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

Adrián Del Río, Woseeok Kim, Carl Henrik Knutsen, Anja Neundorf, Agustina S. Paglayan and Eugenia Nazrullaeva

Corresponding author: Adrián Del Río (a.d.r.rodriguez@stv.uio.no, Norway) is a Marie Skłodowska-Curie Actions 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 the Berlin Social Science Center. His research interests include the origins and effects of elite divisions in autocracies, democratization, and 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 in the Department of Political Science, University of Oslo. He leads the Comparative Institutions and Regimes research group, is a research professor at the Peace Research Institute Oslo, and is a principal investigator of Varieties of Democracy (V-Dem). His research interests include regime change, the economic effects of institutions, and autocratic politics. He is the principal investigator of the "Emergence, Life, and Demise of Autocratic Regimes" project (2020–25), which is funded by a European Research Council Consolidator Grant, and the "Policies of Dictatorships" project (2020–24), 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 (2013–19) and Nuffield College, University of Oxford (2010–12). 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 *(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 the* 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* American Political Science Review. *Dr. Paglayan received her PhD from Stanford University.*

Eugenia Nazrullaeva (E.Nazrullaeva@liverpool.ac.uk, United Kingdom) is a lecturer in the Department of Politics, University of Liverpool. Previously, she was 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.

doi:10.1017/S1537592725103058

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, as well as their 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.

Keywords: dataset construction, transparency, replication, cross-national dataset, research methods, validity of inferences, education systems

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: these datasets eventually become public and enable not only their creators but also other researchers to study a wide range of questions.

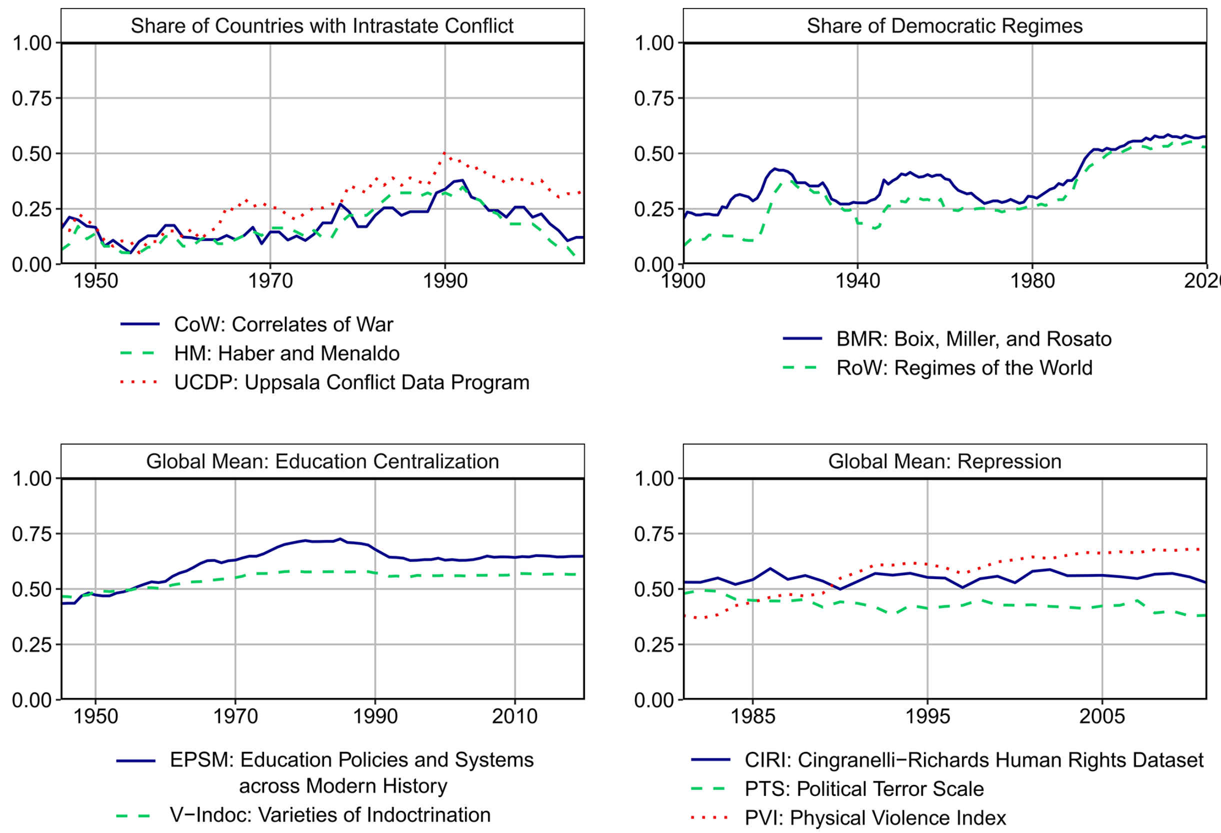
Making appropriate descriptive and causal inferences based on datasets created by other academics, however, is not straightforward. It requires understanding how the dataset was constructed—how variables capture multidimensional concepts, how each dimension is operationalized, what information and sources 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—for example, a *majority of adult males* must be able to vote in Boix, Miller, and Rosato's (2013) classification, whereas *universal male* suffrage or, simply, *universal* suffrage is required by Skaaning, Gerring, and Bartusevičius (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"),[1] 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: intrastate 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 inference. For example, depending on which data source we use, the share of countries experiencing intrastate conflict in recent years oscillates between 5% (Haber and Menaldo 2011) and 33% (Uppsala Conflict Data Program; see Pettersson 2022).[2] Global democratic decline took place during the 1960s according to one democracy measure (Boix, Miller, and Rosato 2013), but not according to another (Regimes of the World, which builds on V-Dem data; see Coppedge et al. 2023). The degree of education centralization varied more from 1945 to 1995 according to one measure (Education Policies and Systems across Modern History [EPSM]; see Del Río, Knutsen, and Lutscher 2024) than according to another (Varieties of Indoctrination [V-Indoc]; see Neundorf et al. 2023). Finally, the level of repression around the world increased (Physical Violence Index; see Coppedge et al. 2023), remained similar (Cingranelli-Richards Human Rights Dataset; see Cingranelli, Richards, and Clay 2021), or declined (Political Terror Scale; see Gibney et al. 2022) in 2011 relative to 1981, depending on the choice of repression measure. Our point is not that measures of the same concept (or at least similar concepts) 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.[3]

