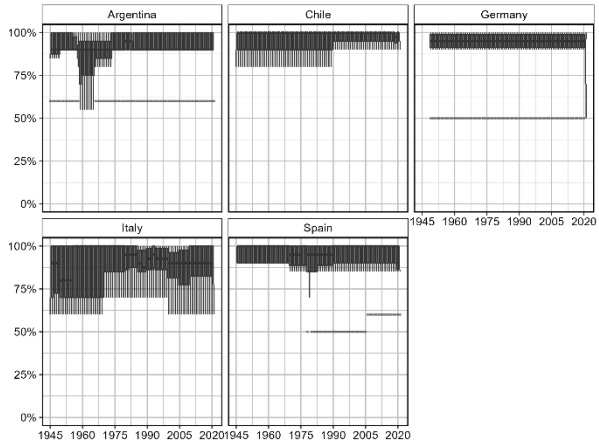
A. Centralization of the curriculum (Fig 2)



B. Politicized teacher hiring (Fig 3)



C. Religious content (Fig. 4)



Note: For each of the countries, we plot the distribution of coders' self-reported confidence levels (percent) over time. This is a box and whisker plot, which shows the interquartile range (25th and 75th percentiles) as well as the median of the distribution in each year. We can see that confidence levels are high for the experts who coded this question for Germany, Italy, and Spain (except for higher uncertainty between the late 1960s and the late 1980s), while there is quite substantive variation in confidence levels of coders for Argentina and Chile.

Appendix D. Guidelines for dataset creators and data users

This section provides a checklist for the construction and use of datasets. These recommendations are the product of our experience coding historical datasets, especially in education. We do not claim these are *the best practices* that everyone should follow to produce valid and reliable datasets. Not everyone possesses the resources and time to follow all these suggestions when producing datasets.

Instead, we intend to increase awareness among data users and makers about a number of issues during the data collection process and methods employed, as well as provide potential solutions. We invite data creators and users to use the checklist to discuss various measurement challenges, trade-offs, and resulting specific advantages or disadvantages as clearly and transparently as possible. Table 2 in the paper provides an overview by summarizing our guidelines, and Appendix D1 and D2 provide more detailed suggestions from each guideline, examples of how to implement them, and discussions of their potential benefits.

D1. Guidelines for dataset creators

Our suggestions for dataset creators focus on the following five areas: codebooks, intercoder reliability, triangulating sources, having data sources/coding type tailored to concepts, and reporting uncertainty. We have also provided examples of how we (or other dataset creators) have implemented the suggestions and their expected benefits to help readers implement the advice. Not all suggestions will work for some topics or source types.

Guideline 1. As part of the codebook precisely defines key concepts, their dimensions, and the measurement level/scales of all variables.

- *Example*: Why do you use a specific conceptualization of, let's say, education centralization over another? And why do you focus on some dimensions over others? The Varieties of Democracy Codebook (Coppedge et al. 2022) is exemplary in this particular regard, and we refer, e.g., to the introduction of V-Dem's five main democracy indices at the top of the codebook. Yet, codebooks should, all else equal, be concise and easy for users to read and navigate (and V-Dem's codebook is not exemplary in terms of its length). Thus, providing additional documentation, and especially dealing with more involved or nuanced conceptual discussions in, say, dataset papers or other documents may be a good strategy for balancing brevity versus detail in codebooks.

- *Benefits*: Explicit documentation of definitions and conceptual choices helps clarify the theoretical assumptions that measures draw on (and thus aid interpretation also of the measures), enhances transparency about the phenomenon under study, and makes it easier for users of data to compare your new measures with existing ones. Conceptual groundwork may also guide and ease downstream efforts on how to best frame question items and make

