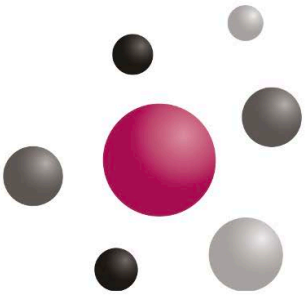


V-Dem INSTITUTE

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

Lars Pelke and Janika Spannagel

June 2023

Working Paper

SERIES 2023:142

THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (V-Dem) is a unique approach to conceptualization and measurement of democracy. The headquarters – the V-Dem Institute – is based at the University of Gothenburg with 20 staff. The project includes a worldwide team with 5 Principal Investigators, 22 Project Managers, 33 Regional Managers, 134 Country Coordinators, Research Assistants, and almost 4,000 Country Experts. The V-Dem project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries to:

V-Dem Institute
Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, Box 711
405 30 Gothenburg
Sweden
E-mail: contact@v-dem.net

V-Dem Working Papers are available in electronic format at https://www.v-dem.net

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

Lars Pelke[1, 2] and Janika Spannagel[3]

[1]Friedrich-Alexander-University Erlangen-Nürnberg
[2]Research Associate, V-Dem Institute, University of Gothenburg
[3]Freie Universität Berlin, Cluster of Excellence "Contestations of the Liberal Scripts (SCRIPTS)."

June 2023

*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. The authors are especially grateful to Johannes von Römer, Oskar Ryden, and Daniel Pemstein for their help in managing the data analysis and providing code and the raw data. In addition, we thank Katrin Kinzelbach for her valuable comments on an earlier version of this article.

†**Corresponding author**: Lars Pelke lars.pelke@fau.de

Abstract

This paper reviews the data quality of the first systematic global measurement of academic freedom, namely 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), this article examines 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.

Keywords: Academic Freedom Index, validity, reliability, academic freedom, Bayesian IRT

1 Introduction

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, p. 1), 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, covering a time span since 1900. Now that the dataset is released in its fourth iteration (Coppedge et al., 2023a) 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.

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, 2022). A further article discussed alternative measurements and data sources on academic freedom (Spannagel, 2020). The present paper expands on these contributions and offers a more in-depth evaluation to thoroughly assess the validity and reliability of the AFI as an instrument to measure academic freedom around the world. 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 paper focus on (1) content validity, (2) the data generation process, and (3) convergent validity.

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, 2022). The present paper 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 (section 2).

With regard to the data generation process, we discuss the validity and reliability of how the AFI data are collected and aggregated. Namely, the data rely on assessments of close to 2,200 country

experts that are aggregated using a customized Bayesian Item Response Theory model (Pemstein et al., 2023) controlling for respondents' individual coding behavior. 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). We will therefore focus on discussing the elements that are specific to the AFI. Special attention will be given to the investigation of coder disagreements and biases in the AFI data as an analytical tool to assess the validity and reliability of a data generation process that relies on expert assessments (section 3).

The convergent validity assessment serves to compare the AFI measure to the only 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* (section 4). Finally, we summarize the findings to draw conclusions on the AFI's strengths, weaknesses, and how its distinct characteristics should shape its application (section 5).

2 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), i.e. 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*, 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 reflect key aspects of academic freedom as defined in international law (Spannagel and Kinzelbach, 2022, 5f.).

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 together with general definitions, a specific clarification text and five defined response levels on an ordinal scale from 0 to 4[1] is presented to country experts (see all details in codebook at Coppedge et al., 2023b, pp. 233–237).

[1] E.g., ranging from *(0): Completely restricted. When determining their research agenda or teaching curricula, scholars are, across all disciplines, consistently subject to interference or incentivized to self-censor* to *(4): Fully free. When determining their research agenda or teaching curricula, scholars are not subject to interference or incentivized to self-censor.*

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 Bayesian measurement model (see more details below).

2.1 Review of conceptual decisions and frequent inquiries

Although the five indicators provide a thorough conceptual framework for the index, the authors concede that the inclusion of additional aspects would have been thinkable (Spannagel and Kinzelbach, 2022, p. 6). 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 frequently asked questions and criticisms that are 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 with regards to the institutional autonomy of higher education institutions as well as campus integrity, which, unlike tenure, are included in the AFI.[2] Compared to these two institutional aspects, tenure would be a far more specific indicator that can limit global comparability given the diversity of higher education sectors and the varying role that provisions such as tenure play across these different contexts. Autonomy and campus integrity, on the other hand, appear universally relevant to the protection of academic freedom. They can be considered as integral parts of a multidimensional understanding of academic freedom that includes both individual and institutional aspects. That said, users of the AFI that wish to focus on academic freedom more narrowly can decide to exclude the institutional indicators when working with the data (see further exploration of this point in section 2.2).

[2]In the definition of institutional autonomy, the AFI indicator follows the Lima Declaration of 1988 in asserting that it "means the independence of institutions of higher education from the State and all other forces of society, to make decisions regarding its internal government, finance, administration, and to establish its policies of education, research, extension work and other related activities." It is thus largely congruent (though less detailed) with other major assessment efforts of university autonomy, such as the European University Association's Autonomy Scorecard (Pruvot, Estermann, and Popkhadze, 2023), which assesses organizational, financial, staffing and academic autonomy.

A second potentially omitted aspect implied by the authors is the diversity of, non-discrimination in, and equal access to higher education. The omission of such aspects is justified in the introductory paper 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, 2022).[3] 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.

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; Zavale, 2022; Sawyerr, 2004), there is arguably a difference between the suppression, redistribution or conditionality of funding motivated by political or economic interests – as captured by indicators on interference with research agendas and on institutional autonomy – and the mere absence of resources for academic research. In this discussion, one should also 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. Taking both these 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 understanding (an active promotion of academic freedom by the state and other actors).

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. Macfarlane, 2012; Abdel Latif, 2014). 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

[3]These include in particular V-Dem's exclusion indicators, some of which assess the access to public services distributed by socio-economic position, gender, urban-rural location, political group, or social group, respectively; another indicator specifically assesses education inequality (albeit only with respect to basic education); further, there are indicators that indirectly capture discrimination, such as the freedom of religion.

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 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. Puaca, 2022; Butler and Mulgan, 2013). 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, p. 233).

In practical terms, however, the ability of the AFI to capture these issues may be limited. On the 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 all in adequate detail.

Finally, the lack of disaggregation between disciplines and higher education institutions is also a frequently 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 more 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. The inevitable aggregation of the academic freedom indicators that is imposed by a country-level measurement may introduce a certain fuzziness into the concept. However, Spannagel and Kinzelbach argue substantively that they

> "are aiming to assess the integrity of the academic community as a whole, and consider it to be dangerous to excuse or relativize the infringements on some subjects by the freedom of others – precisely because the targeting of a few sensitive subject areas is a

known pattern of repression, and often spreads a culture of fear throughout the academic community. What is more, infringements on institutional autonomy affect all academics, regardless of their discipline." (Spannagel and Kinzelbach, 2022, p. 9)

Yet the authors also acknowledge that "the quality of restrictions on the academic sector as a whole is different depending on whether only some or all disciplines are targeted, which is why we didn't choose to focus on only the worst-off subject areas." Instead, the scope of infringements across academic disciplines is incorporated in the response scale of the first two indicators (see more details Spannagel and Kinzelbach, 2022). In terms of the imposed 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 address the level and potential sources of coder disagreement further below (section 3.6).

On the whole, academic freedom remains a latent construct that cannot be measured directly by statistical indicators or objective measures. The five AFI indicators present overall 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.

2.2 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 (the five different indicators) and to incorporate measurement uncertainty into the proposed measure. The AFI ranges from 0 (no academic freedom) to 1 (full academic freedom) and consists of point estimates from the BFA accompanied with uncertainty measures. We discuss the choices in terms of index-level aggregation in section 3.4, but first want to show how the five different dimensions reflect one underlying systematized concept, namely academic freedom.

For this analysis, we use BFA,[4] 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, p. 433). With the BFA, we assess how the five V-Dem academic freedom measures, namely *freedom to research and teach*, *freedom of academic exchange and dissemination*, *institutional*

[4]See McMann et al. (2022) for a detailed explanation and comparison to frequentist factor analysis.

autonomy, *campus integrity*, and *freedom of academic and cultural expression*, 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, p. 433).

Table 1 indicates that all five indicators strongly load on a single dimension.[5] Factor loadings in Table 1 indicate how much an indicator of academic freedom explains of the Academic Freedom Index, while the uniqueness is the variance that is not shared with the other indicators (i.e. that is unique to the 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 (uniqueness). *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 large share of variance is unaccounted for. This is not surprising as it deviates somewhat from a strict academic freedom perspective by focusing on academics' (and artists') freedom of expression. Some users of the data wishing to focus on academic freedom more narrowly may therefore choose to exclude this indicator.[6]

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

[5]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 so that we can run a Gelman & Rubin diagnostic. Table 1 is based on V-Dem version 13.

[6]The lesser fit may also partly be explained by the technical fact that this indicator comes from the *Civil Liberty* V-Dem survey, while the other four dimensions come from the *Civic and Academic Space* V-Dem survey. The two surveys are coded by a partly different set of experts and therefore are also likely to have different empirical distributions of raters' errors. This issue is further explored in Appendix D.

In addition, we test whether a two-factor model explains more variance in the manifest variables than the unidimensional factor model (used by the AFI) by using frequentist factor analysis presented in Tables B1 and B2 (Supplementary 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 dimensions, as the latter represent individual freedoms and the former represent institutional rights of universities that protect them from outside interference. In addition, as noted above, 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. Further below (in 3.4), 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.

3 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, p. 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. The authors recommend to examine an array of elements in the data generation process, including the dataset management structure, data sources, respondent coding procedures, aggregation models, case coverage, as well as (where multiple respondents are used) inter-coder disagreement and intra-coder biases. 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. evaluated in

their paper. We will therefore quickly go over the shared characteristics that were already discussed in their paper, and focus on the elements that are specific to the AFI indicators.

3.1 Data Management Structure and Data Sources

In terms of data management structure, McMann et al. (2022, 434f.) 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 (more on this see below). Events-based data, another frequently used data source on academic freedom that records incidents of academic freedom violations, also suffers from a whole range of selection biases and are generally unsuitable to assess less repressive (and even the most repressive) contexts. Other potential sources include data collecting 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, p. 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." V-Dem and the AFI avoid this problem – as does Freedom House and its D3 indicator, the AFI's sole comparable alternative – by using one expert data collection effort to generate indicators on academic freedom (see also Coppedge et al., 2020). Experts may still draw upon the above-mentioned data sources (where available), but they can contextualize and correct the information using their accumulated expertise.

That said, McMann et al. also caution against the fact that V-Dem country experts often respond to various questions relating 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, p. 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, p. 435), and also implies

that researchers should avoid putting indicators relating to the same concept on both sides of a regression equation. However, in Supplementary Appendix D, we analyze how raters' errors correlate across indicators. The findings reveal that while raw errors in rater scores correlate highly across expert ratings, this appears to stem largely from differential item functioning (DIF). Table D2 in Supplementary Appendix shows only minor evidence of raters' error correlations across indicators after correcting responses for DIF (see also Section D in Supplementary Appendix for detailed discussion of DIF).