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.[4] 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 EPSM dataset (Del Río, Knutsen, and Lutscher

Figure 1
Comparisons of Different Measures of the Same Concept



Sources: BMR (Boix, Miller, and Rosato 2013); CIRI (Cingranelli, Richards, and Clay 2021); CoW (Sarkees and Wayman 2010); EPSM (Del Río, Knutsen, and Lutscher 2024); HM (Haber and Menaldo 2011); PTS (Gibney et al. 2022); PVI (Coppedge et al. 2023); RoW (Coppedge et al. 2023); UCDP (Pettersson 2022); V-Indoc (Neundorf et al. 2023).
Note: CIRI and PTS are ordinal variables that have been rescaled to a unit interval using min-max scaling to facilitate comparisons.

2024), the 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 developed through the experiences of these 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 online 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[5] and causal inferences (e.g., Casper and Tufis 2003). 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 that 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 explicit 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

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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, Gemenis (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 nonclassical measurement error.[6] 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., Knutsen et al. 2024; Little and Meng 2024). For instance, we highlight how different data types 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.

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 online 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 contents 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 to 2020; V-Indoc relies on country-expert assessments across 160 countries from 1945 to 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 and 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 has taken between three and six months in the case of HEQ.

Relatedly, while using secondary sources can enable dataset creators to expand the geographic and temporal scope of their dataset,[7] 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 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 to 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[8] 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 politicized teacher firing or indoctrination

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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 research assistants. 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 (four on compulsory education, seven on ideological content teaching, seven on school autonomy, three on teacher training) *Costs*: approx. $1,000 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 de jure education policies + Does not rely exclusively on language expertise + Relatively quick	– Relies 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 one million) – Only categorical variables or ordinal scales – Requires resource-demanding 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 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*: five countries, beginning with the first year when each country's national government started to regulate the curriculum or teacher training and recruitment, up to 2015 *Indicators*: 39 indicators (five on curriculum, 34 on teacher training and recruitment) *Costs*: approx. $7,200 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

(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 onward 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 nonexpert coding; see, e.g., Knutsen et al. 2024).[9]

Overall, the inherent tensions between breadth and accuracy (given resource constraints) and the choices made by each research team concerning which goal to prioritize resulted in EPSM and V-Indoc accomplishing a substantially broader coverage than HEQ in much shorter time, but at the potential cost of accuracy. As we discuss in the Advice for Dataset Users section, such

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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.*
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 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 those that provide the greatest cross-national or temporal coverage.
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.

Note: * Online appendix D1.1 includes a detailed checklist for best practice on creating codebooks.

features of the data and the implications of the discussed trade-offs should also be considered by data users conducting 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. This resulted in an average cost of approximately $1,000 per country. V-Indoc relied on one postdoctoral scholar 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; a research assistant then cross-checked the initial coding against the primary sources for all countries, which led to a back-and-forth with consultants before arriving at the final coding for an average cost of $7,200 per country.

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 to mitigate these issues and try to generalize different insights through a set of guidelines for future dataset makers, which we detail and concretize further in online appendix D1 and summarize in table 2 in the concluding section.

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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 questions to mitigate multidimensionality.

