Quality Assessment of the Academic Freedom Index: Strengths, Weaknesses, and How Best to Use It

Lars Lott and Janika Spannagel

This article reviews the data quality of the first systematic global measurement of academic freedom, the Academic Freedom Index (AFI), by using a data quality assessment approach proposed by McMann et al. (2022). By analyzing three distinct components of data quality (content validity, the data generation process, and convergent validity), we examine the specific strengths and potential shortcomings of the AFI. The findings indicate that the AFI does well in terms of its theoretical embeddedness (within some conceptual limits), of the transparent data generation process, and the handling of expert assessments, as well as of its temporal and spatial coverage. A critical assessment of the level of disagreement between expert coders further shows that there are few systematic predictors, providing no evidence for problematic biases among AFI coders. Overall, we conclude that the data quality of the AFI is comparatively high but that it could be further increased by recruiting even more experts and thereby enhancing the Bayesian IRT model's performance.

Corresponding author: Lars Lott is a Postdoctoral Researcher at Friedrich-Alexander-Universität Erlangen-Nürnberg (lars.lott@fau.de). He is also a Research Associate at the V-Dem Institute. His research interests include authoritarian regimes, democratization and autocratization, the political economy of inequalities and academic freedom. He has published his work in Democratization, Higher Education, Swiss Political Science Review, Studies in Conflict & Terrorism, European Policy Review, *and* Contemporary Politics, *among others.*

Janika Spannagel is a Postdoctoral Researcher at Freie Universität Berlin, Germany, where she currently studies the diffusion and contestation of academic freedom norms at the Cluster of Excellence "Contestations of the Liberal Script (SCRIPTS)" (janika.spannagel@fu-berlin.de). She previously co-developed the Academic Freedom Index. Her background is in researching human rights and political repression. She has published two books, including International Attention and the Protection of Human Rights Defenders: Campaigning for Agents of Change *(Routledge 2023) and* University Autonomy Decline: Causes, Responses, and Implications for Academic Freedom *(Routledge 2022), and in* Quantity & Quality *and* The International Journal of Human Rights, *among others.*

The global challenge of contested academic freedom has gained increasing attention in public discussions as well as the scholarly literature in recent years (Enyedi 2018; Kaczmarska 2020; Kinzelbach et al. 2023; Kinzelbach, Lindberg, and Lott 2023; Lerch, Frank, and Schofer 2024; Lott 2024; Mendes 2020; Taylor, Kunkle, and Watts 2023, among others), observing both severe and subtle encroachments on academic freedom in countries around the world. The creation of the Academic Freedom Index (AFI), which was first released in March 2020, closed a significant gap in the comparative measurement of abstract concepts of governance and democracy. Although "the last three decades have seen a boom in the development of social science indicators and indices" (Croissant and Pelke 2022, 137), the topic of academic freedom, and in particular its exploration as a multidimensional concept, had largely been overlooked. The new AFI data, curated in the Varieties of Democracy dataset, not only constitutes the first conceptually focused and comprehensive measurement approach to academic freedom, but it also offers extensive coverage: it currently provides assessments for 180 countries and territories worldwide, and covers the time since 1900.

The report released alongside the latest data iteration (Kinzelbach, Lindberg, and Lott 2024) shows that 23 countries worldwide are currently in episodes of decline in academic freedom, while the situation is improving in only ten countries. In 2023, 3.6 billion people lived in

doi:10.1017/S1537592724001968

places where academic freedom is very severely restricted and where scholars may experience violent reprisals for demonstrating independence in their research, teaching, or public expression. The dataset helps to shed light on such deeply concerning developments, which have consequences both for the affected populations and for the global science community. Yet the data also allow for systematic research into changes in academic freedom levels over time, as well as their causes and consequences. Since the first release of the AFI in 2020, a range of studies have already made use of the dataset to delve into various aspects of academic freedom, including its global development (e.g., Kinzelbach et al. 2023; Lott 2024), the connection to democracy (e.g., Berggren and Bjørnskov 2022; Kratou and Laakso 2022; Pelke 2023), decline in university autonomy (e.g., Roberts Lyer, Saliba, and Spannagel 2023), and the social foundations of academic freedom in liberal international institutions (Lerch, Frank, and Schofer 2024).

Now that the dataset is released in its fifth iteration (Coppedge et al. 2024) and has established itself as a leading instrument for evaluating and comparing academic freedom levels in countries across the world, it is time to take stock of its quality and performance. Having ourselves participated in the development of the index (Spannagel) or in the subsequent third-party funded project of the AFI (Lott), we are familiar with the ins and outs of the dataset, both from a conceptual and an application standpoint. Our aim in this article is to undertake a comprehensive critical analysis of the data quality of the Academic Freedom Index and provide guidance to users and readers on its strengths and weaknesses, as well as certain design choices, to support further robust scholarly inquiry.[1]

An earlier paper already introduced the new AFI dataset in detail, explaining the rationale behind its indicators and providing a cursory analysis of content and convergent validity (Spannagel and Kinzelbach 2023). A further article discussed alternative measurements and data sources on academic freedom (Spannagel 2020). The present article expands on these contributions and offers a more in-depth evaluation to thoroughly assess the validity and reliability of the AFI. By exploring its specific strengths and shortcomings, it seeks to highlight how the data should best be used and what needs to be taken into account in their interpretation. In doing so, we largely follow the useful step-by-step guide for measurement quality assessments recently introduced by Kelly McMann and her co-authors (McMann et al. 2022), though we aim to complement rather than reproduce evaluations that have already been done elsewhere. The three proposed assessment steps that also structure our article focus on 1) content validity, 2) the data generation process, and 3) convergent validity. One of the key pieces of advice we have for users is that they should take measurement uncertainty into account when using the AFI —for this reason, we finish the article with 4) a practical guide on how to incorporate such uncertainty in statistical analyses.

The content validity assessment addresses the theoretical construct of academic freedom that underlies the AFI measurement, drawing and expanding on the explanations provided in the introductory article (Spannagel and Kinzelbach 2023). The present article further complements this using Bayesian factor analysis to investigate to what extent the different indicators capture the higher-level theoretical concept that the AFI intends to measure (Content Validity Assessment). With regard to the data generation process, we discuss the validity and reliability of how the AFI data are collected and aggregated. Since the AFI is generated in a very similar process as other V-Dem indices, this assessment step is in large parts congruent with the analysis of V-Dem corruption measures provided by McMann et al. (2022) and we will therefore focus on discussing elements specific to the AFI. Special attention will be given to the investigation of coder disagreements and biases in the AFI data (Data Generation Process Assessment). The convergent validity assessment serves to compare the AFI measure to the only (somewhat) comparable measure of academic freedom available today, namely Freedom House's indicator D3 on *academic freedom and the freedom of the educational system from extensive political indoctrination* (Content Validity Assessment). Finally, the last section shows how to handle uncertainty when using the AFI in statistical analyses. Afterwards, we summarize the findings to draw conclusions on the AFI's strengths, weaknesses, and how its distinct characteristics should shape its application (Conclusion).

By doing so, this article contributes to the literature in different ways. First, it provides an in-depth data quality analysis of the AFI showing how to use this measure of a latent concept in (causal) inference studies. Second, it analyzes the data generation process for V-Dem's academic freedom indicators and thereby complements studies on the data quality of the V-Dem dataset (e.g., Knutsen et al. 2024; Marquardt and Pemstein 2023; McMann et al. 2022; Pemstein et al. 2023; Weitzel et al. 2023), which has become a landmark measurement of democracy. Thereby our article adds to an extensive debate about V-Dem's democracy measures (Knutsen et al. 2024; Little and Meng 2024a, 2024b; Treisman 2024; Weidmann 2024). Third, it provides a hands-on guide for how to use the AFI and incorporate measurement uncertainty, which can be also applied to other V-Dem indicators.

Content Validity Assessment

Content validity refers to the extent to which a measurement (the AFI) captures all aspects of a given topic it is designed to measure (academic freedom), including relevant and excluding irrelevant parts. The AFI relies on five indicators that cover different aspects of a country's de facto academic freedom, namely the *freedom to research and teach*,

https://doi.org/10.1017/S1537592724001968

the *freedom of academic exchange and dissemination*, the *institutional autonomy* of higher education institutions, *campus integrity*, and the *freedom of academic and cultural expression.* The introductory paper explains that the indicators were chosen and formulated on the basis of a review of the literature and in-depth discussions with academics, policymakers and academic freedom advocates, and reflects key aspects of academic freedom as defined in international law (Spannagel and Kinzelbach 2023, 3973f.).

Each of the indicators is formulated as a question (e.g., "To what extent are scholars free to develop and pursue their own research and teaching agendas without interference?"), which is presented to country experts together with general definitions, a specific clarification text, and five defined response levels on an ordinal scale from 0 to 4 (see all details in the codebook at Coppedge et al. 2023b, 233–237).

For each country, multiple experts assess the indicator on an annual basis. The ratings of individual coders are aggregated into country-year scores for each indicator (and in a second step for the index) using a customized Bayesian Item Response Theory model (Pemstein et al. 2023) controlling for respondents' individual coding behavior (more details follow).

Review of Conceptual Decisions and Frequent Inquiries

Although the five indicators provide a thorough conceptual framework for the index, the authors concede that the addition of further aspects would have been thinkable (Spannagel and Kinzelbach 2023, 3974). Moreover, several years after the establishment of the dataset and based on our participation in many discussions focused on the AFI, we can identify some questions and criticisms that are frequently directed towards the conceptual composition of the index. In the following, we will address these common points and consider how they may affect the extent to which the AFI measures academic freedom.

In their introductory paper, the authors mention "academics' general job security" (i.e., tenure) as a possible additional indicator, which has formed the focus of other academic freedom studies (e.g., Karran, Beiter, and Appiagyei-Atua 2017). However, conceptually, this aspect arguably falls more under an enabling condition or even a proxy measurement than representing an aspect of academic freedom itself. A similar argument of being an enabling condition could in fact be made regarding the institutional autonomy of higher education institutions as well as campus integrity, which, unlike tenure, are included in the AFI. Tenure would, however, be a far more specific indicator that can limit global comparability given the diversity of higher education sectors and the varying role that tenure plays across these different contexts. Autonomy and campus integrity, on the other hand, appear universally relevant to the protection of academic freedom and can be considered as integral parts of its multidimensional conception. That said, users of the AFI that wish to focus on academic freedom more narrowly as an individual-level right can decide to exclude the institutional indicators when working with the data (see further exploration of this point in the Factor Analysis of AFI Indicators section).

A second potentially omitted aspect implied by the authors is the diversity of, non-discrimination in, and equal access to higher education. They justify the omission of such aspects by a focus on aspects that are "specific to the academic sector," arguing that discrimination is likely to extend beyond the higher education sector and would thus be captured by other indicators in the V-Dem dataset that may be used to complement the AFI (Spannagel and Kinzelbach 2023).[2] Apart from this practical argument, one could also argue that the issue of discrimination does not itself describe the level or *quality* of academic freedom in a given country, but rather who benefits from it and who is excluded. At the same time, such a viewpoint is more precarious if one sees academic freedom as a good that should (at least potentially) benefit everyone, not just the privileged few. In this perspective, the lacking assessment of whether interference and restrictions are distributed unevenly across different groups based on gender, race, or other characteristics (or whether such groups are systematically excluded from or disadvantaged in academia to begin with) presents a gap in the measurement, which could potentially be filled with the addition of a new indicator. In the meantime, some users may choose to complement the AFI with alternative measures of exclusion.

A similar concern relates to the issue of funding. It could be argued that where the funding available for higher education is very low, there cannot be meaningful academic freedom. While it is true that such underfunded higher education sectors typically have little capacity to do research at all (e.g., Altbach 2016; Sawyerr 2004; Zavale 2022), there is arguably a difference between the retaliatory suppression, redistribution, or overall conditionality of funding motivated by political or economic interests—as captured by several AFI indicators—and the mere absence of resources for academic research. The former category of budgetary pressures and retaliations are decidedly within the scope of the AFI, since they are typical instruments of outside interference with research and teaching, exchange and dissemination, as well as university autonomy. The latter category of resource scarcity, however, is not covered by the AFI. In this discussion, one should be mindful of the fact that research funding is not unlimited in any country and that scholars everywhere are faced with the need for prioritization, although to varying degrees. While this consideration is not meant to relativize such differences, it points to the immense difficulty in establishing a common metric that would do justice to this problem.

https://doi.org/10.1017/S1537592724001968

Taking both aspects of non-discrimination and funding together, we can conclude that overall, the AFI tends to cover more a negative understanding of academic freedom (the absence of infringements) rather than a positive one (an active promotion of academic freedom by the state and other actors)—a key aspect of its conceptualization that users should bear in mind.

A further noteworthy omission is the aspect of student rights, in terms of the freedom of learning and the right to participation in university governance. The AFI indicators capture this aspect only indirectly through the freedom of teaching, as well as campus integrity, which describes the absence of surveillance and security infringements on campus. In fact, definitions of academic freedom disagree on whether student rights form a core part of academic freedom or whether academic freedom primarily relates to the rights of scholars (cf. Abdel Latif 2014; Macfarlane 2012). In this sense, the AFI in its current form should be understood as capturing academic freedom mostly in the latter sense. The inclusion of an additional indicator on students could be envisaged for the future—especially if there are plausible expectations that there are cases in which students' freedom diverges significantly from that of scholars.