The Freedom House measure 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 v13, 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] There were so far 2,197 experts in total who have contributed to the AFI indicators.

3.2 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.

In a first step, we provide some basic descriptive statistics for the pool of country experts coding the indicators in the AFI. Table 2 shows descriptive statistics for our expert sample of 2,197 distinct coders.[8] It indicates that at least 53 percent of the expert are men and at least 21.2 percent are women.[9] In terms of level of education, a majority of at least 54.1 percent of the coders have a PhD

[7] See also Table D5 in the Supplementary Appendix for a list of total pairwise coders and unique coders across indicators.

[8] We analyzed coder characteristics for coders that participated (2,003) in the V-Dem's Post-Survey Questionnaires (PSQ), but different coders have different patterns of missingness in V-Dem's PSQ. Note that the missingness is likely not random, so that the known distributions can only give an approximate idea.

[9] A nonbinary option is not given in V-Dem's post-survey questionnaire. However, the remaining 25.8 percent are missing observations.

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	26.2
Reside in Main Country Coded	No	1063	48.4
	Yes	563	25.6
	Unknown	481	21.9
PhD level	No	441	20
	Yes	1189	54.1
	Unknown	568	25.9
Government employee	No	1936	88.1
	Yes	67	3
	Unknown	194	8.8

degree. In addition, the vast majority of experts are not government employees (at least 88.1).[10] Regarding the age of the experts, Table 2 indicates that at least 6.6 percent 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.

Next, we need to address the question of the experts' qualification to code academic freedom issues. 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, p. 19). Yet when it comes to assessing academic freedom issues specifically, the fact that the V-Dem pool of respondents consists largely of academics makes them in fact especially qualified. At least 28.1% of experts reside in the main country that they code (see also Table 2) and most others have spent some time there and/or maintain close professional contacts within the country, meaning that they are very likely to have direct knowledge of the relevant higher education system, its inner workings and constraints.

[10]From the 67 such government employees, 34 were not living in the main country they were coding, while 33 did. 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.)

Moreover, as McMann et al. also note (2022, p. 436), V-Dem's selection of experts follows rigorous criteria designed to assure not only their expertise on the issues and countries they are charged to code, but also their impartiality. The specific expert selection criteria are discussed in Coppedge et al. (2020, Chapter 3.8). In sum, experts are recruited along the following criteria: (1) experts' expertise on the country or countries and surveys 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; (5) diversity in professional backgrounds among the coders chosen for a particular country. The identities of the participating experts are kept anonymous outside the V-Dem's core team. 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 paper.

In addition to the selection criteria, it is important to highlight that with typically more than 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, p. 436). In its v13 version, 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).[11] 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 (Freedom House, 2022, p. 2).[12]

The AFI's reliance on several independent experts per data point mitigates the issue of individual biases to a great extent. However, the high probability of direct personal exposure to academic freedom issues in the given country for a significant share of the expert pool – above noted as an advantage in terms of knowledge and qualification – could under certain circumstances be a source

[11]The individual indicators rely on average on the assessment of 6.09 to 6.51 experts per country-year.

[12]There is no public information on how these analysts and advisers are selected, whether they are mainly based in the USA, in the countries they are assessing, or in third countries, and how detailed and rigorous the review process is. The methodology note merely states that "The analysts' proposed scores are discussed and defended at a series of review meetings, organized by region and attended by Freedom House staff and a panel of expert advisers. The end product represents the consensus of the analysts, outside advisers, and Freedom House staff, who are responsible for any final decisions" (Freedom House, 2022, p. 2).

of collective bias in the data.[13] Such distortions may in most cases be alleviated by the inclusion of experts who reside outside the country in addition to those who live in the country, but they can still be problematic for the reliability of specific data points, as the Brazilian case showcases. For this reason, 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, 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.

3.3 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, 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 (Pemstein et al., 2023; Coppedge et al., 2023c), 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 like the ones used for the AFI indicators.

> "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, p. 436).

[13] 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, 2022, p. 15).

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, p. 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 were 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, p. 436).

McMann et al. (2022, p. 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 further below (section 3.6).

3.4 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 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 belong to each other. 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, p. 91)[14], while formative indicators treat "indicators as determinants of the concept being measured" (Coppedge et al., 2020, p. 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 2.2). 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 Bayesian 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. The individual indicators are available for others to construct their own academic freedom measure using different aggregation rules. To show that the aggregation rules are important considerations that affect the outcome measurement, we briefly discuss how different aggregation rules can be used and how the resulting measures are intercorrelated. In Supplementary Appendix C, we discuss two additional modes of aggregation: an additive academic freedom index (assuming mutually substitutable aspects of academic freedom), and a multiplicative academic freedom index (assuming that institutional autonomy is a prerequisite for academic freedom). 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. In particular, the findings presented in Supplementary Appendix C and in Table 3 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 *institutional autonomy* indicator as a necessary conditions for academic freedom.

[14]See also Treier and Jackman, 2008

Table 3. Correlation between 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

3.5 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, p. 437), provided the data quality is acceptable. 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 particularly 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 that reaches far back in time. By retrospectively covering years since 1900 (or since countries or their higher education system came into existence), it is far more suitable for analyses over time than Freedom House's D3 indicator, available only since 2012. Other sources of data, such as the incident data-based *Academic Freedom Monitor* of the Scholars at Risk Network or the Global Coalition to Protect Education from Attack's events database, are similarly available since 2013 and 2015. The only other data relating to academic freedom that reaches back in time (even to 1789) is the *Academic Freedom in Constitutions* dataset (Spannagel, 2023). However, it only indicates the *de jure* presence of academic freedom provisions in constitutions, not their realization. Outside the AFI, we have therefore no measure to comparatively assess what *de facto* academic freedom levels looked like in the early 21st century, let alone before then; which seriously limits any general conclusions we might want to draw from those types of data.

[15]Freedom House covers 202 countries/territories, among which are a number of microstates and (semi-)autonomous territories that V-Dem does not cover. Among the country units covered by V-Dem, the AFI specifically is currently missing for three historic countries (People's Democratic Republic of Yemen, Republic of Vietnam, Palestine/British Mandate).

Lastly, 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. Problems of changing data collection parameters are even more severe for events data, which makes them entirely inappropriate for systematic temporal (and even spatial) comparisons (see Spannagel, 2020, 199f.).

3.6 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, p. 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 Figure 1 and 2, we estimate the effect that different country characteristics, as well as the number of respondents have on *respondent disagreement*. In the following section, we estimate the effect different that coder and country characteristics have on *respondent ratings*, in order to identify potential systematic biases in the data generation process of the AFI data.

In the first step, 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.[16] 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 the V-Dem's Bayesian IRT model.

Figure 1 shows the standardized regression coefficients[17] that show the estimated effect size of the respective variable on the level of respondent disagreement. Thus, a positive regression coefficient

[16]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.

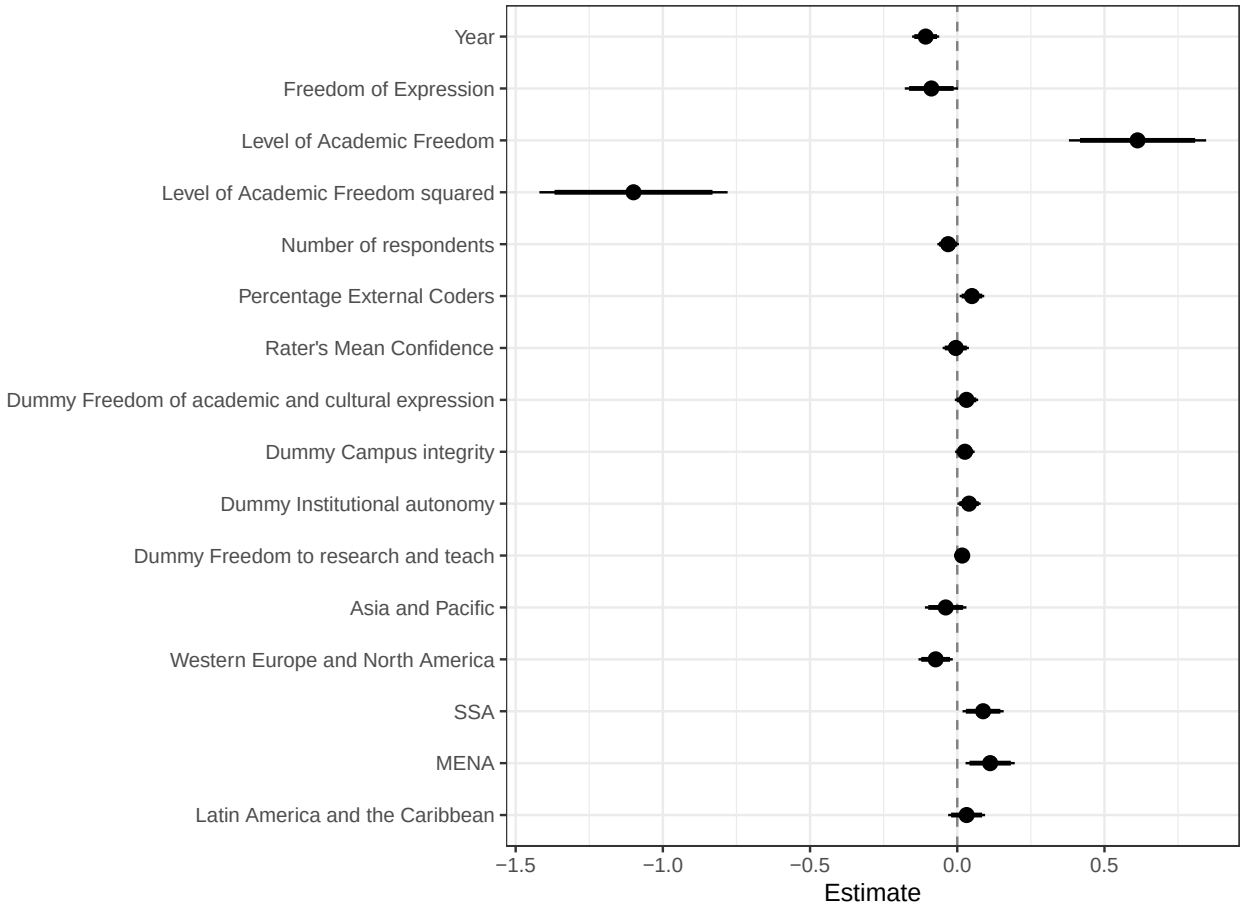
[17]For easier interpretation of the results, all regression coefficients were standardized by two standard deviations in Figure 2 and 3.

indicates that the variable is associated with increased respondent disagreement, while a negative regression coefficient indicates lesser respondent disagreement. Table E1 in the Supplementary Appendix shows the five separate indicators of the AFI, while Figure 1 shows the pooled model.[18]

Figure 1 indicates that respondent disagreement varies slightly depending on the freedom of expression in the country coded, indicating that limited access to information not only influences citizens in their countries but may also affect coders' ratings. Specifically, respondent disagreement is (statistically insignificantly) lower in countries with high levels of access to information (standardized coefficient = −0.09, 95% CI = [-0.178, 0.002]). We further control for the number of respondents per indicator and find that the higher the number of respondents for an indicator per country-date, the lower the overall disagreement between coders (standardized coefficient = −0.031, 95% CI = [-0.068, 0.005]). Moreover, Figure 1 shows that the percentage of external coders (not living in the country coded) increases the respondent disagreement slightly by 0.05 (95% CI = [0.008, 0.092]). 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 or born in the country. Moreover, we test for the raters' mean confidence and find that it does not substantially affect the respondent disagreement (standardized coefficient = 0.005, 95% CI = [-0.05, 0.04]).[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 = −0.107, 95% CI = [-0.154, -0.061]). This result may be surprising given the general idea that "the distant past is harder to code than the present" (McMann et al., 2022, p. 438). However, McMann et al. did also not find evidence for this claim. In reality, this appears thus to be 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. (Knutsen et al., 2023, p. 18) 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 also controls for regional effects and indicates that – compared to Eastern Europe and Central Asia – respondent disagreement is larger in Subsahara Africa and MENA, while respondent disagreement is lower in Western Europe and North America.