Maintaining consistent definitions of evolving terms or even concepts and avoiding multidimensionality and ambiguity in interpretations are, as indicated, often surprisingly difficult. Indeed, these issues may even be hard to detect. Yet several (fairly straightforward) strategies can help to 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 to ensure transparency to data users wondering exactly how questions and categories should be interpreted (or how they align with their specific research questions and contexts), but it also improves intercoder reliability and reproducibility and, crucially, ensures that the dataset provides information that is comparable across space and time.[10] 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 to 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 online 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 developing detailed coding instructions and 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 to 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 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 different coders code different units in the dataset). Absent such strategies, different coders will likely rely on dissimilar 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 on, for example, 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 and studying their relationships with income, state capacity, democracy, or some other feature correlated with the former three concepts.

More generally, low intercoder reliability may cause additional problems for datasets coded by more than one person, so 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 (1) 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; (2) developing and updating a joint rules-of-thumb (RoT)

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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; (3) active communication through a joint web platform and (sometimes) physical colocation when coding, allowing coders to discuss and find joint solutions to challenging cases; and (4) 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 heuristics (preferably made explicit in the RoT document) when approaching similar cases. Nonetheless, avoiding differential interpretations and uses of heuristics across coders is close to impossible to guard against completely, and such between-coder differences 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 item-response theory 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 Coppedge et al. 2020; Pemstein et al. 2025). 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, coders' 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 nontrivial 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 avoid erroneous interpretations of scores and evaluate which kinds of analyses variables may be used for, among other purposes. We list this as one guideline for constructing codebooks, alongside several other pieces of advice indicated in this section, in table D1.1 in the online appendix. Scaling information is provided in the codebooks of the three education datasets, although the information is sometimes insufficiently specified or otherwise problematic.[11]

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 for dealing with the divergence 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 intracoder consistency, the team met and discussed potential sources of coding disagreement and strategies for additional source collection.[12] The goals of this procedure and the wider triangulation strategy were to improve the validity and reliability of the coding and to assess and express remaining uncertainty.

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 nonexperts). 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 code mainly 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-

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

specific knowledge may be extremely time consuming and thus infeasible for any single researcher or research assistant coding numerous cases. Structured (country-topic) expert surveys are therefore 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.[13]

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 help). Limited 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 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). In this case, it makes more sense to train a few coders (e.g., research assistants) to code each country than ask experts to do so.

It is possible to devise strategies that harness benefits from different approaches. The HEQ dataset, for example, relies on country-specific expert historians' local knowledge of the legal educational landscape. This knowledge increases the accuracy and completeness of information about de jure policies, but it also relies 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, data creators can develop protocols for finding and checking with people who have knowledge of the local context to obtain information when particular cases have complexities.

Advice for Dataset Users

This section turns to potential pitfalls and advice for users of datasets, focusing on how different dataset characteristics—and 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.[14] We refer to table 2 for a summary and online 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 in 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 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 contain data that are suited for analyzing aggregate trends.

Same Name, Different Content

As previewed in the introduction, several dataset creators occasionally attempt to capture the same concept but differ on which dimension of that concept 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, Gandhi, and Vreeland 2010), while others incorporate suffrage rights (Boix, Miller, and Rosato 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

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

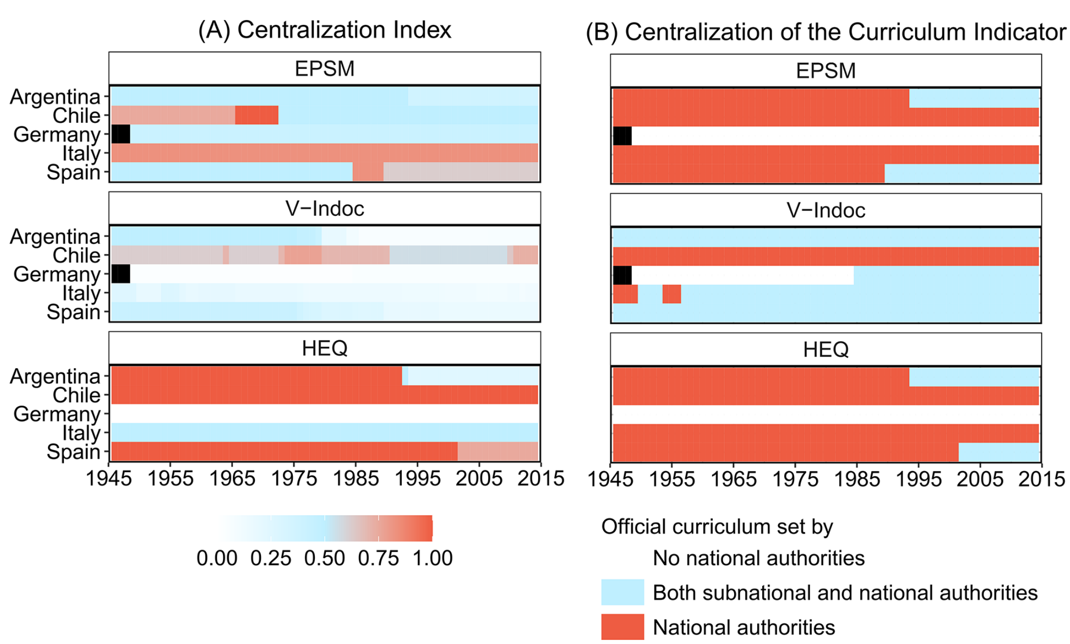
of *education centralization.* Following the emerging literature on education and state building (e.g., Paglayan 2022a; 2022b), education centralization refers to the concentration of authority over education policy decisions in the hands of the national government (Ansell and Lindvall 2020; Del Río, Knutsen, and Lutscher 2024; Neundorf et al. 2024; Paglayan 2021). 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 (Grindle 2004; Kaufman and Nelson 2004; Murillo 1999). 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 (Clarke, Timperley, and Hattie 2003; Zhao 2012), or national school inspection systems (Cermeño, Enflo, and Lindvall 2022).