Another inquiry often made in connection with the AFI's conceptualization is whether it captures the pressures that scholars find themselves under in the context of increasing third-party and performance-based funding, as well as concomitant trends of "managerialism" at universities (e.g., Butler and Mulgan 2013; Puaca 2022). The answer to this question is clearly affirmative from a conceptual standpoint, since the absence of "interference" measured by different indicators refers to influence exerted by "non-academic actors," defined as "individuals and groups that are not a scientifically trained university affiliate," including "individuals and groups such as politicians, party secretaries, externally appointed university management, businesses, foundations, other private funders, religious groups and advocacy groups" (Coppedge et al. 2023b, 233). In practical terms, however, the ability of the AFI to capture these issues may be limited. On one hand, this is due to the global scope of this measurement effort, where the effects of such shifts may be small compared to other types of interference and therefore not be reflected in the data outside the margins of statistical uncertainty. On the other hand, limitations also stem from the fact that the lines between academic and non-academic actors may in the reality of higher education governance be more equivocal than the definition suggests. This complex issue also raises the question whether academic actors, when taking decisions on science governance, procedures, and contents, do always act in the interest of academic freedom. Yet such issues are contentious and highly context-dependent, so that a global comparative measurement like the AFI can hardly be expected to systematically capture them in adequate detail. Moreover, an advantage of keeping the measurement rather narrow and well defined (rather than incorporating various higher education trends that may affect academic freedom) is that it remains useful for those who seek to empirically investigate linkages between academic freedom and trends such as corporatization.

Finally, the lack of disaggregation between academic disciplines as well as higher education institutions is also an occasionally raised concern. Overall, experts are asked to generalize in their assessment across universities and across disciplines for a given country. While on the surface this presents as a question relating to methodology and data collection, the conflation of disciplines in particular has potential ramifications on the conceptual understanding of academic freedom. On this point, Spannagel and Kinzelbach (2023) argue substantively that while it would be misleading to focus only on the worst-off subject areas, it would also be problematic to focus on the better-off subject areas and thus relativize the interference in some disciplines by the freedom of others, when the integrity of academia as a whole is at stake. The AFI authors instead chose a middle ground by factoring potential disciplinary variation into the response scales for the two indicators where this is most pertinent: the *Freedom to research and teach* and the *Freedom of academic exchange and dissemination.* Their lowest two levels distinguish between interference that is consistent "across all disciplines" or consistent "in some disciplines," whereas the other three response levels describe them as occurring "occasionally," "rarely," or "not," regardless of disciplinary variation. In terms of the aggregation across institutions, the problem seems less one of validity than of reliability: by giving experts additional leverage in deciding how to weigh situations at different institutions in the same country, this may increase coder disagreement and introduce additional uncertainty into the measurement. We will further address the level and potential sources of coder disagreement later (see Analyzing Respondent Disagreement).

On the whole, academic freedom remains a latent construct that cannot be measured directly by statistical indicators or objective measures. The five AFI indicators overall present a coherent picture and, several years after their formulation, still seem to strike a legitimate balance between conceptual specificity and global comparability, as well as between conceptual comprehensiveness and finite resources.

Factor Analysis of AFI Indicators

The Academic Freedom Index was built using a Bayesian factor analysis (BFA) model to aggregate the different dimensions of academic freedom and to incorporate measurement uncertainty into the metric. The AFI ranges from 0 (no academic freedom) to 1 (full academic

https://doi.org/10.1017/S1537592724001968

freedom) and consists of point estimates from the BFA accompanied with uncertainty measures. We discuss the choices in terms of index-level aggregation in the Index Level Aggregation, but first want to show how the five different dimensions reflect one underlying systematized concept. For information on replication materials, please refer to Lott and Spannagel (2024).

For this analysis, we use BFA that allows us to incorporate measurement error "in the manifest variables, which themselves were estimated using Bayesian methods, into the model" (McMann et al. 2022, 433). With the BFA, we assess how the five V-Dem academic freedom measures empirically relate to one another. By showing that the five dimensions do in fact empirically load onto an index, we can provide strong empirical support for the theoretical claim of a single underlying systematized concept of academic freedom (compare also McMann et al. 2022, 433).

Table 1 shows that all five indicators strongly load on a single dimension.[3] Factor loadings in table 1 indicate how much of the Academic Freedom Index is explained by a particular indicator, while its uniqueness score is the variance that is not shared with the other indicators (i.e., that is unique to that indicator). For example, a uniqueness of 0.169 for the *Freedom to research and teach* indicator shows that 16.9% of its variance is not shared with the other indicators in the BFA. The higher the factor loading and the lower the uniqueness score for individual indicators, the stronger the empirical evidence that the specific indicator relates strongly to the underlying concept.

Overall, table 1 indicates that the fit to a unidimensional model is adequate as all indicators have strong factor loadings and a large share of their respective variance is accounted for (low uniqueness).[4] *Freedom to research and teach* loads the most strongly on a single dimension with a factor loading of 0.912. The factor loading of *Freedom of academic and cultural expression* is comparatively weaker (but still loads strongly with a factor of 0.814) and a rather large share of variance is unaccounted for. This is not surprising, and indeed desirable,[5] as it conceptually deviates somewhat from a strict academic freedom perspective by focusing on academics' (and artists') freedom of expression.[6] Some users of the data wishing to focus on academic freedom more narrowly may therefore choose to exclude this indicator.

Table 1
Conceptual alignment across V-Dem academic freedom indicators (BFA estimates)

Measure	Loadings	Uniqueness
Freedom to research and teach (**v2cafres**)	0.912	0.169
Freedom of academic exchange and dissemination (**v2cafexch**)	0.912	0.169
Institutional autonomy (**v2cainsaut**)	0.829	0.314
Campus integrity (**v2casurv**)	0.853	0.273
Freedom of academic and cultural expression (**v2clacfree**)	0.814	0.338

In addition, we test whether a two-factor model explains more variance in the manifest variables than the unidimensional factor model by using frequentist factor analysis presented in tables B1 and B2 (in the online appendix). We assume that *Institutional autonomy* and *Campus integrity* load on one dimension, while *Freedom to research and teach*, *Freedom of academic exchange and dissemination*, and *Freedom of academic and cultural expression* load on a second dimension, as the latter represent individual freedoms and the former represent institutional rights of universities that protect them from outside interference. In addition, as noted earlier, a strand of literature argues that institutional features of universities, in particular their autonomy, are a prerequisite for academic freedom (e.g., Matei and Iwinska 2018; Nokkala and Bladh 2014). Within the frequentist factor analysis framework, the one-dimensional model fits the data slightly better than our two-dimensional model, supporting the idea of a complex approach to academic freedom that includes both individual and institutional aspects. Later we test different index aggregation models and discuss theoretical assumptions. However, even though we reject the hypothesis that the two-factor model explains the data best as indicated by the slight improvement of about 0.007% in model fit (1-dim BIC = 166629, 2-dim BIC = 166520), the only very marginal improvement provides justification for both models, depending on the researcher's theoretical assumptions. In sum, the BFA that takes into account the measurement uncertainty provides strong empirical support for the AFI's content validity: all indicators largely reflect a single underlying systematized concept, namely academic freedom.

Data Generation Process Assessment

The validity and reliability of the way in which the data are generated determines whether or not one will obtain an unbiased and reliable measure. McMann et al. (2022, 431) note that, unlike the correctness of individual scores, the quality of the data generation process can actually be observed and evaluated, making this a valuable criterion for the quality of the resulting measurement. In assessing the Academic Freedom Index, we are analyzing a measure that is generated as part of the same data collection effort as the corruption measures that McMann et al. (2022) evaluated in their paper. We will therefore quickly go over the shared characteristics that were already discussed in

https://doi.org/10.1017/S1537592724001968

their paper, and focus on the elements that are specific to the AFI indicators.

Data Management Structure and Data Sources

In terms of data management structure, McMann et al. (2022, 434) highlight positively that unlike many alternative data collection efforts, V-Dem is an entirely academic endeavor, headquartered at the University of Gothenburg, Sweden, and led by an international consortium of scholars based in different locations around the world (Varieties of Democracy Project 2022).

Regarding the sources used for compiling a given measurement, using expert-coded assessments appears by far superior to alternative measurement approaches for generating comparative country-level assessments. Surveys inquiring about academics' personal experiences, for instance, are likely to suffer much more from a self-selection bias than a pool of experts selected on the basis of their expertise (see more on this later). Events-based data, another frequently used data source on academic freedom that records incidents of academic freedom violations, also suffer from a whole range of selection biases, are typically non-representative, and generally unsuitable to assess less repressive (and even the most repressive) contexts. Other potential sources include data collected through institutional self-reporting, as well as legal analyses. All those data sources have specific advantages, but none are suited for comparative assessments of de facto academic freedom at a global scale (see detailed discussion in Spannagel 2020). Furthermore, McMann et al. (2022, 435) argue that "datasets that aggregate information from different sources multiply biases and measurement errors by including those from each source in their composite measure, particularly if measurement errors across data sources are correlated"—a problem that is avoided when relying on one consistent expert data collection effort.

That said, McMann et al. also caution against the fact that V-Dem country experts often respond to several questions that relate to the same measured concept across V-Dem's expert surveys, creating a potential to generate correlated rater error across indicators (McMann et al. 2022, 435). In addition, such "correlated errors could undermine other aspects of our quality assessment, such as the factor analysis in our content validity analysis" (McMann et al. 2022, 435), and also implies that researchers should avoid putting indicators relating to the same concept on both sides of a regression equation. In online appendix D, we analyze how raters' errors correlate across indicators. The findings show that while raw errors in rater scores correlate highly across expert ratings, this appears to stem largely from differential item functioning (DIF).

The Freedom House D3 indicator—the only other expert-coded academic freedom measure available—does not pose this problem since it includes only one single indicator on academic and educational freedom. The AFI's methodologically and conceptually more advanced approach makes it more vulnerable to such issues, while at the same time avoiding a set of important pitfalls from the Freedom House approach. In addition, the AFI was constructed in a way that its indicators are spread across two different expert surveys, which are not necessarily coded by the same experts. As of version 13, the *freedom of academic and cultural expression* indicator (from the *Civil Liberty* Survey) is based on the assessment of 1,838 distinct coders, while the assessments of the four remaining indicators from *Civic and Academic Space* survey are based on up to 1,130 distinct coders. Up to 747 experts rate questions from both surveys.[7] 2,197 experts in total had contributed to the AFI indicators as of version 13.

Expert Characteristics and Qualifications

Next, we need to evaluate the coding procedures, addressing first the characteristics and qualifications of the AFI's pool of country experts. The identities of experts contributing to the V-Dem project are kept anonymous and only few program managers and V-Dem's Institute Director have access to this information (Coppedge et al. 2020, 61). However, researchers with a well-founded interest can apply for confidential access to some coder-level characteristics, such as gender or education level, as we did for this article.

In a first step, we provide in table 2 some basic descriptive statistics for the pool of 2,197 country experts coding the indicators in the AFI, although they are only available for roughly three-quarters of experts on most items.[8] It indicates that at least 53% of the expert are men and at least 21.2% are women, while the gender is unknown for 25.8%.[9] In terms of level of education, a majority of at least 55.6% of coders have a PhD degree, at least 15.3% have a master's degree, while at least 4.9% have no PhD or master's degree. In addition, the vast majority of experts are *not* government employees (at least 88.1%).[10] Regarding the age of the experts, we can see that at least 6.6% are younger than 35 years, while 36.2% of the experts indicated they are between 35 and 49 years old. At least 31% percent of the experts are older than 49 years. Moreover, at least 43% of experts reside in, and at least 51.6% of experts were born in, the main country that they code (see also table 2). Overall, at least 56.8% of experts can be considered as local experts, as they are either residing in or were born in the main country they are coding.[11]

Next, we need to address the question of the experts' qualification to code academic freedom issues. V-Dem's selection of experts follows rigorous criteria, which are discussed in Coppedge et al. (2020, ch. 3.8). In sum, experts are recruited along the following criteria: 1) experts' expertise on the country or countries and surveys

https://doi.org/10.1017/S1537592724001968

Table 2
Descriptive statistics of the expert sample, based on 2,197 distinct experts

		N	%
Gender	Men	1165	53
	Women	465	21.2
	unknown	567	25.8
Age	< = 34 years	146	6.6
	> 34 years and < 50 years	795	36.2
	> = 50 years	680	31
	unknown	576	25.8
Reside in Main Country Coded	Yes	944	43
	No	716	32.6
	Unknown	537	24.4
Born in Main Country Coded	Yes	1134	51.6
	No	530	24.1
	Unknown	533	24.3
Reside in or born in Main Country Coded	Yes	1247	56.8
	No	419	19.1
	Unknown	531	24.1
Education level	PhD	1221	55.6
	Master's degree	336	15.3
	No PhD or master's degree	107	4.9
	Unknown	533	24.3
Government employee	No	1936	88.1
	Yes	67	3
	Unknown	194	8.8

they may be assigned to code is the most important selection criterion; 2) the connection to the country, so that at least three of five experts per country were born in or reside there; 3) "prospective coder's seriousness of purpose, i.e., her or his willingness to devote time to the project and to deliberate carefully over the questions asked in the survey"; 4) experts' impartiality; and 5) diversity in professional backgrounds.