[18]The pooled model controls for indicator-fixed effects, as plotted in Figure 1.
[19]See Appendix I for an overview of the rater's confidence across indicators.
[20]We thank Katrin Kinzelbach for drawing our attention to this point.

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.

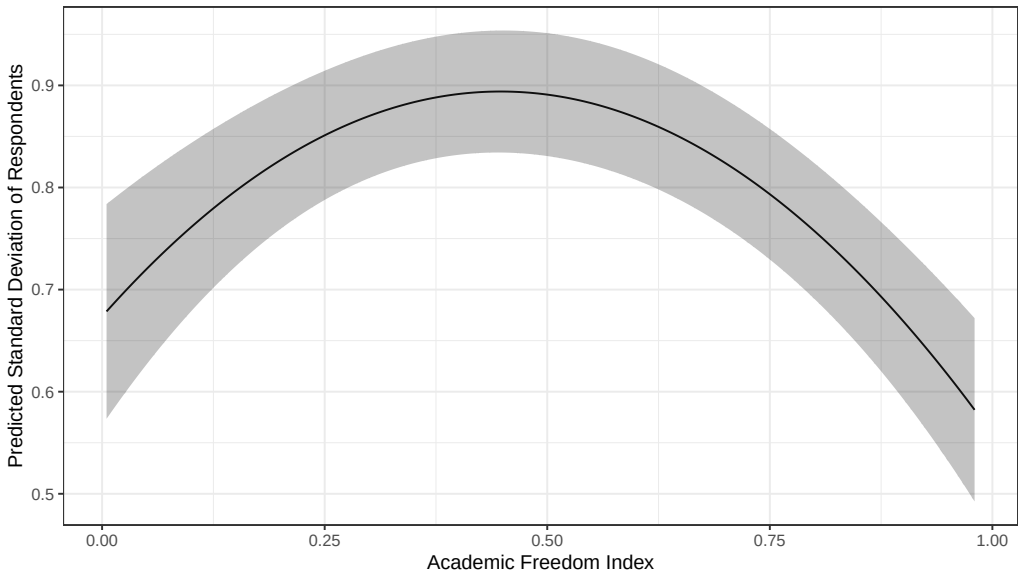
We test also for a nonlinear relationship between academic freedom levels and respondent disagreement by using the quadratic term for the level of academic freedom (standardized coefficient $= -1.1$, 95% CI = [-1.42, -0.78]). The results, which are also plotted in Figure 2, indicate that the greatest disagreement between respondents occurs, in fact, in countries with an Academic Freedom Index between 0.25 and 0.6, while the disagreement is the lowest in countries with well-protected academic freedom. This shows that academic freedom levels between complete absence of academic freedom and mid-levels of academic freedom (<0.75) 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. As a result, we would also expect a comparatively high agreement between experts. Although very low levels of 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, pp. 160–162).

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 a way that we expected.

3.7 Analyzing Individual Respondent Biases

In the next step, we analyze whether there are systematical 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, p. 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

Figure 2. Predicted respondent disagreement by AFI



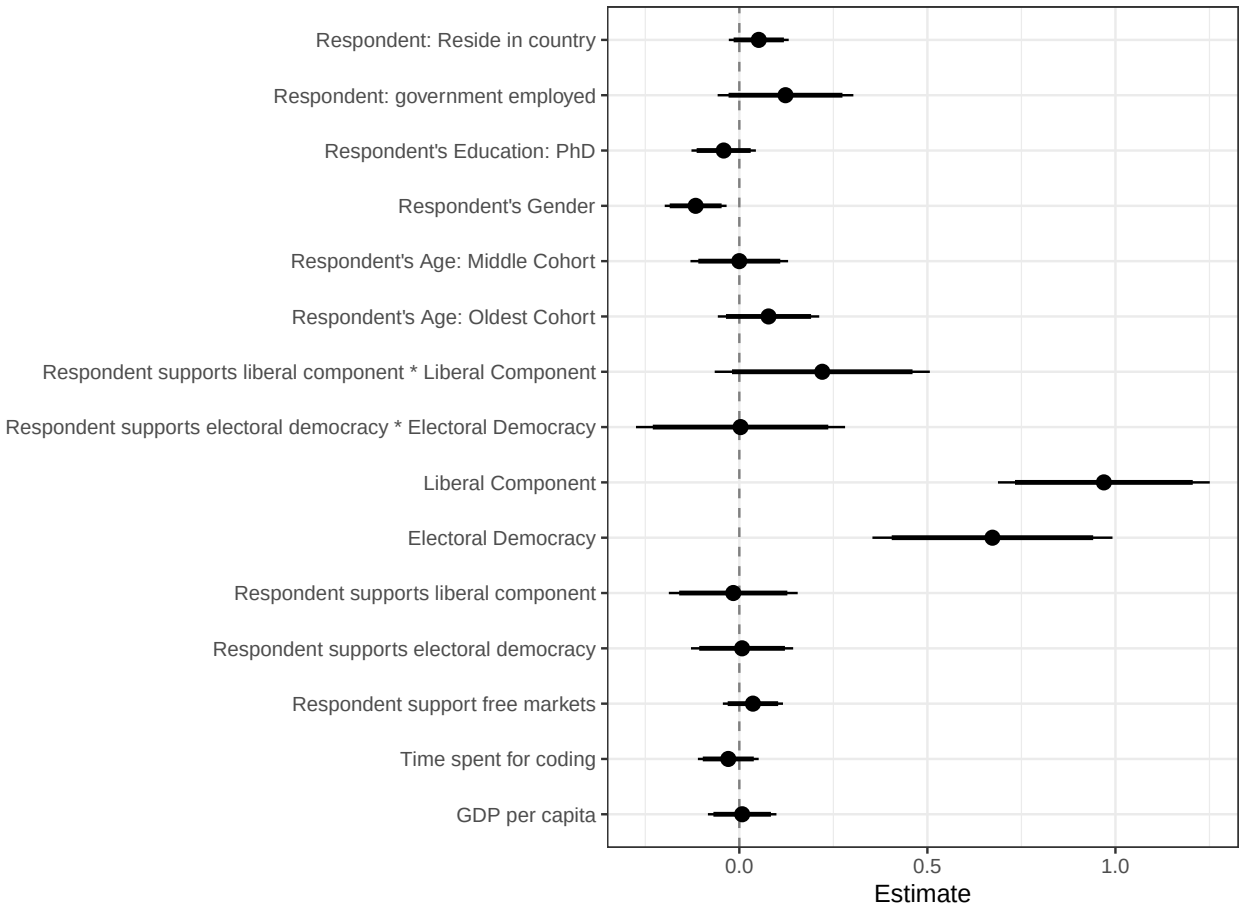
OLS regression with standard errors, clustered on countries.

respondents provide different ratings for academic freedom depending on whether they support (a) the principles of electoral democracy, (b) the principle of liberal democracy, and/or (c) free markets. We also test for a number of other individual-level factors that may influence respondents ratings (see Table 3 for distributions). We further control for the time experts spent on coding.

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' 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. Systematic biases are less problematic when they have a plausible theoretical explanation.

The results shown in Figure 3 and Table E2 in the Supplementary 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 = −0.036, 95% CI = [-0.044, 0.116]), nor respondents' support for electoral democracy (standardized coefficient = 0.007, 95% CI = [-0.129, 0.143]) or liberal democracy (standardized coefficient = −0.016, 95% CI = [-0.187, 0.155]) affect their ratings – indicated by the small and statistically

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.

insignificant effects. Figure 3 and 4 do, however, indicate that respondents rate more democratic (standardized coefficient = 0.673, 95% CI = [0.354, 0.992]) and liberal countries (standardized coefficient = 0.969, 95% CI = [0.688, 1-251]) 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.[21] 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 who reside in the country tend to code academic freedom somewhat higher compared to experts who assess the country from outside (standardized coefficient = 0.052, 95% CI = [-0.028, 0.131]). However, the effect size is substantially small and statistically not significant. Whether respondents are government employees[22] does not increase respondents ratings significantly (standardized coefficient = 0.123, 95% CI = [-0.058, 0.303]), nor does respondents' education level (i.e. PhD or not; standardized coefficient = −0.042, 95% CI = [-0.127, 0.044]).

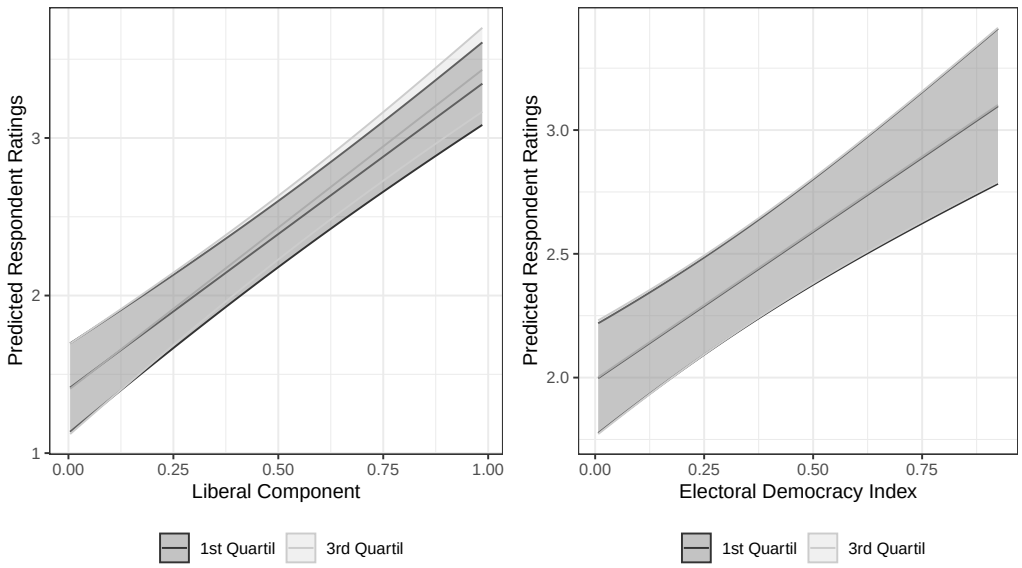
Respondents' gender coefficient, however, is negative and statistical significant at 0.05; female respondents rate academic freedom systematically lower compared to male respondents, all else equal (standardized coefficient = −0.116, 95% CI = [-0.198,-0.034]). To investigate this point further, we plot 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, but here the difference is statistically insignificant.[23] 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 marginal

[21]Standardized coefficient for EDI * Support for EDI = 0.003, 95% CI = [-0.275, 0.281], and standardized coefficient for Liberal component * Support for liberal principle = 0.221, 95% CI = [-0.065, 0.506]).