EPSM, V-Indoc, and HEQ all offer indices that measure education centralization but emphasize different dimensions (see online appendix B for a summary of how these indices are constructed). EPSM focuses on the (de jure) 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.[15] 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 (panel A) 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 (panel B). The darker the cells, the more centralized the education system is.

We can draw several takeaway 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 panel A, 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 (compare panel B), does not entail that it

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



Note: See online appendix B for a description of the centralization indices across the three datasets. The EPSM and V-Indoc datasets have missing values for Germany between 1945 and 1949 as both datasets follow V-Dem's coding of country-years.

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

also controls other aspects of education systems, such as funding and management (only included in the EPSM index). For example, Augusto Pinochet's Chile (1973–90) maintained a centralized national curriculum but engaged in the decentralization of education funding and management to municipalities (Cox 2005).[16] Indicatively, the major differences between the EPSM and HEQ indices in panel A—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 panel B.

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 commonly used in empirical research and require researchers to be deeply familiar with the complex decisions and indicators involved in the creation of those aggregated measures.

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 according to 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.[17] 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 1994, when a new law formally recognized the ability of provinces to have some say over the curriculum.[18]

De jure versus 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 (Paglayan 2019; Przeworski 2004). 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 require to empirically study such questions 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 to accomplish this goal, as we illustrate in this section.

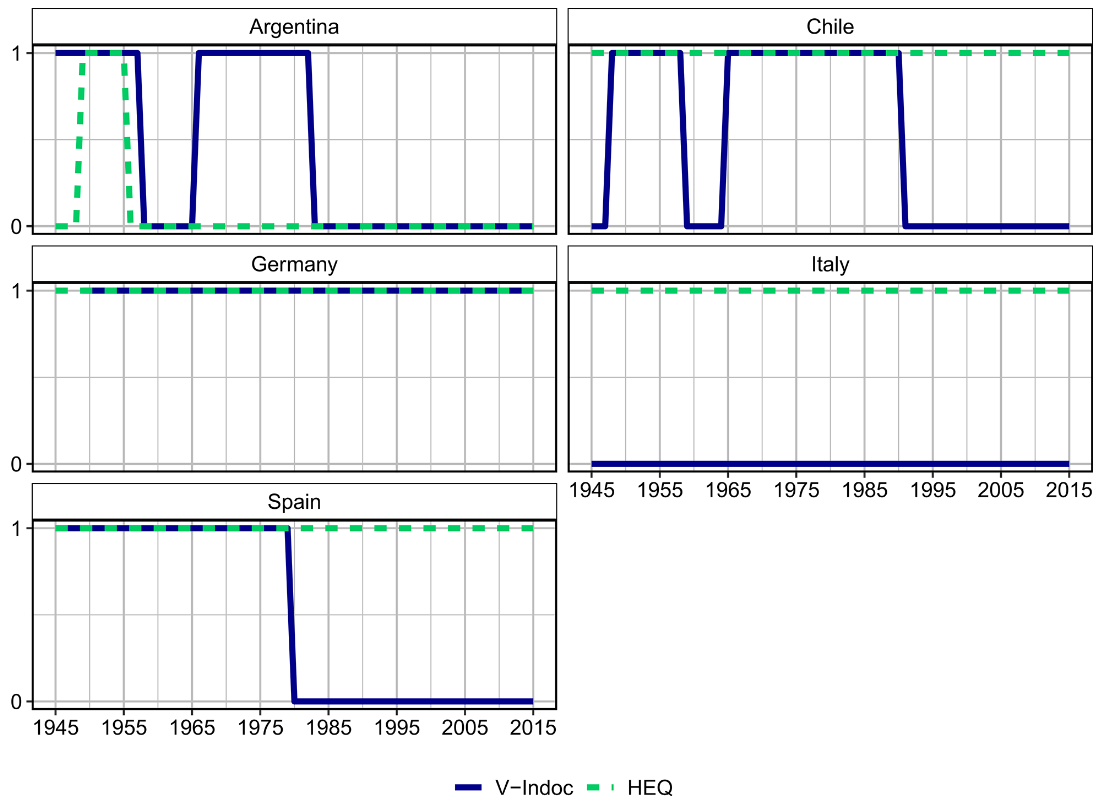
Several factors can affect gaps between de jure policies and de facto practices, including the state's fiscal and administrative capacity, the existence of school inspections, political regime type, conflict, or a country's territorial size (Cermeño, Enflo, and Lindvall 2022; Lopez 2020; Paglayan 2024). 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 draws on measures from V-Indoc and HEQ to demonstrate the relevance of the de jure versus de facto distinction on one specific dimension of education systems: the politicization of teacher recruitment practices.[19] 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 zero when neither of these requirements is present and a value of one 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, political behavior, and/or moral character.[20] The possible answer categories are as follows: rarely or never, sometimes, often, and almost exclusively. To ease comparisons, we dichotomize the V-Indoc indicator: zero means hiring decisions are rarely or never based on politicized criteria, while one combines the three politicized categories (i.e., sometimes, often, and almost exclusively).[21]