In the context of human rights measurement, some have challenged V-Dem's approach to expert selection, arguing that because most respondents have a PhD degree and are therefore likely to be academics, they may not be the best possible experts on human rights abuses—as opposed to human rights advocates, researchers, and lawyers (Brook, Clay, and Randolph 2019, 19). Yet when it comes to assessing academic freedom issues specifically, the fact that the V-Dem pool of respondents consists largely of academics arguably makes them especially qualified. Yet the composition of the country experts pool in terms of their field of study may raise some questions. The V-Dem selection process was originally designed for recruiting social scientists, in particular political scientists, historians, and legal scholars. These experts may be considered "generalists" who are not necessarily higher education scholars and experts in evaluating academic freedom issues. That being said, over the last years, the AFI team has made substantial efforts to diversify the professional background of AFI experts by recruiting large numbers of contributors with specialized expertise.[12] Moreover, as noted earlier, "obtaining diversity in professional background among the coders chosen for a particular country" (Coppedge et al., 2020, p. 59) has always been part of V-Dem's recruitment strategy. This diversity not only includes a mixture of academics and professionals, but also "experts who are located at a variety of institutions, universities, and research institutes since people in institutions sometimes develop a particular collective perspective" (Coppedge et al. 2020, 59).

In addition to the selection criteria, it is important to highlight that with typically at least five expert coders per indicator per country-year (true for 99.58% of country-years), a single respondent's biases cannot drive the resulting estimates (McMann et al. 2022, 436). In its version 13, the AFI rests on assessments by 2,197 coders across the world and across indicators, which translates into an average of 10.56 distinct expert ratings per country-year of the AFI (Min = 3, Max = 31, Median = 10).[13] These numbers and the level of transparency on the data collection process are in stark contrast to Freedom House's approach, which relies on a total of only 128 analysts, i.e., on average 0.61 for each of the 210 countries/territories, and 50 advisers who weigh in as part of a rather vaguely described review process without specific information on the expert selection (Freedom House 2022, 2).

The AFI's reliance on several independent experts per data point mitigates the issue of individual biases to a great extent. However, one could argue that experts residing inside or outside a country, for instance, may rely on different information and assign varying importance to the same information (see also Knutsen et al. 2024; McMann et al. 2022; Pemstein et al. 2023, among others). In fact, all human coders are likely to make inadvertent coding errors, and may base their judgments on irrelevant issues (Weitzel et al. 2023, 8–10), recent events (Weidmann 2024), and historical biases (Weitzel et al. 2023, 8–10). That being said, systematic errors that depend on the personal exposure of experts to academic freedom issues in a given country could be a source of collective bias and introduce systematic error into the AFI data depending on the number of local experts per indicator and country-year.[14] This does not appear to be the case for the AFI, as we will show in detail later.

Indicator Level Aggregation

From the collection of several independent experts' assessments per indicator-country-year follows the need to aggregate them into single indicator-level scores. Importantly, this is not simply done by averaging the individual scores, as this would presuppose that all contributing experts are equally certain about their scores, that they are equally (un)biased, and that they exhibit the exact same coding behavior when confronted with an ordinal scale. These are unrealistic assumptions in any survey context,

https://doi.org/10.1017/S1537592724001968

but arguably even more so when involving experts from countries all over the world (Church 2010). For this reason, V-Dem uses a customized statistical model that relies on Bayesian Item Response Theory (IRT) to aggregate the coder-level scores (Coppedge et al. 2023; Pemstein et al. 2023), an approach that has been shown to outperform the use of simple averages (Marquardt and Pemstein 2018). This IRT model accounts for experts' varying reliability as well as for differential item functioning (DIF), which occurs when experts differ in their perceptions of multi-item scales.

> For example, if respondents provide ordinal ratings and they vary in how they map those ratings onto real cases—perhaps, for example, one respondent has a lower tolerance for corruption than another—then a process that models and adjusts for this issue will outperform a more naive process." (McMann et al. 2022, 436)

Overall, the Bayesian IRT model is able to measure latent—not directly observable—concepts, such as the freedom to research and teach, and provide reliable and comparable expert assessments "while allowing for the possibility that respondents apply ordinal scales differently" (McMann et al. 2022, 436).

In addition, the model uses information from bridge coding (an expert rates multiple countries for many years), lateral coding (an expert codes many countries for one year), and anchoring vignettes (description of hypothetical cases that are rated by experts) to improve the model estimates and comparability within and across countries. The anchoring vignettes are especially useful, because there is no contextual information and all respondents rate the same set of vignettes under a controlled environment. In this way, ratings on these vignettes provide information about how experts understand the ordinal scale and "how they systematically diverge from each other in their coding" (McMann et al. 2022, 436).

McMann et al. (2022, 436) rightly stress that there is no respondent who is free of bias and no expert pool that does not exhibit DIF. However, the approach chosen by V-Dem is specifically designed to address these problems and reduce their imprint on the resulting dataset. In contrast to Freedom House, V-Dem also provides full transparency on the coding process and aggregation procedures. Next to detailed methodological papers, this includes that all individual coder-level ratings are publicly available on the V-Dem website (i.e., data before aggregation by the V-Dem IRT model). Moreover, the final model estimates for each indicator (and the index) are accompanied by upper and lower uncertainty bounds. Roughly speaking, there are two main sources of uncertainty: 1) the indication by experts of a lower level of confidence in their scores when providing their ratings, and 2) the disagreement between the expert coders who assessed the same data point. We will look into the latter in detail later (Analyzing Respondent Disagreement).

Index Level Aggregation

To combine low-level indicators to higher-level measures (indices), V-Dem typically uses a Bayesian factor analysis (BFA) model, when concepts are considered a latent construct, such as components of democracy. Since there are no objective standards of aggregation into a higher-level measure, the most important consideration is the researcher's theory on how the indicators correspond to one another. As Coppedge et al. (2020) discuss, an important consideration for the aggregation is whether indicators are treated as reflective or formative indicators. Reflective indicators are "symptoms of the concept being measured, and are typically estimated by factor analysis" (Coppedge et al. 2020, 91; see also Treier and Jackman 2008), while formative indicators treat "indicators as determinants of the concept being measured" (Coppedge et al. 2020, 91). When indicators are formative, then the specific aggregation choice depends on whether indicators are treated as (partially) mutually substitutable aspects of a given concept (additive aggregation rule) or as individually necessary conditions for it (multiplicative aggregation rule).

The authors of the AFI conceptualized the five indicators as jointly reflecting the latent concept of "academic freedom," where none of the indicators takes precedence over another (see also the discussion in the Factor Analysis of AFI Indicators section). Instead of averaging or multiplying the indicator scores, the AFI therefore uses V-Dem's standard BFA model to aggregate the five indicators into the index. Similarly to the indicator-level aggregation, the BFA model provides measures of uncertainty alongside the index estimates.

Different aggregation choices could legitimately be made depending on researchers' theoretical assumptions or conceptualization of academic freedom, and the individual indicators are available for others to construct their own higher-level academic freedom measure. To show that the aggregation rules are important considerations that affect the outcome measurement, we briefly discuss how the measures resulting from different choices are intercorrelated: the reflective mode of the AFI, an additive academic freedom index, and a multiplicative academic freedom index (conceptualizing institutional autonomy as a necessary condition). Table 3 shows the correlation between these different measures on a country-year level. It indicates that the additive AFI, the multiplicative AFI, and the original AFI are highly correlated. Further findings presented in online appendix C support the original AFI conceptualization and aggregation as they show that the additive aggregation of indicators does not discriminate appropriately at high and low levels of academic freedom compared to the original AFI. Moreover, the multiplicative aggregation approach assigns systematically lower scores compared to the original AFI scores across the whole distribution of scores. As theoretically expected, the multiplicative AFI is more demanding as it formulates the

https://doi.org/10.1017/S1537592724001968

Table 3
Correlation between indices with different aggregation rules

Measure	Pearson's Correlation Coefficient	p-value
AFI and additive AFI	0.986	< 0.001
AFI and multiplicative AFI	0.966	< 0.001
Additive AFI and multiplicative AFI	0.968	< 0.001

institutional autonomy indicator as a prerequisite for academic freedom.

Coverage across Countries and Time

The spatial and temporal coverage of social science indicators are important criteria for the quality of a dataset in view of analyzing phenomena across time and space. As McMann and coauthors note, a reduced number of cases —for instance, the ones that are easier to code—can lead to problematic selection bias. As a result, "maximizing case coverage also improves measurement validity" (McMann et al. 2022, 437), provided the overall data quality is good. Relying on a broad temporal and spatial sample thus reduces potential biases that may result from a short time frame, a small spatial coverage, or a combination of both.

With 180 countries/territories and 123 years covered as of version 13 (a total of 14,976 country-years), the AFI performs exceptionally well in this regard when compared to any other available data source on academic freedom, including the Freedom House measure. Although the latter has comparable global coverage,[15] the AFI is in fact the only data source on de facto academic freedom[16] that reaches far back in time, covering years since 1900 (or since countries or their higher education system came into existence), whereas Freedom House's D3 indicator is only available since 2012.

And last, it is also important to note that V-Dem's data collection and aggregation procedures are consistent for the whole dataset, which is re-released with each annual update. In contrast, Freedom House's methodology has changed over the years—even if only slightly—and as each release only adds the newest year, such changes can create comparability issues over time.

Analyzing Respondent Disagreement

As McMann et al. (2022) argue, the analysis of coder disagreement and biases is a tool to assess the validity and reliability of the data generation process. First, a measure is more reliable when inter-coder disagreement is low. In addition, a low inter-coder disagreement can also indicate the validity of a measure "if one is willing to assume that multiple respondents are unlikely to exhibit identical biases" (McMann et al. 2022, 438). Second, systematic biases in the data can be assessed by analyzing how respondent and country characteristics, such as gender, education and country of residence of a respondent, as well as socioeconomic background factors and general access to information, predict respondents' ratings.

For the Academic Freedom Index, we assess respondent disagreement using a regression framework. In figures 1 and 2, we estimate the effect that different country characteristics, as well as the number of respondents, have on *respondent disagreement.* The dependent variable is the standard deviation of raw ratings among respondent for each country and year. In contrast to McMann et al. (2022), we use the raw ratings among respondents instead of the measurement model-adjusted ratings among respondents.[17] By using these raw scores, we conduct a more conservative test for analyzing respondent disagreement than McMann et al. because we do not account for corrections made by V-Dem's Bayesian IRT model.

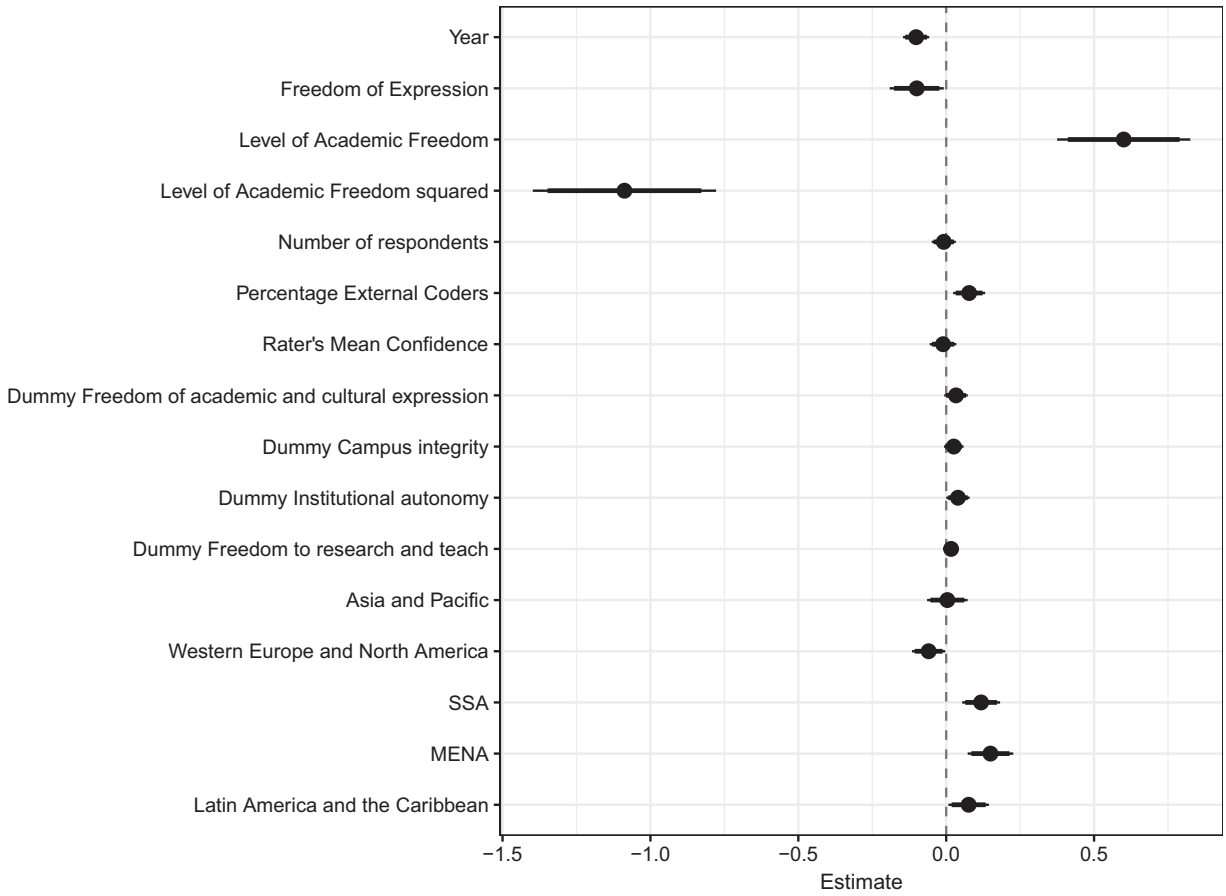
Figure 1 displays the standardized regression coefficients[18] that show the estimated effect size of the respective variable on the level of respondent disagreement. Thus, a positive regression coefficient indicates that the variable is associated with increased respondent disagreement, while a negative regression coefficient indicates lesser respondent disagreement. Table E1 in the online appendix shows the five separate indicators of the AFI, while figure 1 shows the pooled model, controlling for indicator-fixed effects.