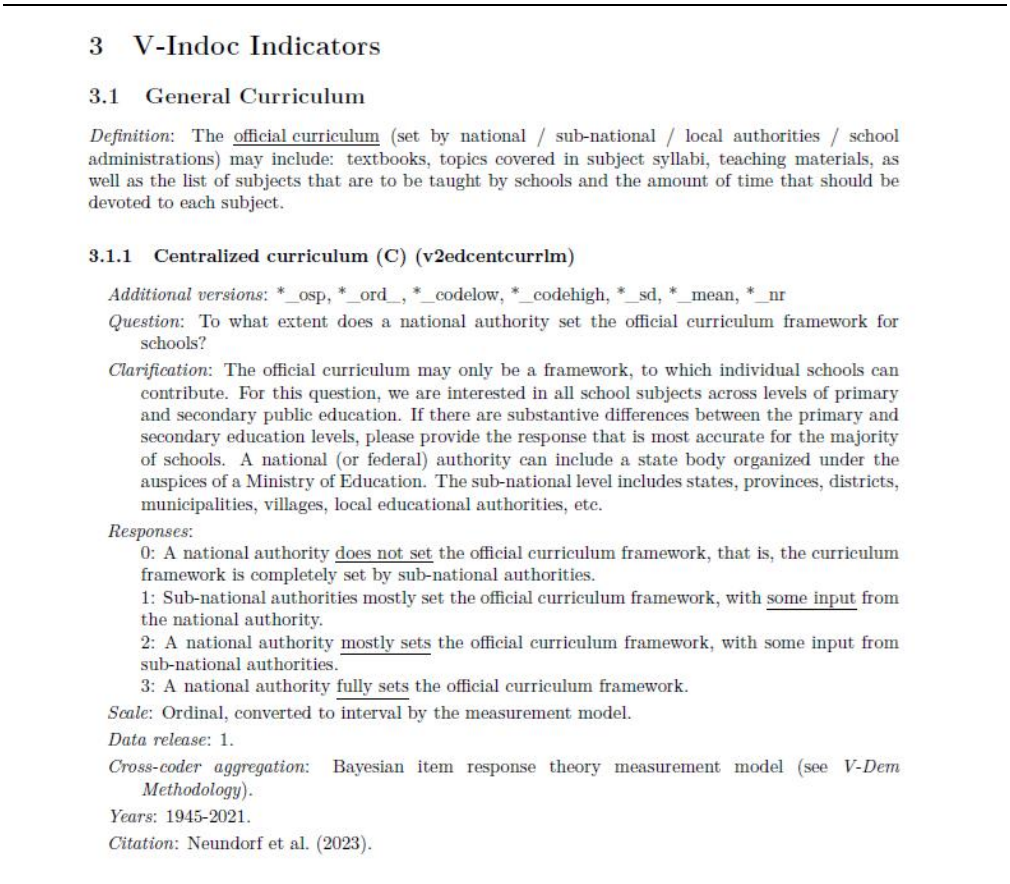
explicit several operational assumptions that may be laid out in the codebook or elsewhere (e.g., a rule of thumb document).

Guideline 2. Specify questions to be coded as much as possible and add clarifications to the main questions if multiple interpretations are plausible, especially if there are concepts that might be well-known for researchers but not for coders or students. Other information about the scaling and, if applicable, estimation method is also informative and worth including.

- *Example*: Figure D1 illustrates this suggestion.

- *Benefits*: Specifications and clarifications help reduce ambiguities regarding the interpretation of concepts and related measures, as well as items' categories. This has benefits both for correct interpretation by dataset users and for inter-coder reliability and internal consistency (when multiple coders code different parts of the dataset).

Figure D1. Example of codebook (V-Indoc) - Clarifications



3 V-Indoc Indicators

3.1 General Curriculum

Definition: The official curriculum (set by national / sub-national / local authorities / school administrations) may include: textbooks, topics covered in subject syllabi, teaching materials, as well as the list of subjects that are to be taught by schools and the amount of time that should be devoted to each subject.

3.1.1 Centralized curriculum (C) (v2edcentcurrlm)

Additional versions: *_osp, *_ord_, *_codelow, *_codehigh, *_sd, *_mean, *_nr

Question: To what extent does a national authority set the official curriculum framework for schools?

Clarification: The official curriculum may only be a framework, to which individual schools can contribute. For this question, we are interested in all school subjects across levels of primary and secondary public education. If there are substantive differences between the primary and secondary education levels, please provide the response that is most accurate for the majority of schools. A national (or federal) authority can include a state body organized under the auspices of a Ministry of Education. The sub-national level includes states, provinces, districts, municipalities, villages, local educational authorities, etc.

Responses:

0: A national authority does not set the official curriculum framework, that is, the curriculum framework is completely set by sub-national authorities.

1: Sub-national authorities mostly set the official curriculum framework, with some input from the national authority.

2: A national authority mostly sets the official curriculum framework, with some input from sub-national authorities.

3: A national authority fully sets the official curriculum framework.

Scale: Ordinal, converted to interval by the measurement model.

Data release: 1.

Cross-coder aggregation: Bayesian item response theory measurement model (see *V-Dem Methodology*).

Years: 1945-2021.

Citation: Neundorf et al. (2023).

Note: Neundorf et al.'s (2023) codebook

Guideline 3. Include one item for each concept's dimension, and if one dimension is already complex, try to break it up into two or more questions (or more categories in categorical questions, when relevant).

- *Example*: Education centralization is a broad concept with multiple dimensions (see section 4.1). EPSM, for example, included funding of schools (two indicators with 10 categories each), curriculum (one indicator, four categories), management of education (one indicator, three categories), and teachers' ideological training (which results from using three categories from two variables, teacher_training_req and teacher_training_source).

- *Benefits*: Breaking complex questions up into more items helps mitigate multi-dimensionality or conflate outcomes in one variable. It is better to have multiple items measuring one concept than one, especially because you can always combine items later on in different indices, depending on the purpose of research. Breaking up complex questions thus allows for precision, increased reliability, and flexibility of use, without any real downsides due to this possibility of aggregating items later on (other than, perhaps, longer codebooks, larger datasets, and more time used for coding).

Guideline 4. Ask some experts on the topic for feedback on the codebook.

- *Example*: The V-Indoc codebook took two years to be developed and went through several rounds of expert assessment and pilot studies.

- *Benefits*: Identify unclear coding formulations and categories, potential problems to code real-world examples, and include further clarifications.

Guideline 5. Conduct pilot studies, selecting diverse countries based on the expected chances to extract the information needed.