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.[22]

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

Figure 3
Trends in Politicized Teacher Recruitment (V-Indoc and HEQ)



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

Peronist regime's 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 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 Juan Perón 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 military dictatorship headed by 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 during two periods after 1955.

Same Concepts but Different Thresholds

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.[23]

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, stand-alone course. The subject need not be about religion exclusively.

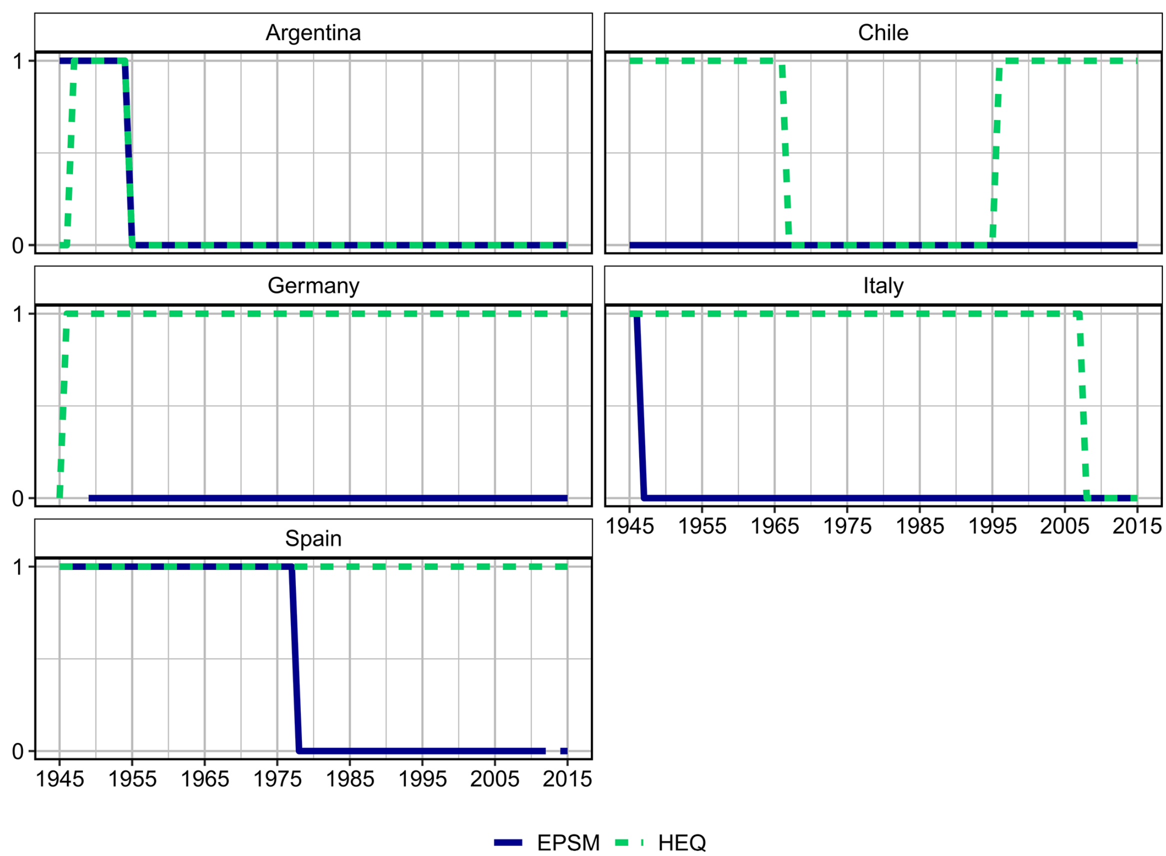
https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

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 forms (large parts of) a stand-alone 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 be the basis of an official school of thought that has the status of an "official" ideology in the regime. The latter 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, stand-alone civics courses are coded). As a result of this coding decision, countries with a compulsory, stand-alone 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 to 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.[24] 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.