Figure 1 indicates that respondent disagreement varies slightly depending on the freedom of expression in the country coded, suggesting that limited access to information may also affect coders' ratings. Specifically, respondent disagreement is lower in countries with high levels of access to information (standardized coefficient = –0.1, 95% CI = [-0.191, -0.008]). We further control for the number of respondents per indicator and country-date and find that it is not associated with the overall disagreement level between coders (standardized coefficient = –0.008, 95% CI = [-0.049, 0.032]). Moreover, figure 1 shows that the percentage of external coders (not living in the country coded) increases the respondent disagreement slightly by 0.078 (95% CI = [0.024, 0.132]). This finding is statistically significant and could be reflective of a slightly lower level of familiarity with the country among external coders as opposed to those residing in the country. Moreover, we test for the raters' mean confidence and find that it does not substantially affect the respondent disagreement (standardized coefficient = –0.011, 95% CI = [-0.056, 0.034]).[19]

In addition, we also test whether the year for the coded country affect coders' disagreement. Figure 1 and table E1 reveals that coder's disagreement is lower for earlier than for recent years (standardized coefficient = –0.102, 95% CI = [-0.146, -0.057]). This result may be surprising given the general idea that "the distant past is harder to code than

https://doi.org/10.1017/S1537592724001968

Figure 1
Predicting respondent disagreement (pooled model)



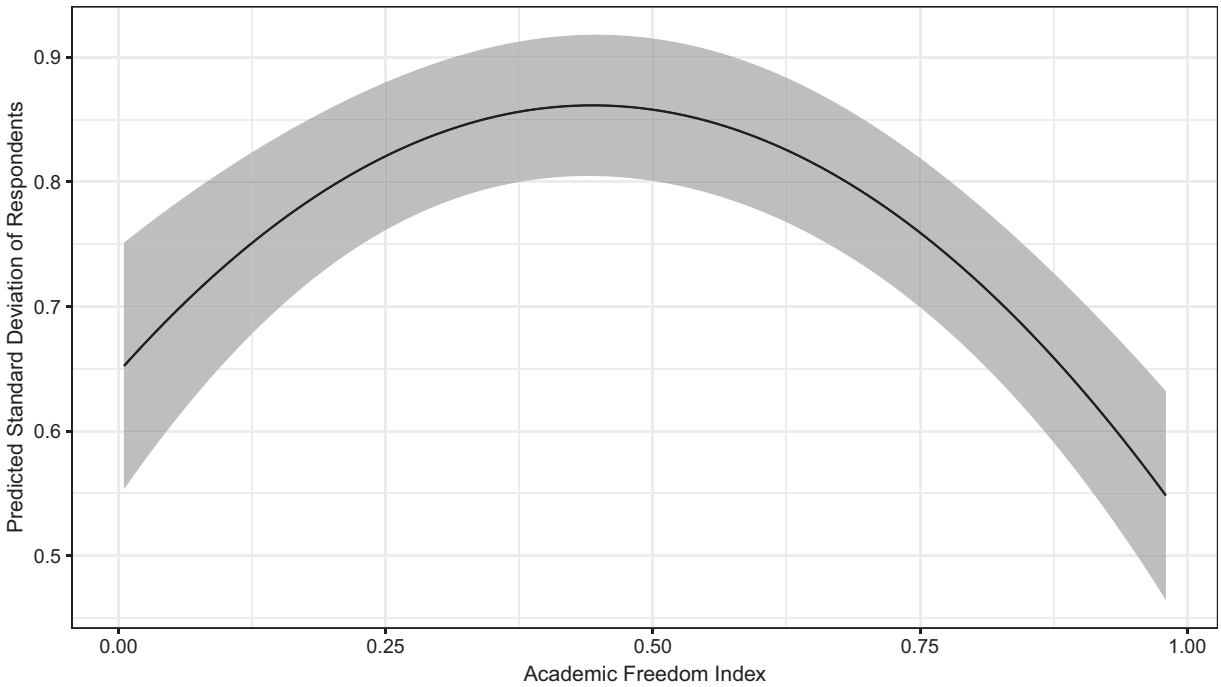
OLS regression with standard errors, clustered on countries. Measure fixed effects are included in the model but omitted from the figure.

the present" (McMann et al. 2022, 438). However, McMann et al. also did not find evidence for this claim. In reality, this is more a matter of perspective and likely depends on the specific pool of experts recruited for the coding. One could in fact plausibly argue that coders are more likely to overestimate the importance of specific events when they code recent years, whereas the larger pattern might become clearer with temporal distance.[20] Knutsen et al. (2024, 166ff) find no evidence for increased pessimism in recent years (also called recency bias) in V–Dem's expert-coded data analyzing V-Dem's indicators for the *Electoral Democracy Index*. Figure 1 further controls for regional effects and indicates that—compared to Eastern Europe and Central Asia—respondent disagreement is larger in Sub-Saharan Africa and MENA, while respondent disagreement is lower in Western Europe and North America. The reasons for these regional differences in respondent disagreement may lie in more time-series fluctuations of academic freedom in some regions compared to others. In other words, volatile academic freedom situations are more likely to generate different experts assessments.

We test also for a nonlinear relationship between academic freedom levels and respondent disagreement by using the quadratic term for the level of academic freedom (standardized coefficient = −1.09, 95% CI = [-1.398, -0.778]). The results, which are plotted in figure 2, indicate that the greatest disagreement between respondents occurs, in fact, in countries with an Academic Freedom Index between 0.25 and 0.6, while the disagreement is lowest in countries with well-protected academic freedom. This shows that mid-levels of academic freedom are most challenging for experts to assess. This may result in more volatile point estimates represented by higher uncertainty intervals. This finding is not particularly surprising: where freedom levels are very high, information availability is likely to be very good, and experts can be relatively confident that relevant issues are known to them. Concurrently, we would also expect a comparatively high agreement between experts. Although very low levels of

https://doi.org/10.1017/S1537592724001968

Figure 2
Predicted respondent disagreement by AFI



OLS regression with standard errors, clustered on countries.

academic freedom are also comparatively easy to identify, the agreement might be somewhat lower because information about pockets of relative freedom is less systematically available and might be assessed differently by different experts. The distinction among middle-lower range freedom levels requires arguably the most in-depth knowledge about the country situation and the most complex decisions on how to factor existing spaces of relative freedom into the overall score, resulting in higher disagreement among coders. Indeed, other research also indicates that that mid-levels of a concept are generally harder to code than cases that show a clear low or high pattern (cf. Coppedge et al. 2020, 160–162). That being said, we also find that this non-linear relationship disappears when using coders' perceptions instead of raw coder assessments. These coders' perceptions control for Differential Item Functioning, which may explain the non-linear relationship between academic freedom and respondent disagreement.

In sum, our findings show that respondent disagreement is not at a critical level and that disagreement varies with the level of academic freedom in ways that we expected.

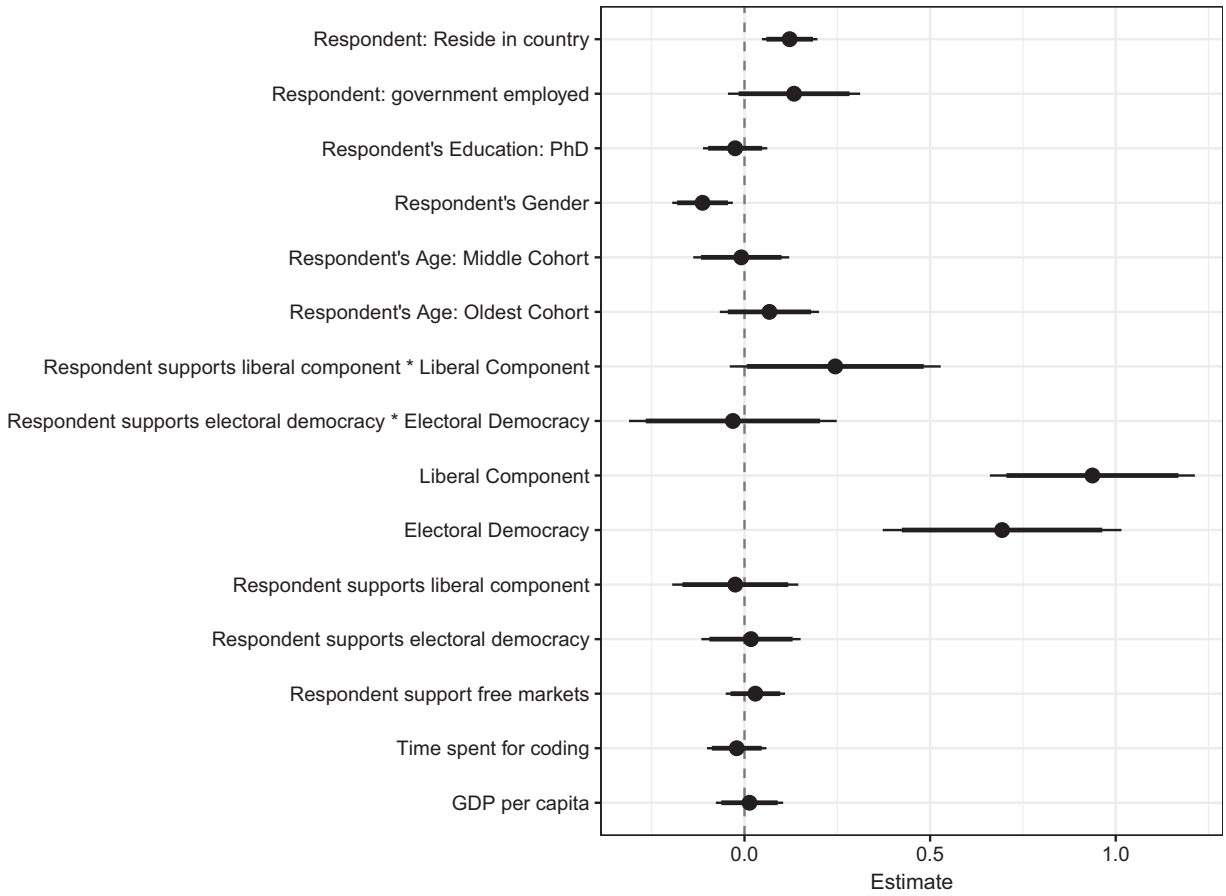
Analyzing Individual Respondent Biases

In the next step, we analyze whether there are systematic biases in the Academic Freedom Index. We first test for what Bollen and Paxton (2000) call "situational closeness" before we evaluate whether there is systematic bias resulting from different coder characteristics. The situational closeness thesis assumes that experts are influenced "by how situationally and personally similar a country is to them" (Bollen and Paxton 2000, 72). To evaluate biases resulting from different respondent characteristics and country characteristics, as well as situational closeness, we use the V-Dem post-survey questionnaire. Figure 3 shows the effects on respondents' ratings of their views of markets and democracy, combined with the coded country's regime characteristics. More concretely, we evaluate whether respondents provide different ratings for academic freedom depending on whether they support a) the principles of electoral democracy, b) the principle of liberal democracy, or c) free markets. We also test for a number of other individual-level factors that may influence respondents ratings (see table 2 for distributions). We further control for the time experts spent on coding.[21]

Figure 3 shows the standardized regression coefficients for the pooled regression analysis with the respondent ratings as the dependent variable. Thus, the point estimates for each explanatory variable (plotted at the y-axis) shows the standardized effects on respondent raw ratings, while the bars represent the 95% and 90% confidence intervals. A positive coefficient indicates a systematic positive effect of this characteristic on the respondents'

https://doi.org/10.1017/S1537592724001968

Figure 3
Predicting respondent ratings with respondent and country characteristics (pooled model)



OLS regression with standard errors, clustered on countries. Measure-fixed effects, year-fixed effects are included in the model but omitted from the figure.

rating, while a negative coefficient indicates that a respondent rates the respective country-date systematically lower. Overall, systematic biases affect the data when the regression coefficients are statistically significant.