[22]See definition in supra note 10.

[23]In the predictions and average marginal effects plots shown in Figure 5A and B, different quantities were estimated compared to the regression coefficients in Figure 4, while the former are non-linear combinations of the latter, resulting in different levels of statistically significance. However, what matters are the substantial effects shown in Figure 5.

Figure 4. Predicted respondent ratings by Democratic Quality and First and Third Quartile of Respondent's Individual Support for Liberal/Electoral Democracy.



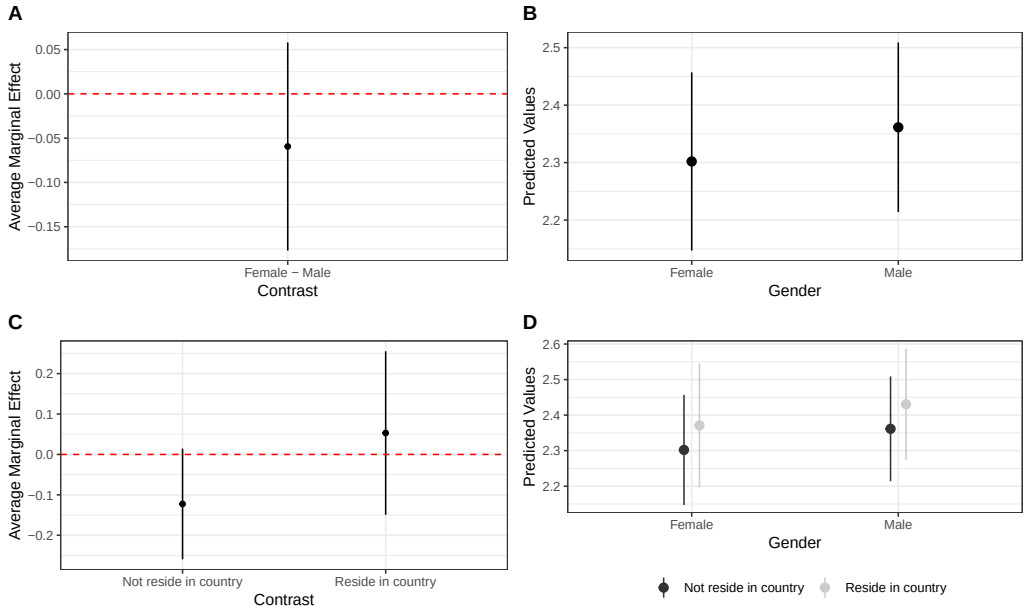
OLS regression with standard errors, clustered on countries. Measure- and year-fixed effects are included in the model.

effects plot in Figure 5C also shows that the difference is still statistically insignificant, though the share of female coders who have contributed to the AFI data is also relatively low overall (less than 30%).

Another possible explanation for these findings 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 (see section 2.1). While the differences are not statistically significant, and therefore not cause for serious concerns, they make a case for directing efforts at further diversifying the pool of expert coders.

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

Figure 5. Average Marginal Effects (A and C) and predicted respondent ratings (B and D) by Respondent's Gender and Respondent's Reside/Born in Country



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

4 Convergent Validity Assessment

In the next step, we analyze to what extent the academic freedom measure correspond 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, 2022). 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 systematic divergence between the chosen measure and the alternatives" (McMann et al., 2022, pp. 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 differ-

ences 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 3.6. 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) and 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 these 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, p. 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 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 Supplementary 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, p. 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 themselves do not establish validity" (Hawken and Munck, 2009, p. 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 Supplementary 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 6. Comparing the V-Dem Academic Freedom Index with Freedom House academic freedom measure (2012-2022).

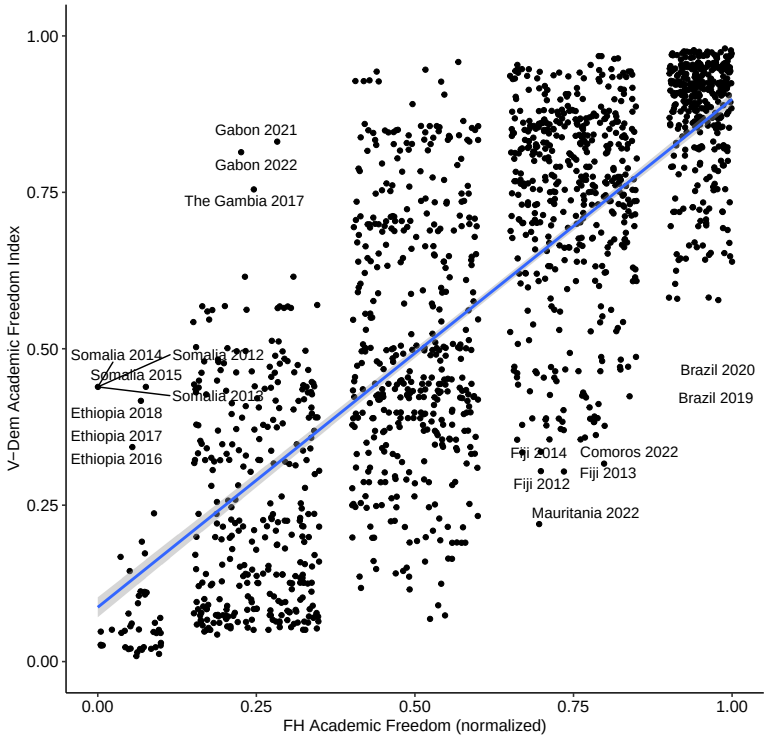
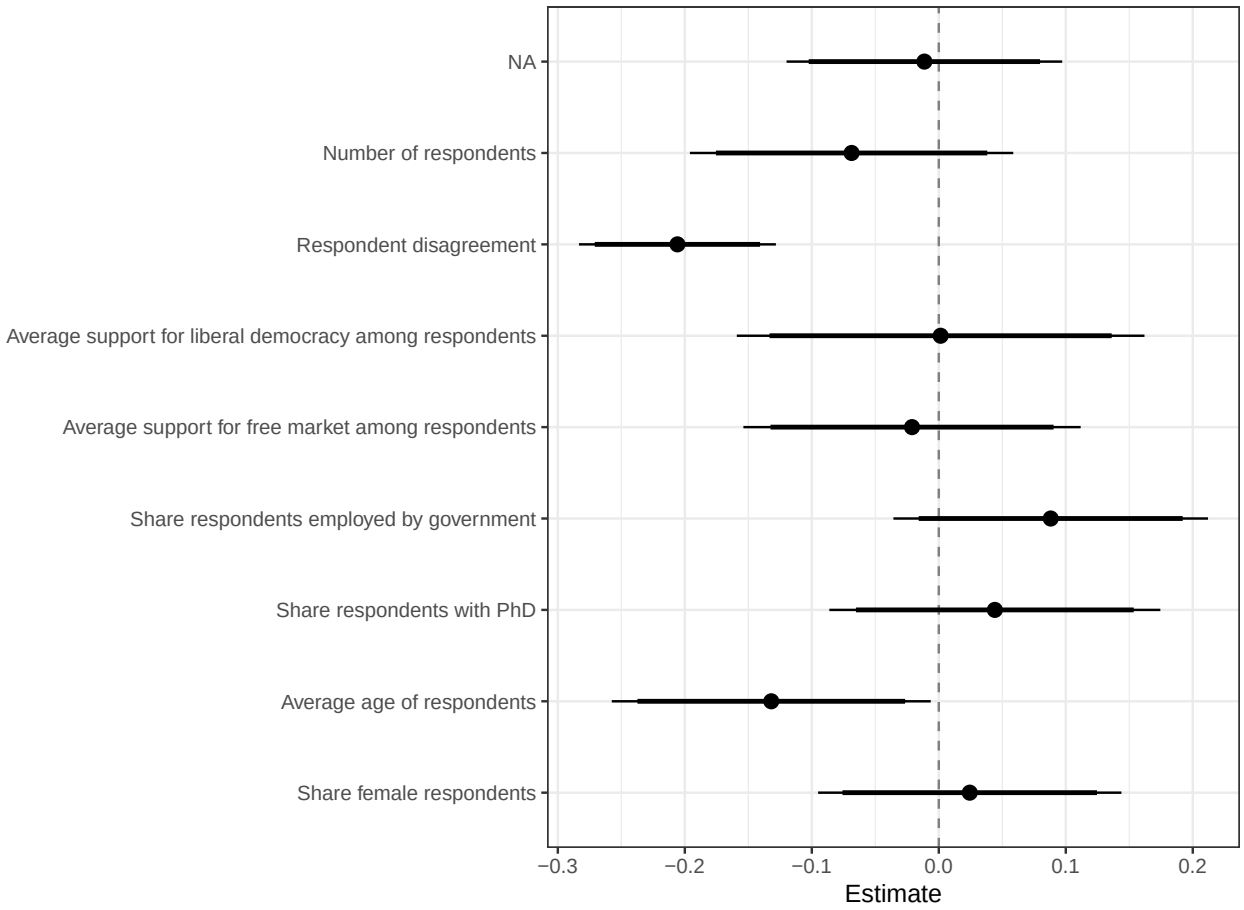


Figure 7 shows that the V-Dem share of female respondents as a predictor of divergence is slightly positive but statistically not significant (standardized coefficient = 0.024, 95% CI = [-0.095; 0.144]). A higher average age of V-Dem respondents (standardized coefficient = −0.132, 95% CI = [-0.258; -0.006]) significantly decreases the absolute difference between V-Dem and FH, while the share of V-Dem respondents with a PhD (standardized coefficient = 0.044, 95% CI = [-0.086; 0.147]) does not significantly affect the absolute difference between V-Dem and FH. In addition, the share of respondents employed by government coefficient is positive but also not statistically significant (standardized coefficient = 0.0.088, 95% CI = [-0.036; 0.212]). Moreover, whether a respondent resides in a country he/she is coding (standardized coefficient = −0.011, 95% CI = [-0.12; 0.097]) and whether respondents support free market (standardized coefficient = −0.021, 95% CI = [-0.154; 0.112]) or liberal democracy (standardized coefficient = 0.001, 95% CI = [-0.159; 0.162]) do not systematically increase the absolute residuals between V-Dem and FH.