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 the fine print in

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



Note: HEQ and EPSM focus on stand-alone 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 zero and one. 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 or dominant models in the history curriculum.

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press

codebooks and other dataset documentation before selecting which measure is most appropriate to use for a particular purpose.

Lessons

This paper has highlighted how various and specific choices on data collection and measurement influence how 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—for example, 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 undercommunicated data collection choices, we also hope that our reflections can promote a shift toward 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, online 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). Sections D1 and D2 in the online appendix 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 RoT 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, the 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 nontrivial 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.

Depending on the data user's research design and goals, our study also highlights how one can fruitfully 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 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.

Finally, 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.

Supplementary material

To view supplementary material for this article, please visit http://doi.org/10.1017/S1537592725103058.

Data replication

Data replication sets are available in Harvard Dataverse at: https://doi.org/10.7910/DVN/FU0U8V

Acknowledgments

We are thankful for the input of editors and reviewers, whose contributions improved this paper. 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 European Political Science Association conference and the Practices of 2024 Comparative-Historical Analysis Conference at Aarhus University. The HEQ database has been funded by numerous centers and programs at Stanford University, including the King Center on Global Development, the Europe Center, Stanford SEED, and the Vice Provost for Graduate Education. V-Indoc was generously funded by the European Research Council Consolidator Grant "Democracy under Threat: How Education Can Save It" (DEMED) (grant no. 865305). EPSM has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant no. 863486).

Notes

1 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).

2 The two datasets use different thresholds of how they define intrastate conflict based on the number of deaths. Haber and Menaldo use a threshold of a thousand deaths while the Uppsala Conflict Data Program uses 25.

3 This example pertains to descriptive inference, but the more general point on the relevance of choice of measure holds also for causal inference. In online 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.

4 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 gross domestic product to COVID-19 deaths, is also fraught with different pitfalls and is the subject of a separate literature (see, e.g., Jerven 2013; Knutsen and Kolvani 2024; Martínez 2022).

5 For a recent discussion on democracy measurement and time trends in global democracy, see Knutsen et al. (2024); Little and Meng (2024).

6 See also Dinas and Gemenis (2010).

7 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.

8 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., the tests conducted by the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), 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.

9 In online appendix C we assess whether coding divergences between V-Indoc and the rest of the datasets are driven by the number of coders that V-Indoc employed and coders' self-reported uncertainty.

10 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. At the same time, dataset creators should be conscious about the possibility of terms being used with different contents in source materials for different contexts, and be able to identify when this might happen.

11 For instance, the scale options listed for V-Indoc includes "dichotomous," which is strictly speaking not a scale option, and combines the ratio levels 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.

12 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 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.

13 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, and regional conferences on education, and collected emails of the authors or participants. Third, we contacted education-related nongovernmental organizations and policy experts outside academia, asking them to circulate our call among their network.

14 See online appendix E for more details on how the indicators are harmonized, including their original scales. For our replication materials, see Del Río et al. (2025).

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

16 See Decree DFL No. 3.166/80 (1980); Decree DFL No. 5.077/80 (1980); Decree DFL No. 13.063 (1980). Available at Biblioteca Nacional del Congreso de Chile.

17 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 state, "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*" (p. 4; emphasis added).

18 The 1993 Federal Law of Education (Ley No. 24.195, available at Biblioteca Nacional del Congreso de Argentina) 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.

19 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.

20 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)."

21 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.

22 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., Knutsen et al. 2024; Little and Meng 2024) rather than how to fruitfully combine insights gained from different measures.

23 V-Indoc also contains information on the presence of religious content in education. But instead of considering stand-alone 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.

24 Decree No. 40 (1980), available at Biblioteca Nacional del Congreso de Chile.

References

Acemoglu, Daron, and James A Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press. DOI: 10.1017/cbo9780511510809.

Ansell, Ben W., and Johannes Lindvall. 2020. *Inward Conquest: The Political Origins of Modern Public Services*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108178440.

Boix, Carles, Michael K. Miller, and Sebastian Rosato. 2013. "A Complete Data Set of Political Regimes, 1800–2007." *Comparative Political Studies* 46 (12): 1523–54. DOI: 10.1177/0010414012463905.

Cantoni, Davide, Yuyu Chen, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang. 2017. "Curriculum and Ideology." *Journal of Political Economy* 125 (2): 338–92. DOI: 10.1086/690951.