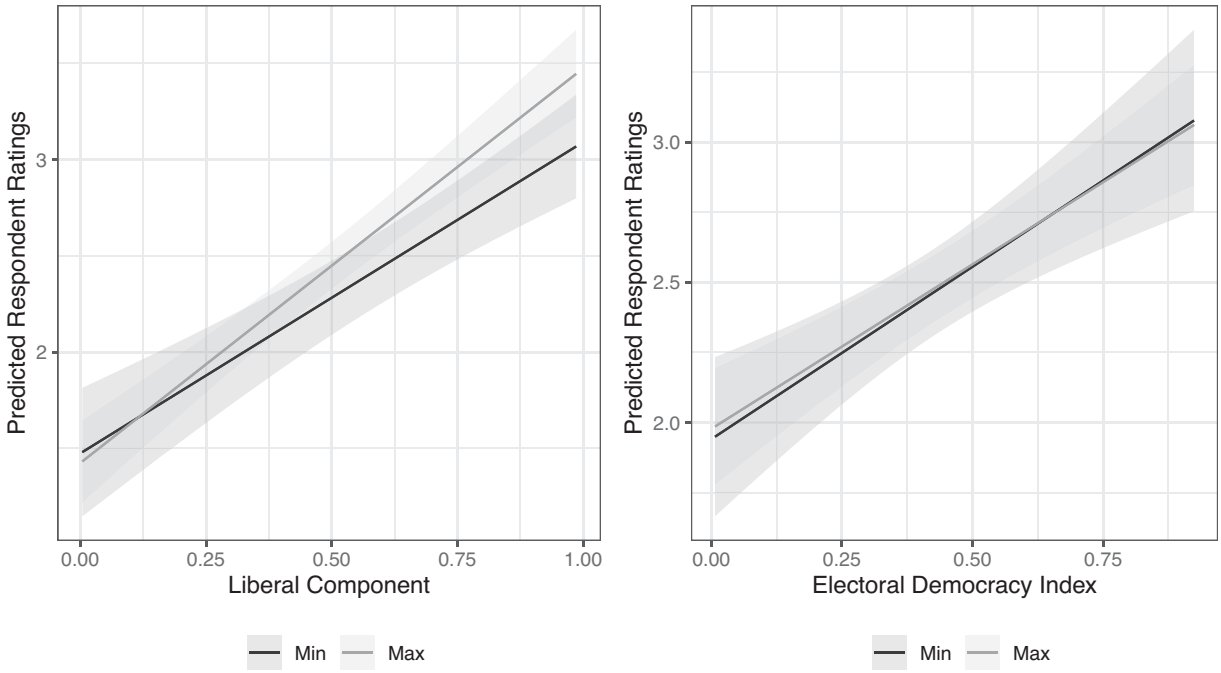
The results shown in figure 3 and table E2 in the online appendix indicate that respondents' situational closeness to the country coded does not result in systematic biases in their ratings. Specifically, figure 3 reveals that neither respondents' support for free markets (standardized coefficient = 0.029, 95% CI = [-0.051, 0.109]), nor respondents' support for electoral democracy (standardized coefficient = 0.017, 95% CI = [-0.116, 0.151]) or liberal democracy (standardized coefficient = –0.025, 95% CI = [-0.194, 0.145]) affect their ratings—indicated by the small and statistically insignificant effects. Figures 3 and 4 do, however, indicate that respondents rate more democratic (standardized coefficient = 0.694, 95% CI = [0.372, 1.016]) and liberal countries (standardized coefficient = 0.937, 95% CI = [0.661, 1.213]) as having more academic freedom, as one would expect. Yet none of the interactions between respondents' views of democracy and the country's regime characteristics are substantially meaningful or statistically significant at the 0.05 level.[22] The positive effect of electoral democracy and the liberal component of democracy does therefore not indicate problematic biases, as the effect is not driven by respondents' individual democratic support. In sum, there is no evidence of ideological biases in respondents' ratings resulting from the context of the country coded.

In addition, we also test for the effects of individual respondent's characteristics on their assessments. Figure 3 shows that respondents being government employees does not change respondents ratings significantly (standardized coefficient = 0.133, 95% CI = [-0.048, 0.321]), nor does respondents' education level (i.e. PhD or not; standardized coefficient = –0.025, 95% CI = [-0.112, 0.061]).

Respondents who reside in the country tend to code academic freedom slightly higher compared to experts who assess the country from outside (standardized coefficient = 0.122, 95% CI = [0.047, 0.197]). While statistically

https://doi.org/10.1017/S1537592724001968

Figure 4
Predicted respondent ratings by Democratic Quality and Minimum and Maximum of Respondent's Individual Support for Liberal/Electoral Democracy



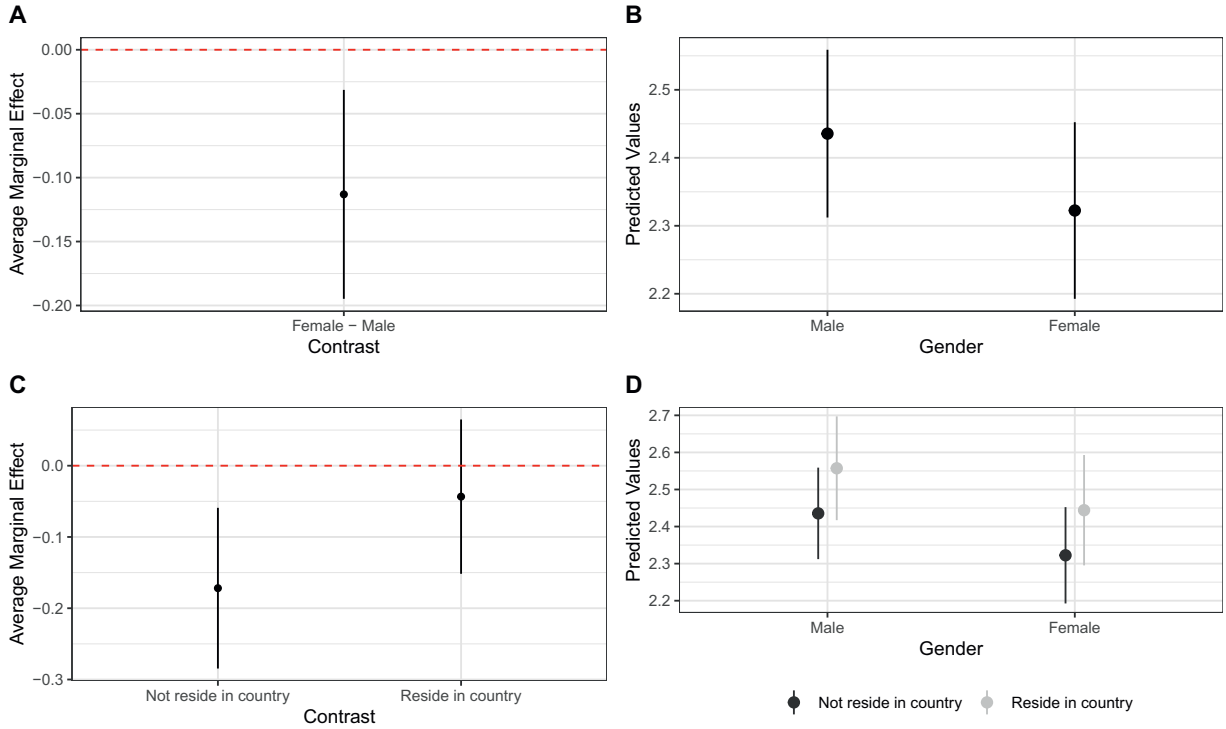
OLS regression with standard errors, clustered on countries. Measure- and year-fixed effects are included in the model.

significant at 0.05, the effect size is substantially small. Nevertheless, since the question of disagreement levels between resident and non-resident experts was discussed earlier, in particular when it comes to local experts being personally affected when academic freedom declines, we explored this question in more detail in the online appendix J. We provide a visual analysis of disagreements between inside/outside experts for three prominent cases of recent academic freedom deteriorations, namely Brazil, India, and the United States. We do find some differences between the two groups of experts, but with different constellations across the three countries. In an additional test, we systematically compare resident and non-resident respondents in cases where academic freedom declines or grows, both of which may personally affect local experts. Here we find no empirical support for the hypothesis that local and non-local experts code systematically differently when academic freedom comes under pressure. Still, we recommend to apply some caution and pay particular attention to the data's uncertainty measures when it comes to recent assessments of ongoing volatile situations—the transparently reported uncertainty interval is a major advantage of the V-Dem approach in this regard. Moreover, future rounds of data collection always allow for retrospect corrections (see Weidmann 2024), both through the recruitment of additional experts who usually code both past and present years, and by giving repeat coders the opportunity to re-evaluate their own past scores.

Respondents' gender coefficient, however, is negative and statistically significant at 0.05; female respondents rate academic freedom systematically lower compared to male respondents, all else equal (standardized coefficient = –0.113, 95% CI = [-0.195,-0.031]). To investigate this point further, we plot in figure 5 the predicted ratings for female and male respondents as well as contrasts between male and female respondents. It similarly indicates that female respondents rate academic freedom slightly lower compared to male respondents. One possible explanation for the differences could be diverging experiences women and men have in terms of their individual academic freedom. Therefore, in figure 5C and D we test the interaction effect between residing in a country and respondents' gender. If the assumption holds true, we would expect to see that women experts who reside in the country rate it systematically lower compared not only to men, but also to women who do not reside in the country. However, the figure shows that the opposite tendency is true—i.e., external female coders tend to assign the lowest scores. The average marginal effects plot in figure 5C shows that the difference is statistically insignificant, though the share of female coders who have contributed to the AFI data is also relatively low overall (less than 30%).

https://doi.org/10.1017/S1537592724001968

Figure 5
Average marginal effects (A and C) and predicted respondent ratings (B and D) by respondent's gender and respondent's reside in country



OLS regression with standard errors, clustered on countries. Measure- and year-fixed effects are included in the model.

Another possible explanation for the slight overall gender difference is that women may not only differ in their individual experience of academic freedom from their male colleagues, but that they might also have a higher awareness of systematic differences in the experiences of others. The divergences in coding could therefore point to the interlinkages between discrimination and academic freedom that are currently not explicitly captured by the AFI (refer to the Review of Conceptual Decisions and Frequent Inquiries section). While the small substantive differences between male and female experts do not cause serious concerns, they make a case for directing efforts at further diversifying the pool of expert coders.

In an additional test, we empirically evaluate the educated guess that coder quality may differ between first-time coders and multiple-time coders.[23] First-time coders may take more care to understand all the concepts in detail than those who have coded the same variables before. On the other hand, coding different V-Dem surveys for the first time for a range of years is more time-consuming than updating the assessment only for the latest year(s) for coders who have previously participated. In figure J4, tables J1 and J2, we test if the coding quality (operationalized as the deviations from the final indicator value in the V-Dem dataset) could be explained by first-time coding. However, the findings indicate that first-time and multiple-time coders do not systematically differ in this respect.

Overall, we can therefore conclude that there is no evidence for systematic biases resulting from individual respondent characteristics that would seriously affect the quality of the Academic Freedom Index.

Convergent Validity Assessment

In the next step, we analyze to what extent the academic freedom measure corresponds to alternative data sources. As mentioned before, among other expert-coded assessments, only Freedom House (FH) measures academic freedom as a separate concept. However, FH's indicator D3 on academic freedom ("Is there academic freedom, and is the educational system free from extensive political indoctrination") does not specify what academic freedom means, focuses mainly on political expression of researchers and students, and conflates higher education with primary and secondary education (Spannagel and Kinzelbach 2023). Since it is the only available cross-national time-series indicator that was not curated by the V-Dem project, we nevertheless use it in this section to conduct a traditional convergent validity assessment in the first step, to then "statistically examine the extent to which observable aspects of the data generation process predict

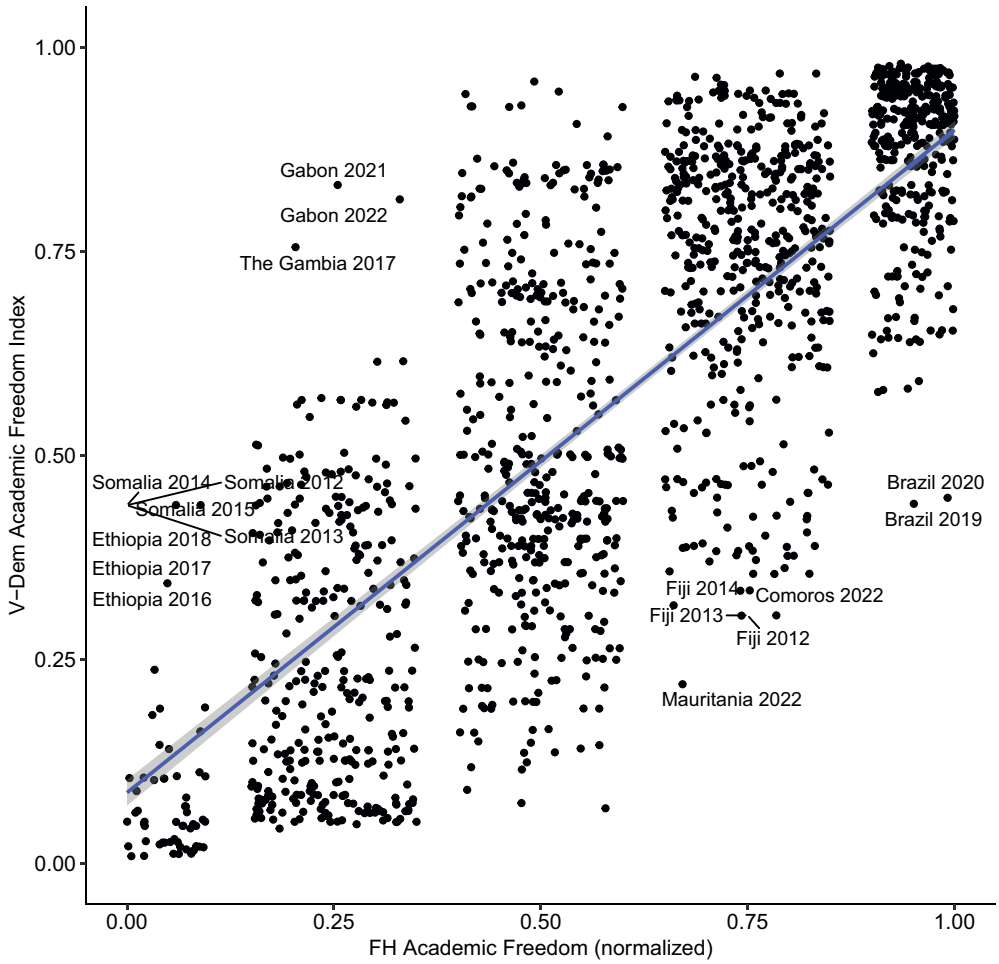
https://doi.org/10.1017/S1537592724001968

systematic divergence between the chosen measure and the alternatives" (McMann et al. 2022, 439–440).