- *Example*: All three datasets conducted pilot studies with real-world examples (see Appendix A). But, if resources are very limited, we suggest creating fictional examples and testing whether coders understand the questions well and how to code them. Examples can be created by discussing the strengths and weaknesses of the codebook.

- *Benefits*: A pilot study on a subset of diverse cases can help identify the scope and depth of the information we can extract to create comparable indicators. However, it can also offer rough estimates of the time and resources needed to complete the data collection process. The more cases are coded, the less time-consuming it will be to find and code data sources. Thus, the time employed in the pilot study could be interpreted as a very conservative estimate.

Guideline 6. Ensure that coders have similar understandings of questions and apply similar thresholds when choosing between categories or use strategies to adjust for dissimilarities. Create a protocol or rule-of-thumb document to provide a set of instructions about how the data collection should proceed and what to do when data sources are unclear. Adding examples of how the researchers have handled tricky cases could be useful to understand the dataset's assumptions better.

- *Example 1*: When working with in-house coders, the EPSM team held a workshop to train coders. In this session, three types of exercises were conducted: 1) code vignettes (hypothetical country scenarios) to ensure coders understand the question similarly; 2) code one country (which was already validated) with *pre-selected* data sources to ensure coders interpret data sources similarly ; 3) let two coders code one country to assess how coders select and understand data sources. Comparisons and discussions of diverging results under points 2) and 3) were also helpful for drafting new rules-of-thumb for coding.

- *Example 2*: The EPSM data collection is accompanied by a rule-of-thumb document that gathers some lessons from coding diverse countries, i.e., identify data sources, code uncertainty, how coders' doubts were solved, and provide general rules to navigate difficult scenarios such as a civil war or regime change. We note that this was a much-consulted document throughout the EPSM data collection, and new information and rules of thumb were added throughout the coding process once encountering new and difficult coding decisions and agreeing on strategies and heuristics for resolving this decision (and similar ones that might appear when coding subsequent cases).

- *Example 3*: V-Indoc designed and asked experts to code anchoring vignettes that have helped them standardize how experts code in general. The responses to these vignettes were used in V-Dem's measurement model to correct for both variations in expert reliability and scale perceptions (Pemstein et al. 2020). Figure D2 provides an example vignette.

- *Benefits*: Provide documented information about tricky cases and how these types of cases were supposed to be interpreted and coded if they reappeared elsewhere in the data collection. These features, in turn, enhance reliability, validity and internal consistency in the coding of different parts of the dataset. A rule-of-thumb document available to (and widely consulted by) all coders is one instrument that may be used to ensure that coders (to the extent possible) operate with the same explicit assumption and coding process, instead of relying on several implicit ones (which might not only differ across coders, but fail to align with the research team's conceptual or other assumptions or original intentions).

Figure D2. Example vignette (V-Indoc) for centralization question

To what extent does a national authority set the official curriculum framework for schools?

Responses:
0: A national authority does not set the official curriculum framework, that is, the curriculum framework is completely set by sub-national authorities.
1: Sub-national authorities mostly set the official curriculum framework, with some input from the national authority.
2: A national authority mostly sets the official curriculum framework, with some input from sub-national authorities.
3: A national authority fully sets the official curriculum framework.

Anchoring Vignette:
0-1: In country X, a national authority is largely excluded from the curriculum design process. A national authority sets minimum curricular standards but does not determine the teaching content. Sub-national authorities determine the teaching content, that is, the contents of the syllabi in all core subjects at each grade level.
1-2: In country X, a national authority sets the core curriculum framework for mathematics and languages. Sub-national authorities set the curricular guidelines for the subjects beyond the core curriculum, that is, science, arts, physical education, and social sciences.
2-3: In country X, a national authority specifies the amount of time that should be devoted to each subject and the teaching content, that is, the contents of the syllabi in all core subjects at each grade level. Sub-national authorities are responsible for the selection of school textbooks available to schools and teachers.

Source: Neundorf et al. (2023; 2024).

Guideline 7. Active communication is key if more than one coder is involved in the data collection process.

- *Example*: The EPSM team used a joint web platform to communicate efficiently with the lead coder and resolve minor issues quickly. Coders also sometimes sat physically together when coding, allowing coders to find joint solutions to challenging cases. Part of this communication strategy was to share results and working papers with coders.

- *Benefits*: Speed up the data collection process as a rapid response (1) avoid coders' doubts will lead them to stop coding, (2) avoid mistakes or that mistakes in past coding may contaminate other coding decisions, (3) provide learning of new and efficient strategies for coding and bolster coders' confidence in their own coding decisions as the team creates a safe and stimulating environment, (4) create a team spirit, which is good for coders psycho-social experience with the work and may also make coders more likely to stay around. The side benefits are that costs associated with training new coders are reduced, and content and experienced coders are, we believe, also more efficient ones.

Guideline 8. If possible, have an external coder cross-checking cases and ask for experts' help.

- *Example 1*: The HEQ dataset construction involved a quality checker that ensured coders' responses were sufficiently justified, adding references to exact paragraphs that motivated the coding.