The only other variable next to age that shows a significant effect is respondent disagreement in the coding, whose coefficient is negative (standardized coefficient = −0.206, 95% CI = [-0.283; -0.128]). Therefore, larger disagreement between V-Dem's respondents is associated with smaller absolute difference between V-Dem Academic Freedom Index and the FH academic freedom measure.

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.

However, this finding may be a statistical artifact and it should not be interpreted as causal. We cannot assess the connection in more detail as FH do not report expert disagreement and the expert consultation happens behind closed doors.[24] 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.

5 Conclusion

This article has explored 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 thereby contributed to different sets of literature. Firstly, we speak to the literature on data quality assessments (e.g., McMann et al., 2022; Seawright and Collier, 2014; Adcock and Collier, 2001; Sartori, 1970; Zeller and Carmines, 1980) by providing one of the first applications of McMann et al's approach. Secondly, we advance research on (how to measure) academic freedom (e.g. Abdel Latif, 2014; Spannagel, 2020; Appiagyei-Atua, Beiter, and Karran, 2016; Grimm and Saliba, 2017; Karran, Beiter, and Appiagyei-Atua, 2017; Pruvot, Estermann, and Popkhadze, 2023) 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, p. 445). This can help inform future research on academic freedom that seeks to make use of this newly introduced measure in substantive analyses.

In this discussion, we give various examples of how different aspects and assumptions of the AFI affect and should inform scholars and practitioners who wish to use the data for substantive research questions. First, in the content validity assessment, we show that the conceptualization of the AFI through five different dimensions, which are reflective indicators for 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

[24] In addition, the findings only include the years between 2012 and 2022.

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 customized ways to measure the latent construct. 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 highly 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 the AFI data does overall 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. This means there is most respondent disagreement at mid-levels of academic freedom, less at low levels, and the least at high levels of academic freedom. Respondent disagreement 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, particularly when examining countries with mid-range or recent volatile levels of academic freedom.

More generally, when working with latent variables, such as academic freedom or democracy, researchers should always take measurement uncertainty into account to avoid presenting inconclusive findings or biased standard errors. For this purpose, the V-Dem dataset provides measurement uncertainty estimates as well as posterior files to enable users to incorporate measurement noise into statistical and substantive analyses.[25]

[25]How measurement noise can drive empirical research is recently discussed in an exchange between Hu and coauthors (2022) and Claassen (2021)

References

Abdel Latif, Muhammad MM (2014). "Academic Freedom: Problems in Conceptualization and Research". In: *Higher Education Research & Development* 33.2, pp. 399–401.

Adcock, Robert and David Collier (Sept. 2001). "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research". In: *American Political Science Review* 95.3, pp. 529–546. DOI: 10.1017/S0003055401003100.

Altbach, Philip G. (2016). "Research Universities in Developing Countries". In: *Global Perspectives on Higher Education*. Ed. by Philip G. Altbach. John Hopkins University Press, pp. 172–198.

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". In: *Journal of Higher Education in Africa/Revue de l'enseignement supérieur en Afrique* 14.1, pp. 85–117.

Bollen, Kenneth A. and Pamela Paxton (2000). "Subjective Measures of Liberal Democracy". In: *Comparative Political Studies* 33.1, pp. 58–86. DOI: 10.1177/0010414000033001003.

Brook, Anne-Marie, K Chad Clay, and Susan Randolph (2019). *Human Rights Measurement Initiative Methodology Handbook*. URL: 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". In: *Victoria U. Wellington L. Rev.* 44, p. 487.

Church, A Timothy (2010). "Measurement Issues in Cross-Cultural Research". In: *The Sage Handbook of Measurement*. Ed. by Geoffrey Walford, Eric Tucker, and Madhu Viswanathan. SAGE, pp. 151–177.

Claassen, Christopher (2021). "The Democracy-Support Nexus Revisited: Accounting for Measurement Error and Simultaneous Effects". In: *Available at SSRN 3924934*.

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, and Yi-ting Wang (2020). *Varieties of Democracy: Measuring Two Centuries of Political Change*. Cambridge, United Kingdom: Cambridge University Press. ISBN: 978-1-108-42483-7.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, 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 (2023b). "V-Dem Codebook v13". In: *Varieties of Democracy (V-Dem) Project*.

— (2023a). "V-Dem [Country-Year/Country-Date] Dataset v13". In: *Varieties of Democracy (V-Dem) Project*. DOI: 10.23696/vdemds23.

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

ernance Indicators (SGI)". In: *European Policy Analysis* 8.2, pp. 136–159. DOI: https://doi.org/10.1002/epa2.1141.

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

— (2023). *Freedom in the World. Methodology Questions*. Bethesda. URL: 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". In: *Interdisciplinary Political Studies* 3.1, pp. 41–75.

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

Karran, Terence, Klaus Beiter, and Kwadwo Appiagyei-Atua (2017). "Measuring academic freedom in Europe: a criterion referenced approach". In: *Policy Reviews in Higher Education* 1.2, pp. 209–239. ISSN: 2332-2969. DOI: 10.1080/23322969.2017.1307093.

Knutsen, Carl Henrik, Kyle L. Marquardt, Brigitte Seim, Michael Coppedge, Amanda Edgell, Juraj Medzihorsky, Daniel Pemstein, Jan Teorell, John Gerring, and Staffan I. Lindberg (2023). "Conceptual and Measurement Issues in Assessing Democratic Backsliding". In: *University of Gothenburg: Varieties of Democracy Institute: V-Dem Working Paper No. 140*. University of Gothenburg: Varieties of Democracy Institute.

Macfarlane, Bruce (2012). "Re-framing student academic freedom: a capability perspective". In: *Higher Education* 63.6, pp. 719–732.

Marquardt, Kyle L and Daniel Pemstein (2018). "IRT Models for Expert-Coded Panel Data". In: *Political Analysis* 26.4, pp. 431–456.

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. by Remus Pricopie Adrian Curaj Ligia Deca. Springer, Cham, pp. 345–368.

McMann, Kelly, Daniel Pemstein, Brigitte Seim, Jan Teorell, and Staffan Lindberg (2022). "Assessing Data Quality: An Approach and An Application". In: *Political Analysis* 30.3, pp. 426–449. DOI: 10.1017/pan.2021.27.

Nokkala, Terhi and Agneta Bladh (2014). "Institutional autonomy and academic freedom in the Nordic context—similarities and differences". In: *Higher Education Policy* 27.1, pp. 1–21.

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". In: *V-Dem Working Paper No. 21. 8th edition*. University of Gothenburg: Varieties of Democracy Institute.

Pruvot, Enora Bennetot, Thomas Estermann, and Nino Popkhadze (2023). *University Autonomy in Europe IV, The Scorecard 2023*. URL: 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. by Richard Watermeyer, Rille Raaper, and Mark Olssen. Edward Elgar Publishing, pp. 106–125.

Sartori, Giovanni (1970). "Concept Misformation in Comparative Politics". In: *American Political Science Review* 64.4, pp. 1033–1053.

Sawyerr, Akilagpa (2004). "African universities and the challenge of research capacity development". In: *Journal of Higher Education in Africa/ Revue de l'enseignement supérieur en Afrique*, pp. 213–242.

Seawright, Jason and David Collier (2014). "Rival strategies of validation: Tools for evaluating measures of democracy". In: *Comparative Political Studies* 47.1, pp. 111–138.

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. by Katrin Kinzelbach. FAU University Press, 175—221.

— (2023). *Academic Freedom in Constitutions (AFC)*. Version V1. DOI: 10.7910/DVN/E8MIMF. URL: https://doi.org/https://doi.org/10.7910/DVN/E8MIMF.

Spannagel, Janika and Katrin Kinzelbach (2022). "The Academic Freedom Index and Other New Indicators Relating to Academic Space". In: *Quantity and Quality*.

Tai, Yuehong 'Cassandra', Yue Hu, and Frederick Solt (2022). "Democracy, Public Support, and Measurement Uncertainty". In: *American Political Science Review*, pp. 1–7.

Treier, Shawn and Simon Jackman (2008). "Democracy as a Latent Variable". In: *American Journal of Political Science* 52.1, pp. 201–217.

Varieties of Democracy Project (2022). *Varieties of Democracy Global Team*. Gothenburg. URL: https://www.v-dem.net/about/v-dem-project/global-team/.

Zavale, Nelson Casimiro (2022). "Academic Freedom in Mozambique". In: *University Autonomy Decline*. Ed. by Kirsten Roberts Lyer, Ilyas Saliba, and Janika Spannagel. Routledge, pp. 92–118.

Zeller, Richard A and Edward G Carmines (1980). *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge University Press.

Appendix

Abstract

This Appendix provides supplementary information and additional analyses to accompany the article *Quality Assessment of the Academic Freedom Index - Strengths, Weaknesses, and How Best to Use It.* Section A includes the wording of the V-Dem academic freedom indicator questions and their answer categories posed to the experts. Section B displays the exploratory factor analysis in the content validity assessment included in the main text. In addition, it also examines the fit of the one-factor model and a number of alternative specifications using frequentist confirmatory factor analysis (CFA) techniques. Section C discusses the index-level aggregation and presents alternative ways to aggregate the indicators to academic freedom indices. Section D expands on the discussion about correlated errors across the Academic Freedom measures included in Section 3.1. In Section E, we provide additional models and findings that accompany the findings presented in the main paper in Figure 1 and 3. We also estimate models with the coder perceptions scores instead of the coder raw scores and thereby show that also the coder perceptions are unbiased in various dimensions (Section G). Section F shows additional information for the convergent analysis applied in the main paper.

A V-Dem Academic Freedom Indicators

A.1 *v2cafres* Freedom to research and teach

Question: To what extent are scholars free to develop and pursue their own research and teaching agendas without interference?
Clarification: Examples of interference include research agendas or teaching curricula being drafted, restricted, or fully censored by a non-academic actor; scholars being externally induced, through possible reprisals, to self-censor; or the university administration abusing its position of power to impose research or teaching agendas on individual academics. It also includes public pressure on academics - offline and online. We do not consider as interference restrictions that are due to research priorities, as well as ethical and quality standards, freely defined by the scholarly community as well as the development of standardized curricula by academics that aim to structure and enhance teaching.
Responses:

- 0: Completely restricted. When determining their research agenda or teaching curricula, scholars are, across all disciplines, consistently subject to interference or incentivized to self-censor.
- 1: Severely restricted. When determining their research agenda or teaching curricula, scholars are, in some disciplines, consistently subject to interference or incentivized to self-censor.
- 2: Moderately restricted. When determining their research agenda or teaching curricula, scholars are occasionally subject to interference or incentivized to self-censor.
- 3: Mostly free. When determining their research agenda or teaching curricula, scholars are rarely subject to interference or incentivized to self-censor.
- 4: Fully free. When determining their research agenda or teaching curricula, scholars are not subject to interference or incentivized to self-censor.