Traditional Convergent Validity

Figure 6 shows the association between the V-Dem Academic Freedom Index and the FH academic freedom indicator. It presents the statistical association for the years between 2012 and 2022 that are available in both the V-Dem and the FH dataset. Figure 6 reveals that divergence between V-Dem and FH is relevant across all levels of academic freedom. The differences are the highest for cases of mid-level academic freedom when looking at the V-Dem Academic Freedom Index, where V-Dem disagreement is also the greatest as shown in the Analyzing Respondent Disagreement section. In addition, we can depict visual outliers, for example Ethiopia (in 2016 to 2018), Gabon in 2021 and 2022, as well as the Gambia in 2017, which all score systematically higher in the AFI than in the FH assessment. At the same time, Brazil (2019 and 2020), Fiji (2012 to 2014), and Mauritania (2022) score systematically higher in the FH assessment than in the AFI, as plotted in figure 6. Some of this divergence may stem from the fact that in their scoring process, FH uses a country's score from the previous year "as a benchmark for the current year under review," meaning that scores only tend to be changed as a result of major developments. Though they note that "gradual changes … are occasionally registered" (Freedom House 2023, 2), this makes the FH scores far less sensitive to incremental improvements or deteriorations than the V-Dem measure. On a substantive level, as noted earlier, the FH indicator conceptually encompasses not only higher education but also primary and secondary education, which could distort

Figure 6
Comparing the V-Dem Academic Freedom Index with Freedom House academic freedom measure (2012–2022)



https://doi.org/10.1017/S1537592724001968

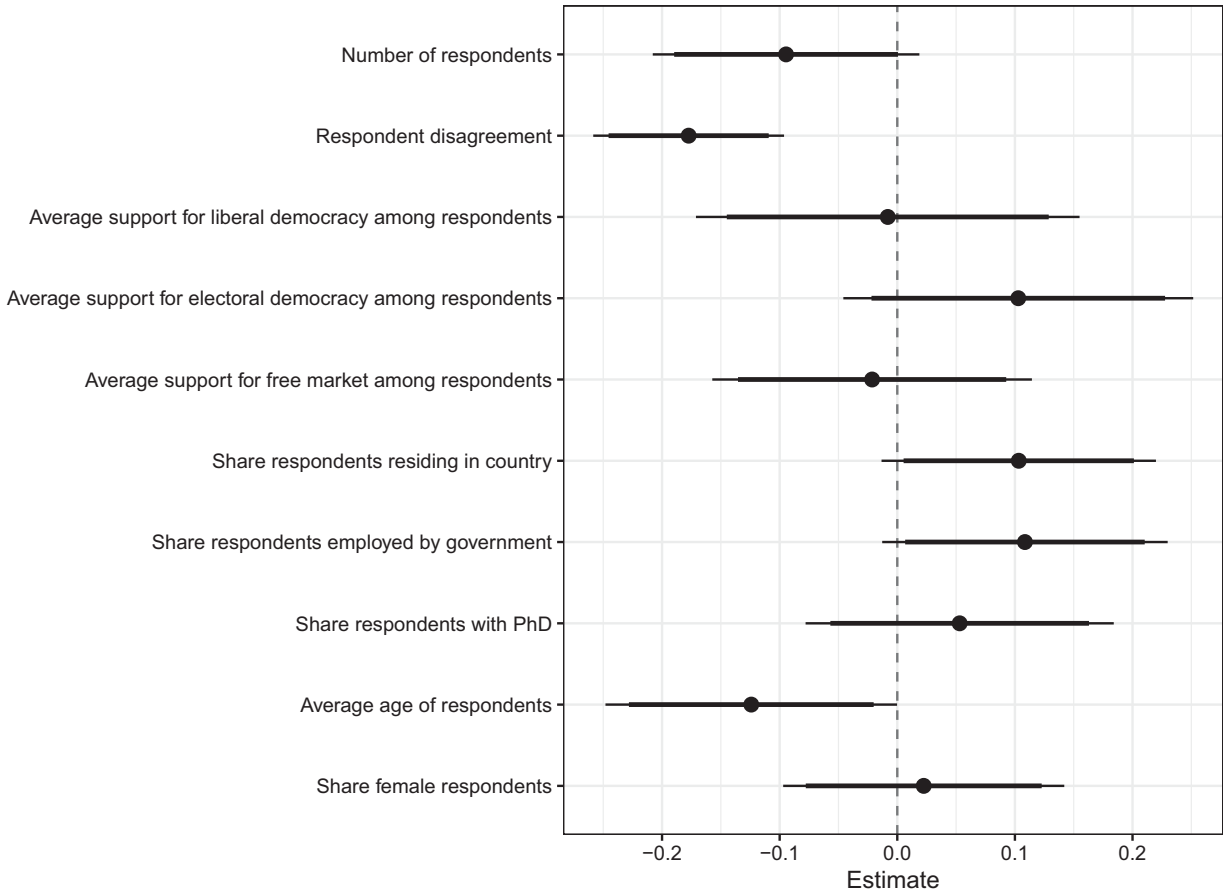
the assessment. At the same time, the correlation coefficient of 0.854 further indicates that the two measures disagree in a number of cases, but it also shows overall evidence of convergent validity. Figure F1 in the online appendix presents the statistical association for each year separately. It validates the main findings from figure 6.

Statistical Analysis of Measure Convergence

Figure 7 assesses systematic determinants of divergence between V-Dem's Academic Freedom Index and FH's academic freedom indicator. We ask here "whether the composition of V-Dem respondents per country and year, measured with average respondent characteristics, affects the tendency for V-Dem to deviate" (McMann et al. 2022, 441) from FH's indicator of academic freedom. However, we should keep in mind that divergence can also come from the fact that the FH measure is conceptually different from V-Dem's. As Hawken and Munck argue, "Consensus is not necessarily indicative of accuracy and the high correlation ….[by itself does] not establish validity" (Hawken and Munck 2009, 4). In addition, we cannot assess the raw country-year coder scores and coder characteristics from FH, as they are not publicly available and thus are not able to regress raw coder scores on each other. However, we can examine the systematic determinants of divergence between both measures. Figure 7 presents the results of the regression analysis (presented in detail in table E2 in the online appendix). The dependent variable is absolute residuals from regressing V-Dem AFI indicators as a pooled model on the FH academic freedom measure (table F1 shows the regression analysis for each indicator separately).

Figure 7 shows that the V-Dem share of female respondents as a predictor of divergence is slightly positive but statistically not significant (standardized coefficient

Figure 7
Explaining deviations from FH academic freedom indicator with aggregate respondent characteristics (pooled model)



OLS regression with standard errors, clustered on countries. The dependent variable is the absolute residuals from regressing each V-Dem measure on Freedom House's D3 indicator on academic freedom and educational system. Year-fixed effects and measure-fixed effects are included in the model but omitted from the figure.

https://doi.org/10.1017/S1537592724001968

= 0.022, 95% CI = [-0.097; 0.142]). A higher average age of V-Dem respondents (standardized coefficient = −0.124, 95% CI = [-0.248; -0.000]) significantly at the 0.05 level decreases the absolute difference between V-Dem and FH, while the share of V-Dem respondents with a PhD (standardized coefficient = 0.053, 95% CI = [-0.078; 0.184]) does not significantly affect the absolute difference between V-Dem and FH. Whether respondents support free market (standardized coefficient = −0.021, 95% CI = [-0.153; 0.114]) or liberal democracy (standardized coefficient = 0.102, 95% CI = [-0.046; 0.252]) also does not systematically increase the absolute difference between V-Dem and FH. The share of respondents employed by government coefficient is positive but only borderline statistically significant (standardized coefficient = 0.108, 95% CI = [-0.013; 0.23]). However, whether a respondent resides in a country he/she is coding (standardized coefficient = 0.103, 95% CI = [-0.014; 0.22]) is significantly at the 0.1 level associated with the deviation from FH academic freedom indicator. This indicates that the larger the share of local coders, the larger the absolute difference between V-Dem's AFI and FH, which could potentially point to deficiencies in local knowledge of coders at Freedom House.

The only other variable next to age and share of respondents residing in the country that shows a significant effect is respondent disagreement in the coding, whose coefficient is negative (standardized coefficient = −0.178, 95% CI = [-0.258; -0.096]). Therefore, larger disagreement between V-Dem's respondents is associated with smaller absolute difference between the AFI and the FH academic freedom measure. This finding may, however, be a statistical artifact and it should not be interpreted as causal. We cannot assess the connection in more detail as FH does not report expert disagreement and the expert consultation happens behind closed doors. Overall, however, the pattern is clear: there are few systematic predictors of the deviations between FH and V-Dem Academic Freedom Index among the coder characteristics.

Incorporating Measurement Uncertainty of Latent Variables

In this final section, we provide a practical guide on how to include measurement uncertainty in regression models, when working with latent variables, such as academic freedom. Even if it is not yet a standard procedure in political science, it is recommended to take measurement uncertainty of latent variables in regression models and other types of data analysis into account. However, empirical contributions that do so remain scarce and works from Fariss et al. (2022); Schnakenberg and Fariss (2014); Tai, Hu, and Solt (2024); and a few others, are still exceptions. As a general rule of thumb, taking measurement uncertainty into account when working with such latent variables helps to prevent inconclusive findings and biased standard errors, and to reduce the risk of type-I and type-II errors (Fariss et al. 2022, 580).

In this section, we summarize the approach suggested by Schnakenberg and Fariss (2014) and provide R code for applying this approach to V-Dem data. To incorporate measurement uncertainty in regression models, users can use either V-Dem's posterior files (available upon request) or the point estimates and standard deviations[24] in V-Dem's main dataset. Users first duplicate the baseline dataset m times (e.g., $m = 1{,}000$). In a second step, users assign m random draws from the posterior distribution of the latent variable to each country–year observation in the baseline dataset.[25] In a third step, users estimate m regression models and then combine these m models according to Rubin's rule (1987). It is also possible to estimate different quantities of interest, such as predictions, comparisons, and slopes.

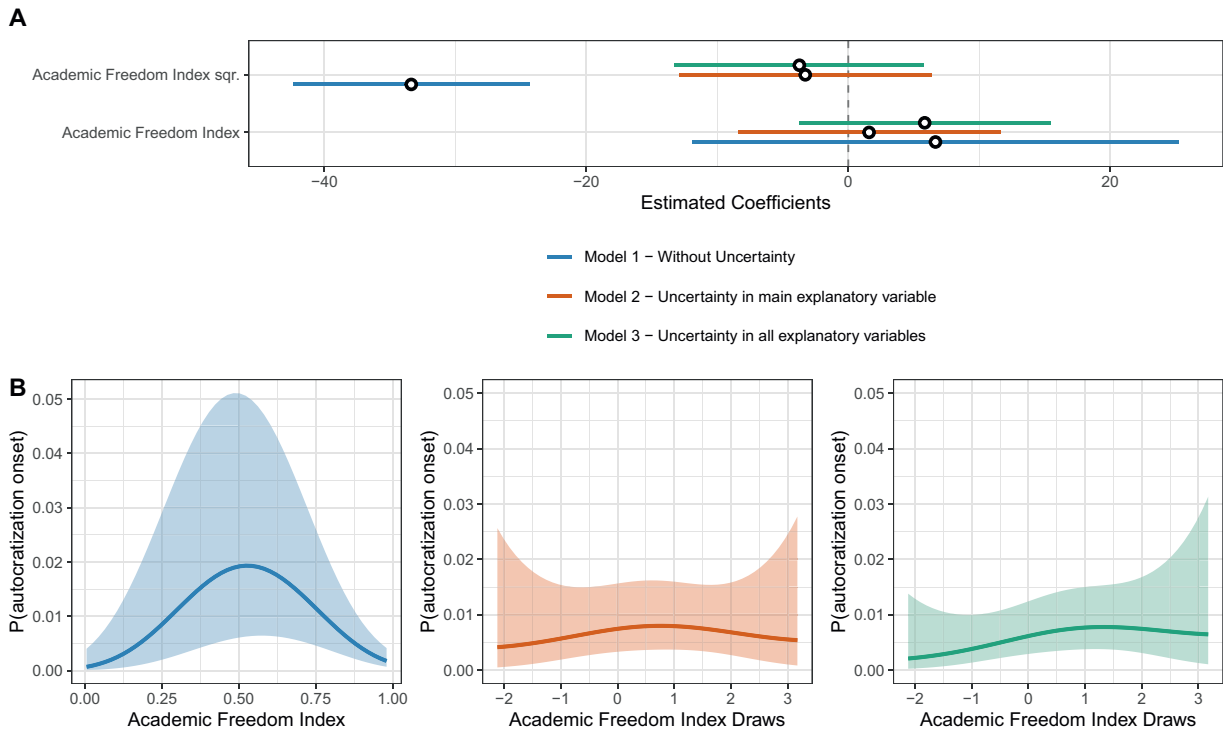
To illustrate the importance of incorporating measurement uncertainty in the regression analysis, we replicate a study on the relationship between autocratization and academic freedom (Pelke 2023). The study tries to investigate the influence of academic freedom (measured by the AFI) on the onset of an autocratization episode (measured by the Episode of Regime Transformation dataset [Maerz et al. 2024]). Using V-Dem data and binomial-response GLM models, the author shows a nonlinear relationship between academic freedom and the onset of autocratization.[26] To summarize it briefly, the original results indicate that high levels as well as low levels of academic freedom reduce probability of an onset of autocratization, while intermediate levels show the highest probability. The author argues that the inverted U-shape relationship may be counterintuitive at the first glance. However, he argues that "low academic freedom is also empirically often associated with low levels of democracy, which means that incumbents have little incentive to further autocratize in these situations" (Pelke 2023, 1022).