- *Example 2*: The EPSM dataset construction involved a second coder going through all original codings, after a batch of countries were finalized, and discussing and adjusting the coding after that. Also, in several cases, the EPSM team subsequently consulted with country experts on coding decisions and availability and soundness of data sources, prior to making a coding decision. This happened whenever the team was unable to reach a consensus or data sources were scarce.

- *Benefits*: Ensure that coders have a similar understanding of concepts and apply similar heuristics when approaching similar cases.

Guideline 9. Assess the extent to which the dataset has been coded consistently and make transparent the strengths and limitations of the dataset. Assigning cases to coders based on regional, language, or historical period expertise does not necessarily mitigate all biases. Some factors could affect our ability to gather data sources, understand it, and produce a coding decision, and sometimes these correlate with other factors that may be of interest (e.g., democracy, war). We encourage data makers to systematically analyze the extent to which scores are affected by: data sources employed, regions, and historical period coded, number of coders employed, or coders' confidence in the coding decision.

- *Example 1*: When working with human coders, Weidmann (2024) shows that recent dramatic events in a country just prior to the coding have a small but visible impact on coder ratings, but primarily for those variables that are directly related to the observed event. Also, Weidman (2016) shows how news reports can bias estimates on protests.

- *Example 2*: Del Río et al. (2024) show the extent to which the number of data sources to code countries and the levels of coding uncertainty is affected by levels of democracy, economic development, historical periods, state capacity, and regions.

- *Example 3*: V-Dem's web-platform infrastructure allows experts to report the level of coding confidence, encompassing a sliding scale from 0-100 that experts use to express their level of confidence for all observations that they code (experts may mark all or a subset of their years coded for a given question, and then apply the confidence sliding scale; see V-Dem's methodology document for details).

- *Benefits*: Being transparent about these issues also helps future data collection efforts to replicate or expand the dataset, as well as identifying ways to improve coding.

Guideline 10. If possible, use multiple data sources to inform your coding decisions.

- *Example*: The EPSM dataset uses a combination of primary and secondary sources. However, one can also compare primary sources coming from different entities. An example of this data triangulation comes from Paglayan's (2022a) dataset on primary school enrollment rates, covering 42 countries in Europe and Latin America from 1828 to 2015. To construct the dataset, Paglayan gathered information on the number of students enrolled in primary school from at least three different sources per country. When all three sources were aligned, the researcher was confident about the information's accuracy. However, when there were differences between sources, the researcher had to decide which source (if any) was more credible. In this case, Paglayan prioritized the statistics assembled by local historians specializing in their country's education history over statistics contained in cross-national datasets.

- *Benefits*: Triangulating sources could be useful in getting a comprehensive picture of a case and enabling more valid and reliable coding. Inspecting the sources of disagreements (if any) could open a fruitful discussion about the advantages/disadvantages of relying on a type of data source and create alternative measures, such as measures of coding confidence. Moreover, when using secondary data sources, we reduce the risk of missing important information (e.g., secondary literature tends to focus on the most famous education laws and neglect the relevant laws and regulations that, while less famous, formed part of the *de jure* educational landscape).

Guideline 11. Think critically about and discuss the strengths and weaknesses of different types of data sources, before devising strategies for how to search for and use sources. When using secondary sources, keep in mind that authors of these texts might have used different conceptualizations of key concepts or invoked different (explicit or implicit) assumptions. A careful read to detect "red flags" and awareness about how particular concepts and terms are used by authors with different (disciplinary, theoretical, methodological and other) backgrounds are key to detecting such issues.

- *Example 1*: Often, the secondary literature assumes incorrectly that a *de facto* education practice was grounded in a *de jure* policy. For example, we often find different understandings of what compulsory, universal, and free education is (see del Río (et al., 2024, pp. 12-15) for a discussion based on the case of Swedish compulsory education law and Bolivia).

- *Example 2*: Coding *de facto* information requires input from experts instead of RAs. Structured expert surveys are thus a way to collect and codify extensive and comparable cross-country information.

- *Benefits*: Expose the advantages and disadvantages of different data sources, be aware of potential limitations, and, accordingly, develop a protocol for reducing measurement errors and biases that are expectedly associated with the type(s) of data source you use.

Guideline 12. Include references in the dataset and facilitate access to the data sources.

- *Example 1*: Geddes et al.'s (2018) dataset on authoritarian regimes or Paglayan's (2021) dataset on state involvement in education provides short descriptions and references to the sources employed to justify the coding decisions.

- *Benefits*: Enhance transparency, replicability, and reliability of the dataset.

Guideline 13. Make your dataset publicly available (including online data exploration) and create ways to obtain feedback from data users, especially to improve coding.

- *Example 1*: Datasets like Whogoverns (Nyrup & Bramwell 2020) and V-Indoc (Neundorf et al. 2023), have made their dataset available online, but also gone further by eliciting interaction with data users. On the website, they provide emails to contact the dataset creators for inquiries or feedback, but someone can also sign up for a newsletter to obtain news associated with the dataset, new releases, or events associated with exploring the dataset.

- *Example 2*: In the case of V-Indoc dataset, the website used the Shiny app to explore the dataset across countries and variables

- *Benefit*: Increase the chances of more tests of the reliability and validity of your coding decisions and obtain additional information to improve coding in less clear/challenging cases. Another benefit is likely increased use of your data by other researchers. By facilitating ways to explore your dataset, you can reach a non-expert audience in statistical methods, but experts on the dataset's subject. This increases the reach and public exposure of your dataset. At times, visual tools can help the audience to spot potential coding errors quickly.