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

A.2 *v2cafexch* Freedom of academic exchange and dissemination

Question: To what extent are scholars free to exchange and communicate research ideas and findings?
Clarification: Free academic exchange includes uncensored access to research material, unhindered participation in national or international academic conferences, and the uncensored publication of academic material. Free dissemination refers to the unrestricted possibility for scholars to share and explain research findings in their field of expertise to non-academic audiences through media engagement or public lectures.
Responses:

- 0: Completely restricted. Academic exchange and dissemination is, across all disciplines, consistently subject to censorship, self-censorship or other restrictions.
- 1: Severely restricted. Academic exchange and dissemination is, in some disciplines, consistently subject to censorship, self-censorship or other restrictions.
- 2: Moderately restricted. Academic exchange and dissemination is occasionally subject to censorship, self-censorship or other restrictions.
- 3: Mostly free. Academic exchange and dissemination is rarely subject to censorship, self-censorship or other restrictions.

- 4: Fully free. Academic exchange and dissemination is not subject to censorship, self-censorship or other restrictions.

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

A.3 *v2cainsaut* Institutional autonomy

Question: To what extent do universities exercise institutional autonomy in practice?
Clarification: Institutional autonomy "means the independence of institutions of higher education from the State and all other forces of society, to make decisions regarding its internal government, finance, administration, and to establish its policies of education, research, extension work and other related activities" (Lima Declaration). Note that institutional autonomy does not preclude universities from accepting state or third party funding, but does require that they remain in charge of all types of decisions listed above. Institutional autonomy does also not preclude a public oversight role by the state over universities' spending of public funds.
Responses:

- 0: No autonomy at all. Universities do not exercise any degree of institutional autonomy; non-academic actors control decision-making.
- 1: Minimal autonomy. Universities exercise only very limited institutional autonomy; non-academic actors interfere extensively with decision-making.
- 2: Moderate autonomy. Universities exercise some institutional autonomy; non-academic actors interfere moderately with decision-making.
- 3: Substantial autonomy. Universities exercise institutional autonomy to a large extent; non-academic actors have only rare and minimal influence on decision-making.
- 4: Complete autonomy. Universities exercise complete institutional autonomy from non-academic actors.

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

A.4 *v2casurv* Campus integrity

Question: To what extent are campuses free from politically motivated surveillance or security infringements?
Clarification: "Campus" refers to all university buildings as well as digital research and teaching platforms. Campus integrity means the preservation of an open learning and research environment marked by an absence of an externally induced climate of insecurity or intimidation on campus. Examples of infringements of campus integrity are politically motivated on-campus or digital surveillance, presence by intelligence or security forces, presence of student militias, or violent attacks by third parties, if specifically targeting universities to repress academic life on campus. Note that we are only interested in politically motivated infringements and targeted attacks on campus integrity, not in non-political security concerns or proportionate security measures taken on campus to address these.
Responses:

- 0: Completely restricted. Campus integrity is fundamentally undermined by extensive surveillance and severe intimidation, including violence or closures.
- 1: Severely restricted. Campus integrity is to a large extent undermined by surveillance and intimidation, at times including violence or closures.

- 2: Moderately restricted. Campus integrity is challenged by some significant cases of surveillance or intimidation.
- 3: Mostly free. Campus integrity is to a large extent respected, with only minor cases of surveillance or intimidation.
- 4: Fully free. Campus integrity is comprehensively respected; there are no cases of surveillance or intimidation.

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

A.5 *v2clacfree* Freedom of academic and cultural expression

Question: Is there academic freedom and freedom of cultural expression related to political issues?
Clarification: No clarification
Responses:

- 0: Not respected by public authorities. Censorship and intimidation are frequent. Academic activities and cultural expressions are severely restricted or controlled by the government.
- 1: Weakly respected by public authorities. Academic freedom and freedom of cultural expression are practiced occasionally, but direct criticism of the government is mostly met with repression.
- 2: Somewhat respected by public authorities. Academic freedom and freedom of cultural expression are practiced routinely, but strong criticism of the government is sometimes met with repression.
- 3: Mostly respected by public authorities. There are few limitations on academic freedom and freedom of cultural expression, and resulting sanctions tend to be infrequent and soft.
- 4: Fully respected by public authorities. There are no restrictions on academic freedom or cultural expression.

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

B Two-dimensional factor analysis

Table B1. Conceptual Alignment across V-Dem academic freedom indicators (Unidimensional Frequentist Factor Analysis)

Measure	Loadings	Uniqueness
Freedom to research and teach **v2cafres**	0.971	0.058
Freedom of academic exchange and dissemination **v2cafexch**	0.977	0.045
Institutional Autonomy **v2cainsaut**	0.897	0.195
Campus integrity **v2casurv**	0.925	0.144
Freedom of academic and cultural expression **v2clacfree**	0.87	0.244

Table B2. Conceptual Alignment across V-Dem academic freedom indicators (Two-dimensional Frequentist Factor Analysis)

Factor	Measure	Loadings	Uniqueness
Dimension 1	Freedom to research and teach **v2cafres**	0.971	0.058
	Freedom of academic exchange and dissemination **v2cafexch**	0.977	0.045
	Freedom of academic and cultural expression **v2clacfree**	0.897	0.195
Dimension 2	Institutional Autonomy **v2cainsaut**	0.925	0.144
	Campus integrity **v2casurv**	0.87	0.244

C Index-level aggregation

In this section, we present two alternative ways to aggregate indicators that we originally used to construct the Academic Freedom Index as a latent construct. The two alternative aggregation rules presented here assume that the indicators are formative indicators for the academic freedom concept. As described in the main text, the second choice is whether indicators are treated as (partially) mutually substitutable aspects of a given concept (additive aggregation rule) or as individually necessary conditions (multiplicative aggregation rule) for it. The equation used to construct the *additive academic freedom index* is defined as:

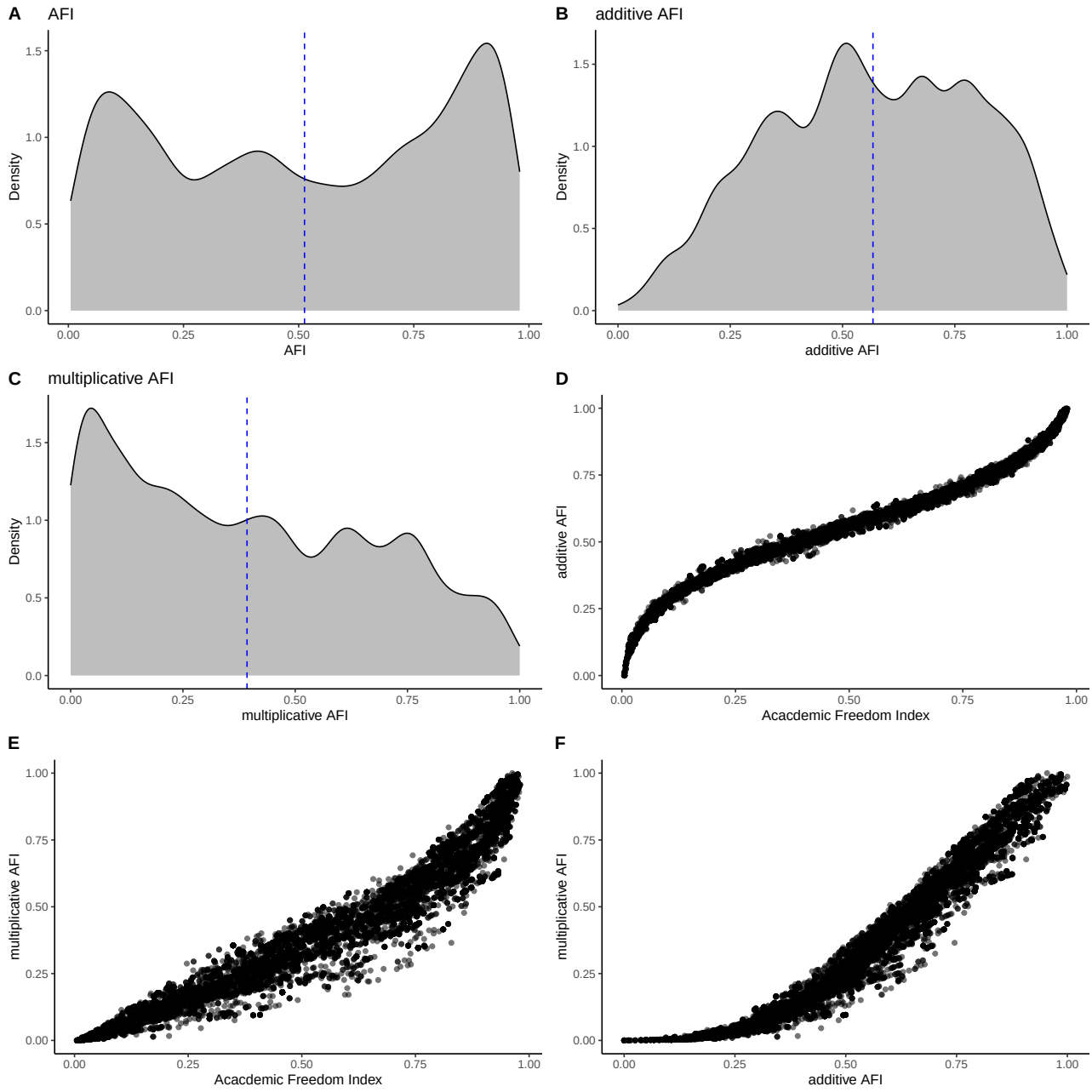
$$\begin{aligned} Additive\ AFI = (v2cafres + v2cafexch + v2clacfree + \\ v2cainsaut + v2casurv)/5 \end{aligned} \tag{1}$$

The equation used to construct the proposed *multiplicative academic freedom index* is defined as:

$$\begin{aligned} Multiplicative\ AFI = v2cainsaut_{osp} * \\ ((v2cafres_{osp} + v2cafexch_{osp} + v2clacfree_{osp} + v2casurv_{osp})/4) \end{aligned} \tag{2}$$

Figure C1 presents the empirical distribution of all three proposed academic freedom measures along with the country-year based correlations between them. Figure C1D contrasts the *academic AFI* and the original AFI, demonstrating that they the *academic AFI* discriminate at two different ends of the underlying academic freedom index distribution. Figure C1C reveals that the multiplicative AFI is left skewed and thus is more demanding compared to the original AFI and the additive AFI as theoretically expected. Figure C1E contrasts the original AFI and the *multiplicative AFI*. It reveals that the multiplicative AFI scores are systematically lower compared to the original AFI. A similar pattern can be observed when comparing the additive and the multiplicative AFI. Overall, the graphical presentation supports the notion that different aggregation procedures lead to different indices.