In the original study, the author assumes—along many others—that the explanatory variable and controls are measured without uncertainty, which is indeed not the case. We compare regression models in which uncertainty is omitted or included in different ways. In Model 1, all measurement uncertainty is omitted (as in the original study). Model 2 includes measurement uncertainty in the main independent variable, while in Model 3, measurement uncertainty is incorporated for all predictors. We take $m = 1{,}000$ random draws from the posterior distribution of latent variables, namely academic freedom, GDP per capita, GDP growth, population size, regional democracy level, and legislative and judicial constraints on the executive; then we estimate $m = 1{,}000$ regression models and combine the results according to Rubin's rule (1987).

Figure 8 illustrates that if we do not account for uncertainty in the measurement of the main explanatory

https://doi.org/10.1017/S1537592724001968

Figure 8
Comparing the effect of academic freedom on autocratization probability across models with and without including latent variable uncertainty



Note: Figure A plots the point estimates for academic freedom and academic freedom squared (lagged by one year) on the probability of autocratization. The bars represent 95% confidence intervals, which are calculated with clustered standard errors. Model 1 (blue line) regresses the point estimates for the latent academic freedom variable on the probability of autocratization. Model 2 (orange line) regresses 1,000 draws from the latent academic freedom variable on the probability of autocratization. Model 3 (green line) uses 1,000 draws from all latent explanatory variables, including academic freedom, GDP per capita, GDP growth, population size, regional democracy level, and legislative and judicial constraints on the executive. Figure B plots the predicted onset probabilities of autocratization for all three models.

variable, we risk getting biased estimates and confidence intervals. Model 1 shows a non-linear relationship with a statistically significant point estimate for academic freedom squared, while Models 2 and 3 indicate that the bias in the estimates is driven predominantly by measurement uncertainty in the (main) explanatory variables. The point estimates and confidence intervals in Models 2 and 3 are comparable. Figure 8B plots the predicted onset probabilities of autocratization for all three models. It is comparable to figure 4 in the original study. Accordingly, when taking the measurement uncertainty into account, we come to a substantively different conclusion than the author, namely that academic freedom levels are not associated in an inverted U-shape relationship with the onset of autocratization.

In online appendix K, we further illustrate the importance of incorporating uncertainty in the measurement of latent constructs by illustrating the effect of democratization on academic freedom in a simple Two-Way Fixed-Effects (TWFE) design. In this case, the point estimates of democratization are comparable, while we underestimate the confidence intervals when not taking into account measurement uncertainty. Even if doing so is computationally expensive and methodologically ambitious, it is very useful and may reduce the risk of type-I and type-II errors (Fariss et al. 2022, 580). In sum, we demonstrate that not accounting for measurement uncertainty "can lead to unaccounted for attenuation bias in regression coefficients" (Fariss et al. 2022, 582).

Conclusion

This article has explored the data quality of the Academic Freedom Index by using different tools for assessing content validity, the data generation process, and convergent validity. We used the road map introduced by McMann and coauthors (McMann et al. 2022) and are contributing to different sets of literature. First, we speak to the literature on data quality assessments (e.g., Adcock and Collier 2001; McMann et al. 2022; Sartori 1970; Seawright and Collier 2014; Zeller and Carmines 1980) by providing one of the first applications of McMann et al.'s suggested approach. Second, we advance research

https://doi.org/10.1017/S1537592724001968

on (how to measure) academic freedom (e.g., Abdel Latif 2014; Appiagyei-Atua, Beiter, and Karran 2016; Grimm and Saliba 2017; Karran, Beiter, and Appiagyei-Atua 2017; Pruvot, Estermann, and Popkhadze 2023; Spannagel 2020) by assessing the data quality of the Academic Freedom Index in detail and with rigor. We thus inform substantive research "about how strengths and limitations of a chosen measure might affect the findings of substantive research, or more specifically, the conditions under which substantive conclusions might be more or less robust" (McMann et al. 2022, 445). Third, and most importantly, our analyses will help inform future research on academic freedom that seeks to make use of this newly introduced measure in substantive analyses.

Different aspects and assumptions of the AFI affect how it should be used, and scholars and practitioners who wish to use the data for substantive research need to be aware of them. First, in the content validity assessment, we show that the conceptualization of the AFI using five different indicators that are understood as reflective of the latent construct of academic freedom, is empirically valid. At the same time, there are conceptual limits to the AFI that need to be taken into account, such as its focus on academic freedom as a negative, not a positive freedom, and its currently only indirect inclusion of the student perspective. In addition, we discuss critically how the different indicators can be aggregated to customized measures and show how to include measurement noise into the aggregation of different indicators with Bayesian factor analysis. Among the alternatives tested, the AFI seems to deliver the best results. However, if researchers disagree with the theoretical assumption that the five indicators are symptoms of the latent concept of academic freedom, they are able to aggregate the available indicators in alternative ways. For instance, there might be reasons to consider specific indicators as determinants of academic freedom, as necessary conditions, or mutually substitutable indicators. That said, prior to using alternative aggregation methods in substantive research, a critical assessment of the chosen measure's strengths and limitations is recommended.

Second, the AFI's data generation process as part of the V-Dem project, a serious and renowned academic endeavor, inspires confidence in its general data quality. Moreover, the V-Dem experts represent diverse backgrounds in terms of their geographic location and expertise, and as academics they seem particularly qualified given their likely intimate knowledge of the country's higher education system. Overall, these findings suggest that the academic freedom data can be applied across contexts and are valid for countries around the world.

Third, the findings from the inter-respondent disagreement analysis as well as the correlates of respondent ratings tell us that overall the AFI data does not exhibit systematic biases that stem from country or respondent characteristics, as far as we can tell from the data available. In particular, the analysis of the inter-respondent disagreements indicates that country contexts with more freedom of expression show less respondent disagreement. In addition, respondent disagreement has a nonlinear relationship with the level of academic freedom. Specifically, respondent disagreement is highest at mid-levels of academic freedom, lesser at low levels, and lowest at high levels of academic freedom. Respondent disagreement, alongside other measurement issues, may lead to more uncertainty in the AFI measure, so we strongly advise users of the data to consider the statistical uncertainty of the predicted scores, as illustrated in this article.

Supplementary material

To view supplementary material for this article, please visit http://doi.org/10.1017/S1537592724001968.

Data Availability Statement

Research documentation and data that support the findings of this study are openly available at the *Perspectives on Politics* Dataverse: https://doi.org/10.7910/DVN/US8MUW. The scripts and data are also available on Code Ocean: https://codeocean.com/capsule/5176300/tree/v1. The reproduction materials contain all data that is necessary to computationally reproduce the results presented in this article and the supplementary appendix, except of specific variables of V-Dem's post-survey questionnaire (PSQ). This PSQ data (including the gender, age, country of residence, government employment, and education level) contains potentially identifiable personal information and is therefore subject to legal restrictions preventing us from making it public. Every person who would like to reproduce the original results using the PSQ data (table 2, figures 1, 2, 3, 4, 5, and 7) can submit a request to the V-Dem Institute for access to the PSQ data. The reproduction materials include scripts using simulated PSQ data to show the computationally reproducibility of our results.

Acknowledgments

The authors are especially grateful to Johannes von Römer, Oskar Ryden, Linnea Fox, and Daniel Pemstein for their help in managing the data analysis and providing code and the raw data. In addition, they thank Katrin Kinzelbach, Ann-Marie Clark, participants at ISA 2024, as well as the reviewers and editors of *Perspective on Politics* for their valuable comments on an earlier version of this article.

The author(s) declared no potential conflicts of interest with respect to the research, authorship, or publication of this article.

This research was funded by the Volkswagen Foundation [grant number A138109], PI: Katrin Kinzelbach and Staffan I. Lindberg. Janika Spannagel's work on this paper was funded by the Deutsche Forschungsgemeinschaft (DFG, German research Foundation) under Germany's

Excellence Strategy [grant EXC 2055]. The V–Dem measurement process was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden and the Swedish National Infrastructure for Computing at the National Supercomputer Center in Sweden, partially funded by the Swedish Research Council through grant agreements No. 2022-06725 and No. 2018-05973.

Notes

1 This article is based on Version 13 of the V-Dem dataset (Coppedge et al. 2023a).
2 These include in particular V-Dem's exclusion indicators, some of which assess the access to public services or education inequality among different groups.
3 The Bayesian factor analysis model was fitted here to each draw of the V-Dem measurement model (i.e., one draw from the posterior of each manifest variable, covering every country-year) using Markov Chain Monte Carlo (MCMC) methods. In order to capture posterior uncertainty, we run the Bayesian factor model 200 (ITER) times with different posterior draws from the variables and 10,000 sampling iterations. We divide these runs into four groups, each with the same initial values, and for convergence purposes we treat each group as a separate chain to allow for a Gelman & Rubin diagnostic.
4 The factor loadings of the AFI are in fact stronger than the factor loadings reported for the electoral and for the liberal dimensions of democracy in the V-Dem dataset (Coppedge et al. 2020, ch 5).
5 Too much congruence of the factor loadings and too low uniqueness would indicate that the different indicators all capture the same underlying questions and that expert coders do not differentiate adequately between these indicators.
6 The lesser fit may also partly be explained by the technical fact that this indicator is included in a different V-Dem survey than the others and thus coded by a partly different set of experts; see also online appendix D.
7 See also table D5 in the online appendix for a list of total pairwise coders and unique coders across indicators.
8 Coder characteristics are only available for coders that participated in V-Dem's Post-Survey Questionnaires (PSQ) (n=2,008); those who participated have different patterns of missingness in V-Dem's PSQ. The missingness is likely not random, so the known distributions can only give an approximate idea.
9 A nonbinary option is not provided by V-Dem.
10 From the 67 such government employees, 25 were not living in the main country they were coding, while 41 were. We define government employees here as coders who indicated to belong to one of the following entities (in V-Dem's v2zzemploy indicator of the post-survey questionnaire): 1: The current executive (presidential administration/cabinet). 2: A ministry, board, or agency within the central government. 3: A ministry, board, or agency within the local/regional government.
11 This is comparable to the percentage of local experts across all V-Dem data.
12 Unfortunately, the experts' field of study is not collected in V-Dem's PSQ, so we cannot provide systematic statistics.
13 The individual indicators rely on average on the assessment of 6.09 to 6.51 experts per country-year.
14 The authors of the AFI's introductory paper discuss this possibility using the example of Brazil, whose scores seem to have deteriorated disproportionately under Jair Bolsonaro's presidency, compared both to other countries during the same period and to Brazil's own historic records (Spannagel and Kinzelbach 2023, 15).
15 Freedom House covers 202 countries/territories, among which are a number of microstates and (semi-) autonomous territories that V-Dem does not cover.
16 The *Academic Freedom in Constitutions* dataset (Spannagel 2023) goes back to 1789, but it only documents the *de jure* presence of academic freedom provisions in constitutions, not their realization.
17 Measurement-model adjusted ratings are transformations of parameters from the IRT model and can be seen as rater "perceptions" of a latent score after adjusting for DIF by using the posterior simulations.
18 For easier interpretation of the results, all regression coefficients were standardized by two standard deviations in figures 1 and 3.
19 Refer to online appendix I for an overview of the rater's confidence across indicators.
20 We thank Katrin Kinzelbach for drawing our attention to this point.
21 This post survey item, named *v2zztimespent*, measures the focused work time an expert spent to complete the coding, including the preparation and the time spent in data entry tool. It is based on the self-declaration of the expert.
22 Standardized coefficient for EDI * Support for EDI = −0.031, 95% CI = [-0.311, 0.248], and standardized coefficient for liberal component * Support for liberal principle = 0.244, 95% CI = [-0.04, 0.529]).
23 We thank the anonymous reviewer who drew our attention to this point.
24 Identified as "_sd" at the end of a variable name.
25 A code example is provided in the online appendix and the replication materials.
26 In this replication study, we use the original AFI rather than the proposed index constructed by Pelke and estimate binomial-response GLM models without

https://doi.org/10.1017/S1537592724001968

Firth's method of bias reduction to reduce computation time. Therefore, the original results and the results presented in Model 1 slightly differ.