D1.1. Checklist for writing codebooks

In this appendix, we have noted some key points to keep in mind when constructing codebooks. In this subsection, we detail, in the form of a concretized checklist, the different features that we consider should be included in a codebook in order to, e.g., enhance intercoder reliability as well as misinterpretation (and potential misuse) of indicators by dataset users. Appendix Table D1.1 provides this checklist.

Appendix Table D1.1. A checklist for codebook construction

Components	Rationale
Concepts	Define precisely any ambiguous terms, notably including the key concepts or concept dimensions to be measured.
Opearationalizations	Specify the dimensionality of concepts and how different indicators relate to specific dimensions. Describe the approach to measurement and the relevant details of how data was collected (both general and variable-specific information).
Scaling	Provide information on the measurement level and scale for each variable. If relevant, describe any methods used to re-scale variables.
Scores	For categorical variables, briefly explain or exemplify the different categories/scores.
Clarifications	Where needed, provide further clarifications that allow users to understand how difficult or ambiguous cases have been coded. This includes any heuristics applied for coding particular sets of cases.
Unit of analysis and coverage	Specify, and if relevant, justify, the unit of analysis as well as the temporal, geographical, or other scope covered by each variable.
Indices	If the dataset contains indices constructed from two or more indicators, describe and justify the inclusion of relevant indicators, the aggregation rules, and scaling. A discussion on the extent to which these technical features match the theoretical concept to be measured by the index may also be useful.
Aggregation	Whenever scores on indicators represent aggregated information that has been reduced to one number, notify this and clarify the rules applied for aggregation (e.g., averaged across the population making up the unit or the minimum/maximum value observed in the population).. For example, the education datasets discussed in our article code country-year observations and have different ways to aggregate across territory in federal systems where the relevant education-system feature may differ across regions. Such aggregation rules should be made explicit for each relevant variable.
Uncertainty	Highlight potential sources of uncertainty, and if relevant, how dataset users may identify and account for uncertainty in the coding (e.g., detail any explicit measures of uncertainty contained within the dataset and how they should be interpreted).
Type of coding and sources	Describe the mode of data collection and who collected the data (country-experts, in-house researchers or research assistants, etc.). If relevant, list the (main) sources employed.
Organization and transparency	Provide a table of contents, and include, if relevant, separate sections to address both general issues pertaining to the wider dataset or specific (groups of) indicators to minimize redundancy while including all relevant

Appendix Table D1.1. A checklist for codebook construction

	information. For instance, key terms that pertain to a large number of indicators may be contained in a separate glossary, whereas definitions or other clarifications that are specific to a single indicator may be placed adjacent to this indicator.

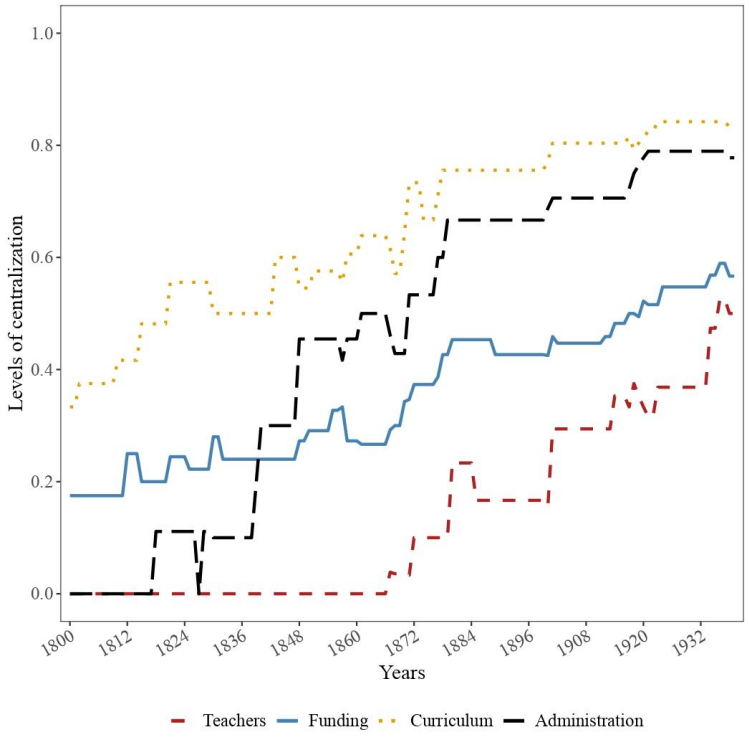
D2. Guidelines for data users

Guideline 14. A careful reading of articles introducing the dataset, codebook, and other documentation can reveal key assumptions underlying the dataset (e.g., threshold assumptions, underlying dimensions of the operationalizations), which are important for inference and interpretation. Datasets often have to make several simplifying assumptions to make the measures comparable across units (e.g., countries, regions, elites) and over time (e.g., centuries, weeks). These differences may sometimes have important implications for how measures are scored.