Casper, Gretchen, and Claudiu Tufis. 2003. "Correlation versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets." *Political Analysis* 11 (2): 196–203. DOI: 10.1093/pan/mpg009.

Cermeño, Alexandra L., Kerstin Enflo, and Johannes Lindvall. 2022. "Railroads and Reform: How Trains Strengthened the Nation State." *British Journal of*

Political Science 52 (2): 715–35. DOI: 10.1017/s0007123420000654.

Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland. 2010. "Democracy and Dictatorship Revisited." *Public Choice* 143 (1–2): 67–101. DOI: 10.1007/s11127-009-9491-2.

Cingranelli, David L., David L. Richards, and K. Chad Clay. 2021. The CIRI Human Rights Dataset. Version 2014.04.14, November 19. *Harvard Dataverse.* DOI: 10.7910/DVN/UKCPXT.

Clarke, Shirley, Helen Timperley, and John Hattie. 2003. *Unlocking Formative Assessment: Practical Strategies for Enhancing Students' Learning in the Primary and Intermediate Classroom*. Auckland: Hodder Moa Beckett.

Coppedge, Michael, John Gerring, Adam Glynn, Carl Henrik Knutsen, Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, and Jan Teorell. 2020. *Varieties of Democracy: Measuring Two Centuries of Political Change*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108347860.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, et al. 2023. "Codebook." Version 13, March. Gothenburg: Varieties of Democracy (V-Dem) Institute. https://v-dem.net/documents/24/codebook_v13.pdf.

Cox, Cristián D., ed. 2005. *Las políticas educacionales en el cambio del siglo: La reforma del sistema escolar de Chile*, 2nd edition. Santiago de Chile: Editorial Universitaria.

Del Río, Adrián, Carl Henrik Knutsen, and Philipp M. Lutscher. 2024. "Education Policies and Systems across Modern History: A Global Dataset." *Comparative Political Studies* 58 (5): 851–89. DOI: 10.1177/00104140241252075.

Del Río, Adrián, Wooseok Kim, Carl Henrik Knutsen, Anja Neundorf, Agustina Paglayan, and Eugenia Nazrullaeva. 2025. "Replication Data for: Enhancing Transparency and Replicability in Data Collection: Lessons from the Construction of Three Education Datasets." *Harvard Dataverse.* DOI: 10.7910/DVN/FU0U8V.

Dinas, Elias, and Kostas Gemenis. 2010. "Measuring Parties' Ideological Positions with Manifesto Data: A Critical Evaluation of the Competing Methods." *Party Politics* 16 (4): 427–50. DOI: 10.1177/1354068809343107.

Fukuyama, Francis. 2004. *State-Building: Governance and World Order in the 21st Century*. Ithaca, NY: Cornell University Press. DOI: 10.7591/9780801455360.

Gemenis, Kostas. 2012. "Proxy Documents as a Source of Measurement Error in the Comparative Manifestos Project." *Electoral Studies* 31 (3): 594–604. DOI: 10.1016/j.electstud.2012.01.002.

Gibney, Mark, Peter Haschke, Daniel Arnon, Attilio Pisanò, Gray Barrett, Baekkwan Park, and Jennifer Barnes. 2022. The Political Terror Scale, 1976–2021. Dataset, 2022 version. Asheville, NC: Political Terror Scale. https://www.politicalterrorscale.org/Data/Data-Archive.html.

Grindle, Merilee S. 2004. "Good Enough Governance: Poverty Reduction and Reform in Developing Countries." *Governance* 17 (4): 525–48. DOI: 10.1111/j.0952-1895.2004.00256.x.

Haber, Stephen, and Victor Menaldo. 2011. "Do Natural Resources Fuel Authoritarianism? A Reappraisal of the Resource Curse." *American Political Science Review* 105 (1): 1–26. DOI: 10.1017/s0003055410000584.

Jerven, Morten. 2013. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Ithaca, NY: Cornell University Press. DOI: 10.7591/9780801467615.

Kasuya, Yuko, and Kota Mori. 2021. "Re-Examining Thresholds of Continuous Democracy Measures." *Contemporary Politics* 28 (4): 365–85. DOI: 10.1080/13569775.2021.1993564.

Kaufman, Robert, and Joan M. Nelson, eds. 2004. *Crucial Needs, Weak Incentives: Social Sector Reform, Democratization, and Globalization in Latin America*. Baltimore: Johns Hopkins University Press. DOI: 10.56021/9780801880490.

Knutsen, Carl Henrik, Kyle L. Marquardt, Brigitte Seim, Michael Coppedge, Amanda B. Edgell, Juraj Medzihorsky, Daniel Pemstein, Jan Teorell, John Gerring, and Staffan I. Lindberg. 2024. "Conceptual and Measurement Issues in Assessing Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 162–77. DOI: 10.1017/s104909652300077x.