Figure C1. Aggregation to Academic Freedom Index



D Respondent-Correlated Errors

Our experts complete surveys of questions for the Academic Freedom Index and the V-Dem project, often more than one. Therefore, the same experts are likely to provide information about multiple academic freedom questions, such as *Freedom to research and teach* and *institutional autonomy*. The Academic Freedom Index questions appear in two separate surveys. Thus, it is highly likely that the same expert rate at least the questions asked in the *Civic and Academic Space* survey. In addition, those experts that answer the questions that appear in the *Civic and Academic Space* survey are not highly likely to also answer questions in the *Civil liberty* survey, which asks the question on the *freedom of academic and cultural expression*. As shortly discussed in the main paper, this fact may be a weakness of the V-Dem data generation process, "because correlated errors can inflate relationships between V-Dem variables, to an extent unwarranted by the latent variables that these measures attempt to capture" (McMann et al., 2022, p. 9). Therefore, caution is indicated when placing these variables on both sides of a regression equation. In addition, this fact also might mean that "some of the covariance in our factor analytic analyses stems from correlated respondent errors, rather than strong reflection of an underlying factor" (McMann et al., 2022, p. 9). In this step, we examine the extent to which respondent-correlated errors are likely to contaminate the V-Dem academic freedom measures.

We begin with the analysis of raw errors in rater scores. We use version 13 of the V-Dem dataset and calculate the average codings for the raw survey scales for each of the fifth measures and calculate the average deviation from those scores for each rater. Table D1 shows the pairwise-complete correlations between raw rater errors across the fifth scores. The correlations in Table D1 indicates that the are quite high, especially for the *Civic and Academic Space* survey. As a first clue, these high correlations seem to be worrying.

	v2cafexch	v2cafres	v2cainsaut	v2casurv	v2clacfree
v2cafexch	1.000	0.783	0.540	0.581	0.329
v2cafres	0.783	1.000	0.574	0.589	0.367
v2cainsaut	0.540	0.574	1.000	0.524	0.275
v2casurv	0.581	0.589	0.524	1.000	0.328
v2clacfree	0.329	0.367	0.275	0.328	1.000

Table D1. Raw Respondent Error Correlations

However, as McMann et al. argue "correlations in rater errors can stem from both systematic and stochastic sources" (McMann et al., 2022, p. 10). Stochastic errors are those that appear when a expert who gives a country A too high a score on one indicator my make a similar random mistake with respect to another indicator for country A. Therefore, a highly likely driver of correlated errors between these indicators is so-called differential item functioning (DIF). DIF appears when raters with higher standards apply those same high standards to every item they score. As described in Pemstein et al. (2023), the V-Dem measurement model explicitly models and adjusts for this sort of DIF when aggregating expert ratings. Because the V-Dem measurement model account for DIF, the analysis of raw respondent errors is error-prone and can lead to false conclusions. Therefore, we analyze whether model-corrected scores (also called rater perceptions by the V-Dem approach) are highly correlated with each other. Again, we use V-Dem version 13 data to calculate these perceptions from the posterior samples of parameters that V-Dem's ordinal item response theory model simulated for each indicator. For more information on the coder perceptions that are simulated from the posterior distributions see Pemstein et al. (2023).

We use these rater perception estimates to replicate the raw respondent-error correlation analysis from Table D1. After correcting for DIF, we find substantially lower correlations in respondents errors across academic freedom indicators. Table D2 indicates that after controlling for DIF with the V-Dem measurement model, few errors correlate above 0.3 across measures. However, we see some remaining evidence of correlated errors within *Civic and Academic Space* survey, ranging from 0.139 to 0.498. Especially the high correlation between the rater errors of campus integrity (v2casurv) and freedom to research and teach (v2cafres) indicate caution. Overall, Table D2 and Table D1 show that the rater errors are not completely uncorrelated and thus caution is warranted especially with analysis that use *Civic and Academic Space* survey items on both side of the equation.

	v2cafexch	v2cafres	v2cainsaut	v2casurv	v2clacfree
v2cafexch	1.000	0.436	0.188	0.327	0.083
v2cafres	0.436	1.000	0.191	0.493	0.053
v2cainsaut	0.188	0.191	1.000	0.139	0.079
v2casurv	0.327	0.498	0.139	1.000	0.038
v2clacfree	0.083	0.053	0.079	0.0308	1.000

Table D2. Model Adjusted Respondent Error Correlations

	v2cafres	v2cafexch	v2cainsaut	v2casurv	v2clacfree
v2cafres	**1130**	1128	1127	1127	746
v2cafexch	1128	**1130**	1127	1126	747
v2cainsaut	1127	1127	**1128**	1126	745
v2casurv	1127	1126	1126	**1128**	745
v2clacfree	746	747	745	745	**1838**

Table D3. Total pairwise coders, unique coders per indicator in the diagonal

E Examining Respondent Disagreement and Biases

E.1 Respondent Disagreement

Table E1 and Figure D1 extent the analysis from Figure 1 in the main paper to examine the correlates of respondent disagreement for each indicator separately (Model 1-5) in the addition to the academic freedom index (Model 6).

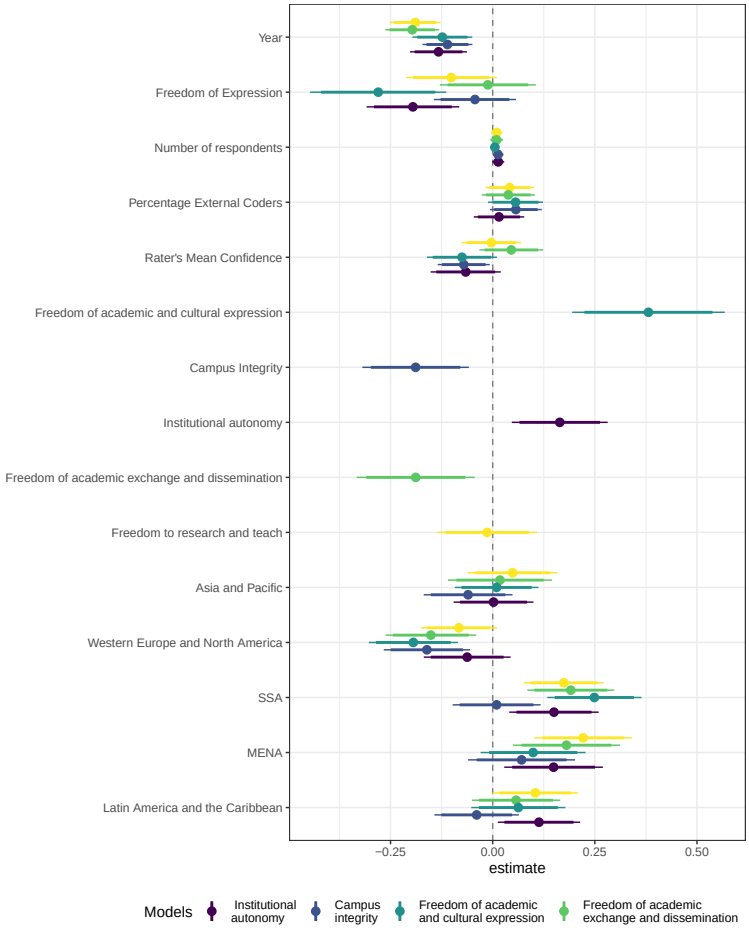
	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	6.163*	6.156*	4.851*	4.189*	4.764*	3.829*
	[4.447; 7.880]	[4.343; 7.969]	[2.929; 6.773]	[2.493; 5.885]	[2.915; 6.614]	[2.515; 5.143]
Year	−0.003*	−0.003*	−0.002*	−0.002*	−0.002*	−0.002*
	[−0.004; −0.002]	[−0.004; −0.002]	[−0.003; −0.001]	[−0.003; −0.001]	[−0.003; −0.001]	[−0.002; −0.001]
Freedom of Expression	−0.160	−0.019	−0.309*	−0.069	−0.443*	−0.139
	[−0.339; 0.018]	[−0.207; 0.170]	[−0.491; −0.127]	[−0.229; 0.092]	[−0.710; −0.176]	[−0.284; 0.006]
Freedom to research and teach	−0.005					0.017
	[−0.046; 0.037]					[−0.008; 0.041]
Number of Coders Freedom to research and teach	0.010					
	[−0.005; 0.025]					
Freedom of academic exchange and dissemination		−0.062*				
		[−0.111; −0.014]				
NoC Freedom of academic exchange and dissemination		0.010				
		[−0.007; 0.026]				
Institutional autonomy			0.057*			0.040*
			[0.016; 0.098]			[0.000; 0.080]
Number of Coders Institutional autonomy			0.014			
			[−0.002; 0.029]			
Campus integrity				−0.061*		0.026
				[−0.103; −0.018]		[−0.008; 0.059]
Number of Coders Campus integrity				0.013		
				[−0.001; 0.027]		
Freedom of academic and cultural expression					0.120*	0.031
					[0.060; 0.179]	[−0.009; 0.071]
NoC Freedom of academic and cultural expression					0.005	
					[−0.005; 0.016]	
Academic Freedom Index						0.984*
						[0.605; 1.364]
Academic Freedom Index sqr.						−1.100*
						[−1.423; −0.777]

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of Coders						−0.002
Academic Freedom Index						[−0.004; 0.000]
Percentage of Extern Coders	0.144	0.132	0.053	0.195	0.193	0.172*
	[−0.066; 0.354]	[−0.097; 0.361]	[−0.164; 0.270]	[−0.027; 0.418]	[−0.044; 0.429]	[0.025; 0.319]
Rater's Mean Confidence	−0.015	0.209	−0.301	−0.323*	−0.416	−0.023
	[−0.360; 0.329]	[−0.157; 0.575]	[−0.702; 0.099]	[−0.619; −0.027]	[−0.898; 0.065]	[−0.235; 0.188]
Latin America and the Caribbean	0.104	0.057	0.113*	−0.039	0.063	0.032
	[−0.003; 0.212]	[−0.054; 0.168]	[0.010; 0.217]	[−0.145; 0.067]	[−0.056; 0.181]	[−0.033; 0.096]
MENA	0.221*	0.181*	0.149*	0.071	0.099	0.112*
	[0.098; 0.344]	[0.046; 0.316]	[0.025; 0.273]	[−0.064; 0.206]	[−0.033; 0.231]	[0.026; 0.198]
SSA	0.174*	0.191*	0.150*	0.010	0.249*	0.088*
	[0.074; 0.274]	[0.082; 0.300]	[0.038; 0.262]	[−0.101; 0.120]	[0.130; 0.368]	[0.016; 0.159]
Western Europe and North America	−0.083	−0.151*	−0.062	−0.161*	−0.194*	−0.073*
	[−0.179; 0.013]	[−0.265; −0.038]	[−0.171; 0.046]	[−0.269; −0.053]	[−0.306; −0.082]	[−0.133; −0.014]
Asia and Pacific	0.049	0.018	0.002	−0.060	0.010	−0.039
	[−0.064; 0.161]	[−0.113; 0.149]	[−0.098; 0.102]	[−0.172; 0.051]	[−0.096; 0.115]	[−0.112; 0.033]
R^2	0.239	0.310	0.176	0.289	0.208	0.242
Adj. R^2	0.238	0.310	0.175	0.288	0.208	0.242
Num. obs.	14682	14674	14663	14666	18978	74512
RMSE	0.299	0.307	0.304	0.297	0.367	0.317
N Clusters	180	180	180	180	183	180

* Null hypothesis value outside the confidence interval.