- *Example 1*: The three datasets on education made different assumptions to account for regional variation when subnational units have the authority on education. If subnational variation exists, HEQ looks at the most populous state (i.e., it looks at North Rhine-Westphalia in the case of Germany). V-Indoc asks experts for "the typical primary/secondary school," while EPSM uses different rules for different questions (e.g., most regions in a federal state should have compulsory education for all in order to say that the country had compulsory education for all).

- *Example 2*: Several dataset creators occasionally attempt to capture the same concept but differ in the dimension of that concept that they (want to) measure. Take education centralization as an example in Appendix Figure D3. The figure compares the historical trends using Ansell and Lindavall's (2020) measure – which focuses on a binary measure of education centralization is based on who has authority over the appointment, promotion, and payment of teachers – and EPSM's (del Río et al. 2024) measures that focus on the extent to which the state has the control over the funding, curriculum, and administration of education. These measures thus capture quite different dimensions of the education centralization concept, with implications for scores and interpretations. The most obvious pattern from Figure D3 is that the timing of centralization reforms depends on the dimension we select.

Figure D3. Comparing education centralization measures from EPSM and Ansell & Lindvall (2020): Averages across 19 countries with data on all measures



Notes: measures are normalized to 0–1: "teachers" refers to Ansell & Lindvall (2020)'s centralization measure, whereas the other measures are from EPSM.
Source: del Rio et al.

- *Implications and suggestions*: The differential scores and patterns following from the two measures comprising different dimensions of the same concept have two implications if we want to study the origins and effects of education centralization reforms. On the one hand, we might want to aggregate measures to have a more comprehensive measurement of the wider concept. However, this strategy entails losing information about particular trends (or if particular aspects of education centralization have particular causes or effects in causal analysis). A potential solution would be to go back to the research's theory and think about empirical implications that would lead researchers to expect that some dimensions are more likely to be relevant than others. Afterward, researchers can conduct a sub-component analysis and assess what dimension is driving trends (or causal relationships).

  On the other hand, if we focus on one dimension only, we might have the problem that focusing on, let's say, the dimension pertaining to teachers might provide a more conservative analysis of the origins and effects of education centralization. If researchers focus on the presence of a centralized state authority that manages education, we might have a more optimistic view of how states centralize the education system. Governments

might have incentives to reform some aspects of education systems before others. Theorizing and empirically testing (using different measures) at the concept-dimension level may allow researchers to assess such a more fine-grained level.

Guideline 15. Prioritize datasets that match the theoretical assumptions and purposes of your research over popular measures or cross-national and temporal scope. In the paper, we highlight the distinction between the *de jure* vs. *de facto* education systems and policies. Dataset users should consider to what extent the measures they use reflect one or both of these categories and keep this feature of the measure in mind when interpreting results. Indeed, studying the relationships between *de jure* and *de facto* measures for similar concepts may give rise to important insights. Researchers are, for example, often interested in understanding the extent to which changes in legislation or formal institutions produce changes in behavior or power relations.

- *Example*: In Appendix B, we find indications that *de jure* measures of education centralization exhibit a less persistent decrease in the aftermath of democratization compared to *de facto* measures.

- *Implications and suggestions*: Opting to use measures capturing *de jure* or *de facto* characteristics should be related to what type of theoretical argument we want to assess or the goals of the study more generally. Some extended practices might not be backed by legislation, and some legislation might not be fully implemented. On the one hand, if our study's goals are related to examining the origins of education systems, we might want to think of a theory that focuses on *de jure* characteristics. On the other hand, if we want to study the effects of studying under a particular education systems on, let's say, political attitudes, researchers might want to use *de facto* data, primarily (it might not be that most citizens were exposed to the education system as prescribed by the legislation, for example).

Guideline 16. When engaging in convergent validation exercises, pay careful attention to conceptual differences underlying measures that may, at first glance, seem to measure similar concepts. Sometimes, similar measures from different datasets may capture the exact same concept, but different thresholds are used to establish the different categories. Analyzing coding divergences could highlight different assumptions.

- *Example*: We refer to discussions in the main article for the coding of "religious education" in EPSM and HEQ for one example. Another example could be two otherwise similar binary measures of electoral democracy including a necessary suffrage criterion, but where one measure assumes that voting rights for all adult males is sufficient for scoring a regime democratic, whereas another requires universal suffrage (i.e., voting rights for all adult men and women).

Guideline 17. If possible, identifying the sources of (dis)agreement in similar measures across datasets could expose different assumptions made by dataset creators and provide nuanced insights that could aid both descriptive and causal inference. Factors that might influence measurement without dataset users being aware of it might be linked to data source availability or the fact that corruption or conflicts made coding more challenging or error-prone for one type of measure than another.

- *Example*: Datasets that rely on news reports might miss relevant cases because they depend on information availability. Some governments control the media and might censor or severely bias the news's content. Protest datasets, for example, may be prone to such and other biases, as protest visibility might depend, e.g., on protest size, protest tactics (violent ones are more easily reported), country size, and several other factors such as salient events (see, e.g., Hellmeier et al. 2018).