Knutsen, Carl Henrik, and Palina Kolvani. 2024. "Fighting the Disease or Manipulating the Data? Democracy, State Capacity, and the COVID-19 Pandemic." *World Politics* 76 (3): 543–93. DOI: 10.1353/wp.2024.a933071.

Knutsen, Carl Henrik, and Simone Wegmann. 2016. "Is Democracy about Redistribution?" *Democratization* 23 (1): 164–92. DOI: 10.1080/13510347.2015.1094460.

Little, Andrew T., and Anne Meng. 2024. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57 (2): 149–61. DOI: 10.1017/s104909652300063x.

Lopez, David. 2020. "State Formation, Infrastructural Power, and the Centralization of Mass Education in Europe and the Americas, 1800 to 1970." Paper presented at the 2020 Annual Meeting of the American Political Science Association, held online, September 9–13.

Marquardt, Kyle L., and Daniel Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26 (4): 431–56. DOI: 10.1017/pan.2018.28.

Martínez, Luis R. 2022. "How Much Should We Trust the Dictator's GDP Growth Estimates?" *Journal of Political Economy* 130 (10): 2731–69. DOI: 10.1086/720458.

Munck, Gerardo L., and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35 (1): 5–34. DOI: 10.1177/001041400203500101.

Murillo, María Victoria. 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. DOI: 10.2307/166226.

Neundorf, Anja, Eugenia Nazrullaeva, Ksenia Northmore-Ball, Katerina Tertytchnaya, Wooseok Kim, Aaron Benavot, Patricia Bromley, et al. 2023. "Varieties of Political Indoctrination in Education and the Media (V-Indoc) Codebook." DEMED Project documentation, March 13. SSRN preprint. DOI: 10.2139/ssrn.4726306.

Neundorf, Anja, Eugenia Nazrullaeva, Ksenia Northmore-Ball, Katerina Tertytchnaya, and Wooseok Kim. 2024. "Varieties of Indoctrination the Politicization of Education and the Media around the World." *Perspectives on Politics* 22 (3): 771–98. DOI: 10.1017/s1537592723002967.

Paglayan, Agustina S. 2019. "Public Sector Unions and the Size of Government." *American Journal of Political Science* 63 (1): 21–36. DOI: 10.1111/ajps.12388.

——. 2021. "The Non-Democratic Roots of Mass Education: Evidence from 200 Years." *American Political Science Review* 115 (1): 179–98. DOI: 10.1017/s0003055420000647.

——. 2022a. "Education or Indoctrination? The Violent Origins of Public School Systems in an Era of State-Building." *American Political Science Review* 116 (4): 1242–57. DOI: 10.1017/s0003055422000247.

——. 2022b. "The Historical Political Economy of Education." In *The Oxford Handbook of Historical Political Economy*, eds. Jeffery A. Jenkins and Jared Rubin, 837–56. Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780197618608.013.45.

——. 2024. *Raised to Obey: The Rise and Spread of Mass Education*. Princeton, NJ: Princeton University Press. DOI: 10.1515/9780691261775.

——. n.d. Historical Education Quality (HEQ) Dataset. Work in progress.

Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Juraj Medzihorsky, Joshua Krusell, Farhad Miri, and Johannes von Römer. 2025. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." Working Paper 2025:21, March. Gothenburg: Varieties of Democracy (V-Dem) Institute. https://www.v-dem.net/media/publications/wp21_2025.pdf.

Pettersson, Therese. 2022. "UCDP Dyadic Dataset Codebook." Version 22.1. Uppsala: Uppsala Conflict Data Program. https://ucdp.uu.se/downloads/dyadic/ucdp-dyadic-221.pdf.

Przeworski, Adam. 2004. "Institutions Matter?" *Government and Opposition* 39 (4): 527–40. DOI: 10.1111/j.1477-7053.2004.00134.x.

Sarkees, Meredith Reid, and Frank Wayman. 2010. *Resort to War: 1816–2007*. Washington: CQ Press. DOI: 10.4135/9781608718276.

Skaaning, Svend-Erik. 2018. "Different Types of Data and the Validity of Democracy Measures." *Politics and Governance* 6 (1): 105–16. DOI: 10.17645/pag.v6i1.1183.

Skaaning, Svend-Erik, John Gerring, and Henrikas Bartusevičius. 2015. "A Lexical Index of Electoral Democracy." *Comparative Political Studies* 48 (12): 1491–1525. DOI: 10.1177/0010414015581050.

Weidmann, Nils B. 2022. "Recent Events and the Coding of Cross-National Indicators." *Comparative Political Studies* 57 (6): 921–37. DOI: 10.1177/00104140231193006.

Zhao, Yong. 2012. *World Class Learners: Educating Creative and Entrepreneurial Students*. Thousand Oaks, CA: Corwin.

https://doi.org/10.1017/S1537592725103058 Published online by Cambridge University Press