References

Abdel Latif, Muhammad. M. 2014. "Academic Freedom: Problems in Conceptualization and Research." *Higher Education Research & Development* 33(2): 399–401. https://doi.org/10.1080/07294360.2014.881766

Adcock, Robert, and David Collier 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3): 529–546. https://doi.org/10.1017/S0003055401003100

Altbach, Philip G. 2016. "Research Universities in Developing Countries." In *Global Perspectives on Higher Education*, ed. Philip G. Altbach, 172–198. Baltimore, MD: Johns Hopkins University Press.

Appiagyei-Atua, Kwadwo, Klaus Beiter, and Terence Karran. 2016. "A Review of Academic Freedom in Africa through the Prism of the UNESCO's 1997 Recommendation." *Journal of Higher Education in Africa/Revue de l'enseignement supérieur en Afrique* 14(1): 85–117. https://doi.org/10.57054/jhea.v14i1.1508

Berggren, Niclas, and Christian Bjørnskov. 2022. "Political Institutions and Academic Freedom: Evidence from across the World." *Public Choice* 190(1): 205–228. https://doi.org/10.1007/s11127-021-00931-9

Bollen, Kenneth A., and Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1): 58–86. https://doi.org/10.1177/0010414000033001003

Brook, Anne-Marie, K. Chad Clay, and Susan Randolph. 2019. *Human Rights Measurement Initiative Methodology Handbook* (https://humanrightsmeasurement.org/wp-content/uploads/2019/06/HRMI-Methodology-Guide-2019-version-2019.06.06.pdf).

Butler, Petra, and Roderick Mulgan. 2013. "Can Academic Freedom Survive Performance Based Research Funding." *Victoria University of Wellington Law Review* 44:487–519.

Church, A. Timothy. 2010. "Measurement Issues in Cross-Cultural Research." In *The Sage Handbook of Measurement*, ed. Geoffrey Walford, Eric Tucker, and Madhu Viswanathan, 151–177. Thousand Oaks, CA: Sage.

Coppedge, Michael, John Gerring, Adam Glynn, Carl Henrik Knutsen, Staffan I. Lindberg, Daniel Pemstein, Brigitte Seim, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, Fernando Bizzarro, Joshua Krusell, Matthew Maguire, Kyle Marquardt, Kelly McCann, Valeriya Mechkova, Farhad Miri, Josefine Pernes, Jeffrey Staton, Natalia Stepanova, Eitan Tzelgov, Yi-ting Wang. 2020. *Varieties of Democracy: Measuring Two Centuries of Political Change*. Cambridge: Cambridge University Press.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Fabio Angiolillo, Michael Bernhard, M., Cecilia Borella, Agnes Cornell, M. Steven Fish, Linnea Fox, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Ana Good God, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Anja Neundorf, Pamela Paxton, Daniel Pemstein, Oskar Rydén, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2024. *V-Dem Country-Year Dataset v14.* https://doi.org/10.23696/vdemds24

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, Ana Good God, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Joshua Krusell, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Natalia Natsika, Anja Neundorf, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Oskar Rydén, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2023a. "V-Dem [Country-Year/Country-Date] Dataset v13." *Varieties of Democracy (V-Dem) Project.* https://doi.org/10.23696/vdemds23

——. 2023b. "V-Dem Codebook v13". *Varieties of Democracy (V-Dem) Project.*

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Kyle L. Marquardt, Juraj Medzihorsky, Daniel Pemstein, Lisa Gastaldi, Sandra Grahn, Josefine Pernes, Oskar Rydén, Johannes von Römer, Eitan Tzelgov, Yi-ting Wang, and Steven Wilson. 2023. "V-Dem Methodology v13." *Varieties of Democracy (V-Dem) Project.*

Croissant, Aurel, and Lars Pelke. 2022. "Measuring Policy Performance, Democracy, and Governance Capacities: A Conceptual and Methodological Assessment of the Sustainable Governance Indicators (SGI)." *European Policy Analysis* 8(2): 136–59. https://doi.org/10.1002/epa2.1141

Enyedi, Zsolt. 2018. "Democratic Backsliding and Academic Freedom in Hungary." *Perspectives on Politics* 16(4): 1067–74. https://doi.org/10.1017/S1537592718002165

Fariss, Christopher J., Therese Anders, Jonathan N. Markowitz, and Miriam Barnum. 2022. "New Estimates of Over 500 Years of Historic GDP

and Population Data." *Journal of Conflict Resolution* 66(3): 553–91. https://doi.org/10.1177/00220027211054432

Freedom House. 2022. *Freedom in the World 2022 Methodology.* (https://freedomhouse.org/sites/default/files/2022-02/FIW_2022_Methodology_For_Web.pdf).

——. 2023. *Freedom in the World. Methodology Questions.* Bethesda. (https://freedomhouse.org/sites/default/files/2023-03/FITW_2023%20MethodologyPDF.pdf).

Grimm, Jannis, and Ilyas Saliba. 2017. "Free Research in Fearful Times: Conceptualizing an Index to Monitor Academic Freedom." *Interdisciplinary Political Studies* 3(1): 41–75. https://doi.org/10.1285/i20398573v3n1p41

Hawken, Angela, and Gerardo L. Munck. 2009. "*Do You Know Your Data? Measurement Validity in Corruption Research.*" Unpublished typescript, Pepperdine University and University of Southern California, Malibu, CA.

Kaczmarska, Katarzyna. 2020. "Academic Freedom in Russia." In *Researching Academic Freedom: Guidelines and Sample Case Studies*, ed. Katrin Kinzelbach, 103–140. Erlangen: FAU University Press.

Karran, Terence, Klaus Beiter, and Kwadwo Appiagyei-Atua. 2017. "Measuring Academic Freedom in Europe: A Criterion Referenced Approach." *Policy Reviews in Higher Education* 1(2): 209–39. https://doi.org/10.1080/23322969.2017.1307093

Kinzelbach, Katrin, Staffan I. Lindberg, and Lars Lott. 2024. "Academic Freedom Index–2024 Update." *FAU Erlangen-Nürnberg and V-Dem Institute.* https://doi.org/10.25593/open-fau-405

Kinzelbach, Katrin, Staffan I. Lindberg, Lars Pelke, and Janika Spannagel. 2023. " Academic Freedom Index–2023 Update." *FAU Erlangen-Nürnberg and V-Dem Institute.* https://doi.org/10.25593/opus4-fau-21630

Knutsen, Carl Henrik, Kyle L. Marquardt, Brigitte Seim, Michael Coppedge, Amanda B. Edgell, Juri 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. https://doi.org/10.1017/S104909652300077X

Kratou, Hajer, and Liisa Laakso. 2022. "The Impact of Academic Freedom on Democracy in Africa." *Journal of Development Studies* 58(4): 809–26. https://doi.org/10.1080/00220388.2021.1988080

Lerch, Julia C., David J. Frank, and Evan Schofer. 2024. "The Social Foundations of Academic Freedom: Heterogeneous Institutions in World Society, 1960 to 2022." *American Sociological Review* 89(1): 88–125. https://doi.org/10.1177/00031224231214000

Little, Andrew T., and Anne Meng. 2024a. "Measuring Democratic Backsliding." *PS: Political Science & Politics* 57(2): 149–61. https://doi.org/10.1017/S104909652300063X

——. 2024b. "What We Do and Do Not Know about Democratic Backsliding." *PS: Political Science & Politics* 57(2): 224–29. https://doi.org/10.1017/S1049096523001038

Lott, Lars. 2024. "Academic Freedom Growth and Decline Episodes." *Higher Education* 88(3): 999–1017. https://doi.org/10.1007/s10734-023-01156-z

Lott, Lars, and Janika Spannagel. 2024. "Replication Data for: Quality Assessment of the Academic Freedom Index: Strengths, Weaknesses, and How Best to Use It." https://doi.org/10.7910/DVN/US8MUW

Macfarlane, Bruce. 2012. "Re-framing Student Academic Freedom: A Capability Perspective." *Higher Education* 63(6): 719–32. https://doi.org/10.1007/s10734-011-9473-4

Maerz, Seraphine F., Amanda B. Edgell, Matthew C. Wilson, Sebastian Hellmeier, and Staffan I. Lindberg. 2024. "Episodes of Regime Transformation." *Journal of Peace Research* 61(6): 967–84. https://doi.org/10.1177/00223433231168192

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

——. 2023. "Estimating Latent Traits from Expert Surveys: An Analysis of Sensitivity to Data-generating Process." *Political Science Research and Methods* 11(2): 384–93. https://doi.org/10.1017/psrm.2021.39

Matei, Liviu, and Julia Iwinska. 2018. "Diverging Paths? Institutional Autonomy and Academic Freedom in the European Higher Education Area." In *European Higher Education Area: The Impact of Past and Future policies*, ed. Adrian Curaj and Ligia Deca, 345–368. Cham: Springer.

McMann, Kelly, Daniel Pemstein, Brigitte Seim, Jan Teorell,Staffan I. Lindberg. 2022. "Assessing Data Quality: An Approach and an Application." *Political Analysis* 30(3): 426–449. https://doi.org/10.1017/pan.2021.27

Mendes, Conrado H. 2020. "Academic Freedom in Brazil." In *Researching Academic Freedom: Guidelines and Sample Case Studies*, ed. Katrin Kinzelbach, 63–102. Erlangen: FAU University Press.

Nokkala, Terhi, and Agneta Bladh. 2014. "Institutional Autonomy and Academic Freedom in the Nordic Context—Similarities and Differences." *Higher Education Policy* 27(1): 1–21. https://doi.org/10.1057/hep.2013.8

Pelke, Lars. 2023. "Academic Freedom and the Onset of Autocratization." *Democratization* 30(6): 1015–39. https://doi.org/10.1080/13510347.2023.2207213

Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Juraj Medzihorsky, Joshua Krusell, Farhad Miri, and Johannes von Römer. 2023. "The V-Dem Measurement Model: Latent Variable Analysis for

Cross-National and Cross-Temporal Expert-Coded Data." V-Dem Working Paper No. 21. 8th ed.
Pruvot, Enora B., Thomas Estermann, and Nino Popkhadze. 2023. "University Autonomy in Europe IV, The Scorecard 2023." (https://eua.eu/downloads/publications/eua%20autonomy%20scorecard.pdf).
Puaca, Goran. 2022. "Institutional Autonomy, Managerialism and the Conditions for Academic Freedom in Swedish Higher Education." In *Handbook on Academic Freedom*, ed. Richard Watermeyer, Rille Raaper, and Mark Olssen, 106–125. Northampton, MA: Edward Elgar Publishing.
Roberts Lyer, Kirsten, Ilyas Saliba, and Janika Spannagel. 2023. *University Autonomy Decline: Causes, Responses, and Implications for Academic Freedom*. Abingdon: Taylor & Francis.
Rubin, Donald B. 1987. *Multiple Imputation for Survey Nonresponse.* Unpublished manuscript.
Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4): 1033–53.
Sawyerr, Akilagpa. 2004. "African Universities and the Challenge of Research Capacity Development." *Journal of Higher Education in Africa/ Revue de l'enseignement supérieur en Afrique* 2(1): 213–42.
Schnakenberg, Krith E., and Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1): 1–31. https://doi.org/10.1017/psrm.2013.15
Seawright, Jason, and David Collier. 2014. "Rival Strategies of Validation: Tools for Evaluating Measures of Democracy." *Comparative Political Studies* 47(1): 111–38. https://doi.org/10.1177/0010414013489098
Spannagel, Janika. 2020. "The Perks and Hazards of Data Sources on Academic Freedom: An Inventory." In *Researching Academic Freedom: Guidelines and Sample Case Studies*, ed. Katrin Kinzelbach, 175–221. Erlangen: FAU University Press.
——. 2023. *Academic Freedom in Constitutions (AFC).* https://doi.org/10.7910/DVN/E8MIMF
Spannagel, Janika, and Katrin Kinzelbach. 2023. "The Academic Freedom Index and Other New Indicators Relating to Academic Space." *Quantity and Quality* 57*:* 3969–89.
Tai, Yuehong, Yue Hu,and Frederick Solt. 2024. "Democracy, Public Support, and Measurement Uncertainty." *American Political Science Review* 118(1): 512–18. https://doi.org/10.1017/S0003055422000429
Taylor, Barrett J., Kelsey Kunkle, and Kimberly Watts. 2023. "Democratic Backsliding and the Balance Wheel Hypothesis: Partisanship and State Funding for Higher Education in the United States." *Higher Education Policy* 36*:*781–803. https://doi.org/10.1057/s41307-022-00286-w
Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201–17. https://doi.org/10.1111/j.1540-5907.2007.00308.x
Treisman, Daniel. 2024. "Psychological Biases and Democratic Anxiety: A Comment on Little and Meng (2023)." *PS: Political Science & Politics* 57(2): 194–97. https://doi.org/10.1017/S1049096523000768
Varieties of Democracy Project. 2022. *Varieties of Democracy Global Team.* Gothenburg. (https://www.v-dem.net/about/v-dem-project/global-team/).
Weidmann, Nils B. 2024. "Recent Events and the Coding of Cross-National Indicators." *Comparative Political Studies* 57(6): 921–37. https://doi.org/10.1177/00104140231193006
Weitzel, Daniel, John Gerring, Daniel Pemstein, and Svend-Erik Skaaning. 2023. "Measuring Electoral Democracy with Observables." https://doi.org/10.31235/osf.io/9kzja
Zavale, Nelson C. 2022. "Academic Freedom in Mozambique." In *University Autonomy Decline*, ed. Kirsten Roberts Lyer, Ilyas Saliba, and Janika Spannagel, 92–118. New York: Routledge.
Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge: Cambridge University Press.