- *Implications and suggestions*: There might be a number of (un)observable factors that might bias our research without data users necessarily being aware of it. Explicitly identifying, empirically assessing, and clarifying the relevance of such factors is thus important. We suggest a number of tests in Appendix D1's **Guideline 9.**

Guideline 18. How suitable an indicator or index might be as a dependent or independent variable depends on the core concept the research aims to measure.

- *Example*: For economists, education is usually a measure of human capital, which is typically an independent variable (e.g., Angrist et al. 2021). This implies an interest in the successful implementation (de facto dimension, like the measurements provided in V-Indoc) or the outcome of education in the form of human capital instead of *de jure* education indicators as EPSM, HEQ, or Bromley et al.'s (2022) World Education Reform Dataset offer.

- *Implications and suggestions*: Prioritize the use of indicators and scaling that match the concept of interest. Is the concept that you envision an interval or categorical? Otherwise, there might be the issue that indices are not really measuring the concept of interest or other sources of biases pointed out in **Guideline 15.** We also refer to Goertz (2020) for further suggestions on concepts, measurement, and scaling.

Appendix E. Indicator harmonization

Although many indicators across the EPSM, V-Indoc, and HEQ datasets overlap in terms of the concepts they capture, the ordinal scales used to code these indicators can vary. Below, we outline how the scales of the indicators we discuss in the main text were harmonized so that they could be compared meaningfully across datasets.

Centralized curricula

The centralized curricula indicator in the EPSM dataset measures the degree to which school curricula are determined by a ministry (or other entity) at the national or sub-national level. The V-Indoc version of the indicator measures the extent to which a national authority sets the official curriculum framework for schools. Both indicators are measured based on an ordinal scale with four levels, but there is no clear one-to-one mapping between these levels. As such, we map each indicator to a three-level scale so that each level represents similar responses. The HEQ data are coded to match the scale of the V-Indoc indicator.

<table>
<tr><th>Scale</th><th colspan="2">EPSM</th><th colspan="2">V-Indoc / HEQ</th></tr>
<tr><td rowspan="2">0</td><td>1</td><td>There is no centralized curricula provided by the national government or by regional government</td><td rowspan="2">0</td><td rowspan="2">A national authority does not set the official curriculum framework, that is, the curriculum framework is completely set by sub-national authorities</td></tr>
<tr><td>2</td><td>There is a centralized curriculum provided by a regional government only</td></tr>
<tr><td rowspan="2">1</td><td rowspan="2">3</td><td rowspan="2">There is a centralized curriculum provided partly by a regional and partly by a national government</td><td>1</td><td>Sub-national authorities mostly set the official curriculum framework, with some input from the national authority</td></tr>
<tr><td>2</td><td>A national authority mostly sets the official curriculum framework, with some input from sub-national authorities</td></tr>
<tr><td>2</td><td>4</td><td>There is a centralized curriculum provided by a national government only</td><td>3</td><td>A national authority fully sets the official curriculum framework</td></tr>
</table>

Politicized teacher recruitment

The political teacher hiring variable in the V-Indoc dataset measures the extent to which hiring decisions for teachers are based on their political views and/or political behavior and/or moral character, which follows a four-level ordinal scale. A similar binary indicator in the HEQ dataset

measures whether applicants must show proof of moral competency, belong to a particular religion, and whether public primary school teachers are required to swear allegiance to the state and/or the constitution (yes/no) or to a particular party or a ruler (yes/no). To facilitate comparability, we map the ordinal scale for the V-Indoc indicator to a binary scale.

<table>
<tr><th>Scale</th><th colspan="2">V-Indoc</th><th colspan="2">HEQ</th></tr>
<tr><td>0</td><td>0</td><td>Rarely or never</td><td>0</td><td>Applicants do not need proof of moral competency or belong to a religion, and teachers do not need to swear allegiance to the state/constitution/party/ruler</td></tr>
<tr><td rowspan="3">1</td><td>1</td><td>Sometimes</td><td rowspan="3">1</td><td rowspan="3">Applicants need proof of moral competency or belong to a religion, or teachers do not need to swear allegiance to the state/constitution/party/ruler</td></tr>
<tr><td>2</td><td>Often</td></tr>
<tr><td>3</td><td>Almost exclusively</td></tr>
</table>

Religious teaching in primary schools

The religious teaching in primary school indicators we use in our main text are sourced from the EPSM and HEQ datasets. Both indicators take on a value of 1 if national laws mandate that religion should be taught as a part of civic education and 0 otherwise. We use the original scales for these indicators as they overlap.

Bibliography

Angrist, N., Djankov, S., Goldberg, P.K. *et al.* Measuring human capital using global learning data. *Nature* 592, 403–408 (2021).

Benavot, Aaron. (2004). *A Global Study of Intended Instructional Time and Official School Curricula, 1980-2000*. UNESCO International Bureau of Education. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000146625.

Bromley, Patricia; Kijima, Rie; Overbey, Lisa; Furuta, Jared; Choi, Minju; Santos, Heitor; Song, Jieun; Nachtigal, Tom; Yang, Marcia, 2023, "World Education Reform Database (WERD)", https://doi.org/10.7910/DVN/C0TWXM, Harvard Dataverse, V3

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Eitan Tzelgov, Luca Uberti, Yi ting Wang, Tore Wig, and Daniel Ziblatt. 2022. "V-Dem Codebook v12" Varieties of Democracy (V-Dem) Project.".

Del Río A., Knutsen C.H. and Lutscher P.M. 2024. "Education policies and systems across modern history: a global dataset." *Comparative Political Studies* 58(5): 851-889.

Geddes, B., Wright, J. and E. Frantz. 2018. *How Dictatorships Work: Power, Personalization, and Collapse*. Cambridge: Cambridge University Press.

Goertz, G. (2020) Social Science Concepts and Measurement: New and Completely Revised Edition, Princeton: Princeton University Press.

Grumbach, J.M. 2023 "Laboratories of Democratic Backsliding." *American Political Science Review* 117(3): 967-984.

Hellmeier, S., N.B. Weidmann and E.G. Rød. 2018. "In the Spotlight: Analyzing Sequential Attention Effects in Protest Reporting." *Political Communication* 35(4): 587–611.

Neundorf, A., et al. 2023. Data and "Codebook: Varieties of political indoctrination in education and the media (V-Indoc)." URL: http://dx.doi.org/10.5525/gla.researchdata.1397

Neundorf, A., et al. 2024. "Varieties of indoctrination (V-Indoc): Introducing a global dataset on the politicization of education and the media." 2023*, Perspectives on Politics*, Online First.

Paglayan, A.S. 2019. "Public Sector Unions and the Size of Government." *American Journal of Political Science* 63(1): 21-36.

Paglayan, A. S. 2021. "The Non-Democratic Roots of Mass Education: Evidence from 200 Years." *American Political Science Review* 115(1): 179-198.

Paglayan, A. S. 2022a. "Education or Indoctrination? The Violent Origins of Public School Systems in an Era of State-Building." *American Political Science Review* 116(4): 1242-1257.

Paglayan, A. S. 2022b. "The Historical Political Economy of Education." In *The Oxford Handbook of Historical Political Economy*, edited by Jeffery Jenkins and Jared Rubin. Oxford University Press.

Paglayan, A. S. n.d. "Historical Education Quality (HEQ) dataset," working in progress

Pemstein, D. et al.. 2020. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." *V-Dem Working Paper* No. 21.

Weidmann, N.B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science*, 60: 206-218.

Weidmann, N. B. 2024. "Recent Events and the Coding of Cross-National Indicators." *Comparative Political Studies*, 57(6), 921-937

Corresponding author: **Adrián del Río** (a.d.r.rodriguez@stv.uio.no, Norway) is MSCA-postdoctoral fellow at the University of Oslo. Before joining Oslo, he was a Humboldt postdoctoral fellow at the Centre for East European and International Studies and Berlin Social Science Centre. His research interests include the origins and effects of elite divisions in autocracies, democratization as well as the effects of education policies. He holds a PhD in Social and Political Science from the European University Institute.

Wooseok Kim (wooseok.kim@glasgow. ac.uk, United Kingdom) is a Leverhulme Trust Early Career Fellow at the University of Glasgow. His research investigates the factors that contribute to regime resilience and performance across democratic and autocratic contexts. Much of his ongoing work on these themes focuses on the role of party systems. He holds a PhD in Political Science from the University of Michigan.

Carl Henrik Knutsen (c.h.knutsen@stv.uio.no. Norway) is a Professor at the Department of Political Science, University of Oslo. He leads the Comparative Institutions and Regimes research group, Research Professor at PRIO and PI of Varieties of Democracy. His research interests include regime change and stability, the economic effects of institutions, autocratic politics, and more broadly, comparative politics and political economy. He is PI of the project "Emergence, Life, and Demise of Autocratic Regimes" project (2020-2025) funded by an ERC Consolidator Grant, "Policies of Dictatorships" (2020-2024), financed by Research Council Norway. He holds a PhD in Political Science from the University of Oslo.

Anja Neundorf (anja.neundorf@glasgow.ac.uk, United Kingdom) is a Professor of Politics and Research Methods at the School of Social and Political Sciences at the University of Glasgow. Before joining Glasgow, she held positions at the University of Nottingham (2013-2019) and Nuffield College, University of Oxford (2010-2012). Professor Neundorf is currently leading a European Research Council Consolidator Grant project on "Democracy under Threat: How Education can Save it" (DEMED).

Agustina S. Paglayan (apaglayan@UCSD.EDU, United States) is an Associate Professor of Political Science at the University of California, San Diego, and the author of Raised to Obey: The Rise and Spread of Mass Education (Princeton University Press, 2024). Her research examining the interplay between education and democracy, autocracy, state-building, and unions has been published in the American Political Science Review and American Journal of Political Science and has received numerous awards from the American Political Science Association, including the APSA Heinz I. Eulau Award for the best article published in the APSR. Dr. Paglayan received her PhD from Stanford University.

Eugenia Nazrullaeva (evgeniya.nazrullaeva@uni-konstanz.de, Germany) is a Postdoctoral Researcher in the Department of Politics and Public Administration at the University of Konstanz. She holds a PhD in Political Science from the University of California, Los Angeles. Her research interests are in the areas of political economy and economic history