Table E1. Linear Models predicting respondents' disagreement (raw coder ratings)

Figure D1. Predicted respondent disagreement (raw coder scores disagreement)



OLS regression with standard errors, clustered on countries.

E.2 Respondent Biases

Table D2 and Figure D2 extend our test for "situational closeness" from Figure 3 in the paper. In addition, they also extent the analysis of systematic bias resulting from different coder characteristics. The respondent-country characteristics interaction are - as in the main paper - insignificant for any of the individual indicators.

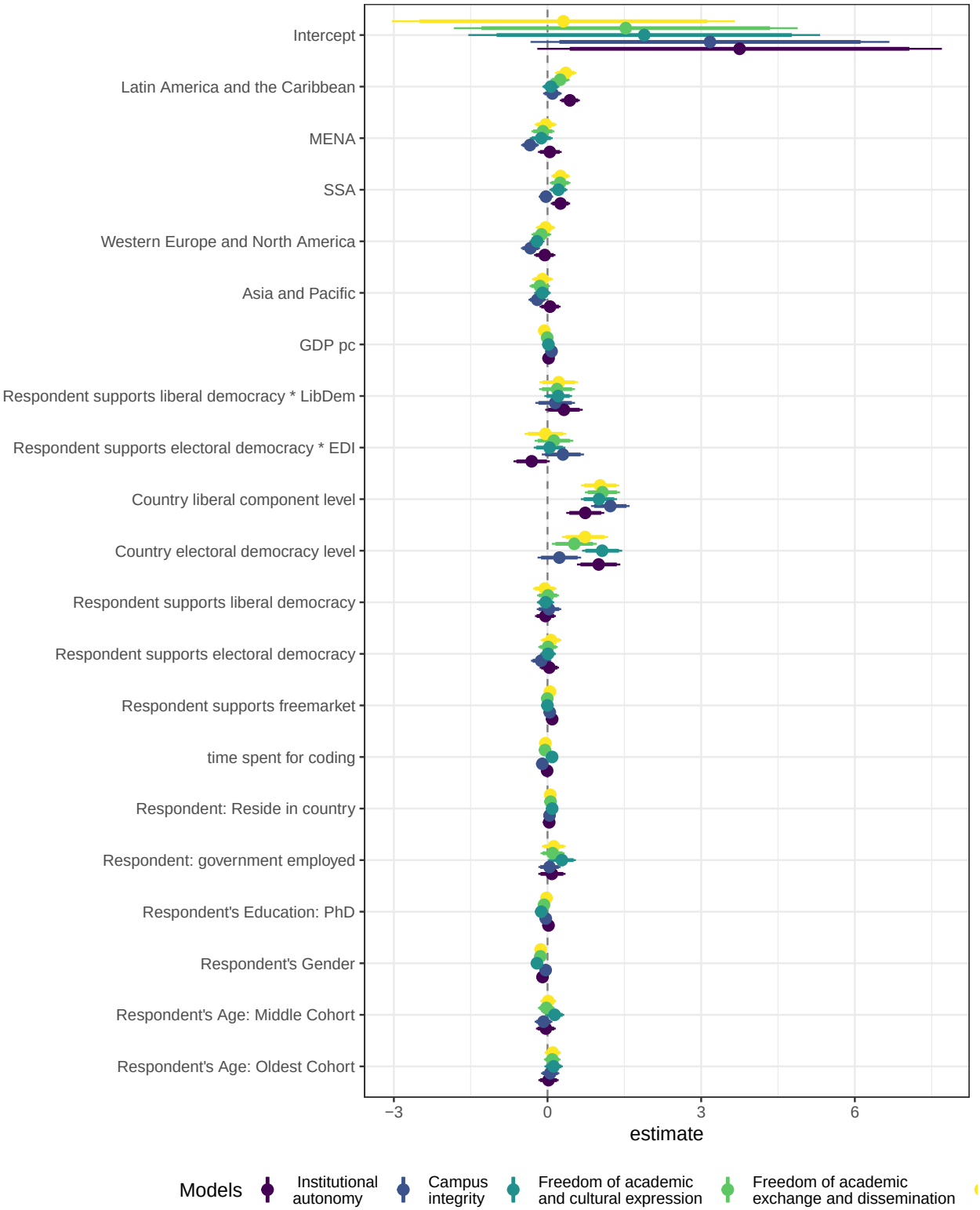
	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.310	1.527	3.750	3.172	1.888	1.032*
	[−3.082; 3.702]	[−1.874; 4.928]	[−0.251; 7.751]	[−0.378; 6.721]	[−1.590; 5.365]	[0.699; 1.364]
Respondent's Gender	−0.133*	−0.139*	−0.098	−0.036	−0.207*	−0.116*
	[−0.244; −0.023]	[−0.251; −0.027]	[−0.200; 0.005]	[−0.149; 0.076]	[−0.298; −0.117]	[−0.199; −0.033]
Respondent's Age: Middle Cohort	0.014	−0.024	−0.032	−0.078	0.139	−0.000
	[−0.145; 0.173]	[−0.191; 0.143]	[−0.230; 0.165]	[−0.254; 0.099]	[−0.054; 0.333]	[−0.133; 0.133]
Respondent's Age: Oldest Cohort	0.102	0.092	0.020	0.052	0.120	0.078
	[−0.061; 0.265]	[−0.073; 0.257]	[−0.183; 0.222]	[−0.130; 0.234]	[−0.066; 0.305]	[−0.060; 0.215]
Respondent: PhD education	−0.021	−0.069	0.019	−0.034	−0.123*	−0.042
	[−0.132; 0.090]	[−0.177; 0.040]	[−0.099; 0.137]	[−0.141; 0.074]	[−0.233; −0.013]	[−0.128; 0.045]
Respondent: Government employed	0.121	0.100	0.086	0.039	0.280	0.123
	[−0.122; 0.364]	[−0.144; 0.344]	[−0.191; 0.362]	[−0.187; 0.265]	[−0.005; 0.564]	[−0.064; 0.309]
Respondent: Reside in country	0.054	0.060	0.032	0.039	0.091	0.052
	[−0.050; 0.158]	[−0.043; 0.162]	[−0.074; 0.139]	[−0.065; 0.144]	[−0.012; 0.194]	[−0.029; 0.132]
Respondent supports free markets	0.020	−0.002	0.037	0.018	−0.000	0.015
	[−0.022; 0.061]	[−0.046; 0.042]	[−0.002; 0.076]	[−0.027; 0.063]	[−0.042; 0.041]	[−0.018; 0.048]
Respondent supports electoral democracy	0.036	0.005	0.018	−0.065	0.005	0.004
	[−0.070; 0.142]	[−0.096; 0.106]	[−0.081; 0.118]	[−0.170; 0.040]	[−0.078; 0.088]	[−0.069; 0.077]
Respondent supports liberal democracy	−0.026	0.005	−0.020	0.014	−0.019	−0.008
	[−0.143; 0.090]	[−0.103; 0.113]	[−0.123; 0.082]	[−0.107; 0.135]	[−0.105; 0.067]	[−0.096; 0.079]
time spent for coding	−0.001	−0.001	−0.000	−0.002	0.002	−0.001
	[−0.002; 0.001]	[−0.002; 0.001]	[−0.002; 0.002]	[−0.004; 0.000]	[−0.000; 0.004]	[−0.002; 0.001]
Country liberal component	1.735*	1.812*	1.245*	2.072*	1.697*	1.671*
	[1.092; 2.379]	[1.228; 2.396]	[0.608; 1.881]	[1.425; 2.718]	[1.090; 2.304]	[1.179; 2.163]
Country electoral democracy level	1.240*	0.885*	1.690*	0.391	1.806*	1.194*
	[0.460; 2.020]	[0.127; 1.643]	[0.959; 2.421]	[−0.344; 1.126]	[1.129; 2.482]	[0.618; 1.770]
Latin America and the Caribbean	0.355*	0.246*	0.436*	0.095	0.066	0.262*
	[0.142; 0.569]	[0.055; 0.437]	[0.236; 0.637]	[−0.091; 0.282]	[−0.105; 0.236]	[0.097; 0.427]
MENA	−0.036	−0.091	0.046	−0.342*	−0.122	−0.102
	[−0.257; 0.185]	[−0.323; 0.141]	[−0.195; 0.286]	[−0.523; −0.162]	[−0.355; 0.110]	[−0.284; 0.079]
SSA	0.254*	0.246*	0.253*	−0.031	0.213*	0.199*
	[0.069; 0.440]	[0.040; 0.452]	[0.060; 0.446]	[−0.176; 0.114]	[0.033; 0.393]	[0.054; 0.343]
Western Europe and North America	−0.039	−0.120	−0.054	−0.332*	−0.206*	−0.122
	[−0.234; 0.156]	[−0.315; 0.075]	[−0.268; 0.160]	[−0.523; −0.142]	[−0.356; −0.056]	[−0.273; 0.029]

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Asia and Pacific	−0.098	−0.150	0.050	−0.199*	−0.099	−0.088
	[−0.309; 0.113]	[−0.357; 0.056]	[−0.161; 0.262]	[−0.376; −0.023]	[−0.266; 0.068]	[−0.253; 0.076]
GDP pc	−0.003	−0.000	0.001	0.003	0.001	0.000
	[−0.007; 0.002]	[−0.005; 0.004]	[−0.003; 0.005]	[−0.002; 0.008]	[−0.004; 0.006]	[−0.004; 0.004]
Respondent supports liberal democracy * LibDem	0.097	0.082	0.142	0.066	0.092	0.098
	[−0.072; 0.266]	[−0.075; 0.240]	[−0.023; 0.307]	[−0.108; 0.240]	[−0.030; 0.214]	[−0.031; 0.226]
Respondent supports electoral democracy * EDI	−0.018	0.056	−0.138	0.133	0.018	0.001
	[−0.204; 0.167]	[−0.116; 0.227]	[−0.299; 0.024]	[−0.055; 0.321]	[−0.124; 0.159]	[−0.128; 0.131]
Dummy Freedom to research and teach						−0.114*
						[−0.142; −0.086]
Dummy Institutional Autonomy						−0.119*
						[−0.146; −0.093]
Dummy Campus Integrity						−0.131*
						[−0.174; −0.087]
Dummy Freedom of academic and cultural expression						−0.309*
						[−0.367; −0.251]
R^2	0.492	0.488	0.420	0.487	0.556	0.488
Adj. R^2	0.492	0.488	0.420	0.486	0.556	0.487
Num. obs.	81324	80992	80755	80183	60987	384241
RMSE	0.914	0.905	0.932	0.918	0.947	0.928
N Clusters	175	175	175	175	177	178

* Null hypothesis value outside the confidence interval.

Table D2. Linear Models predicting respondents rating with country characteristics

Figure D2. Predicting respondent ratings with respondent and country characteristics (raw coder ratings)

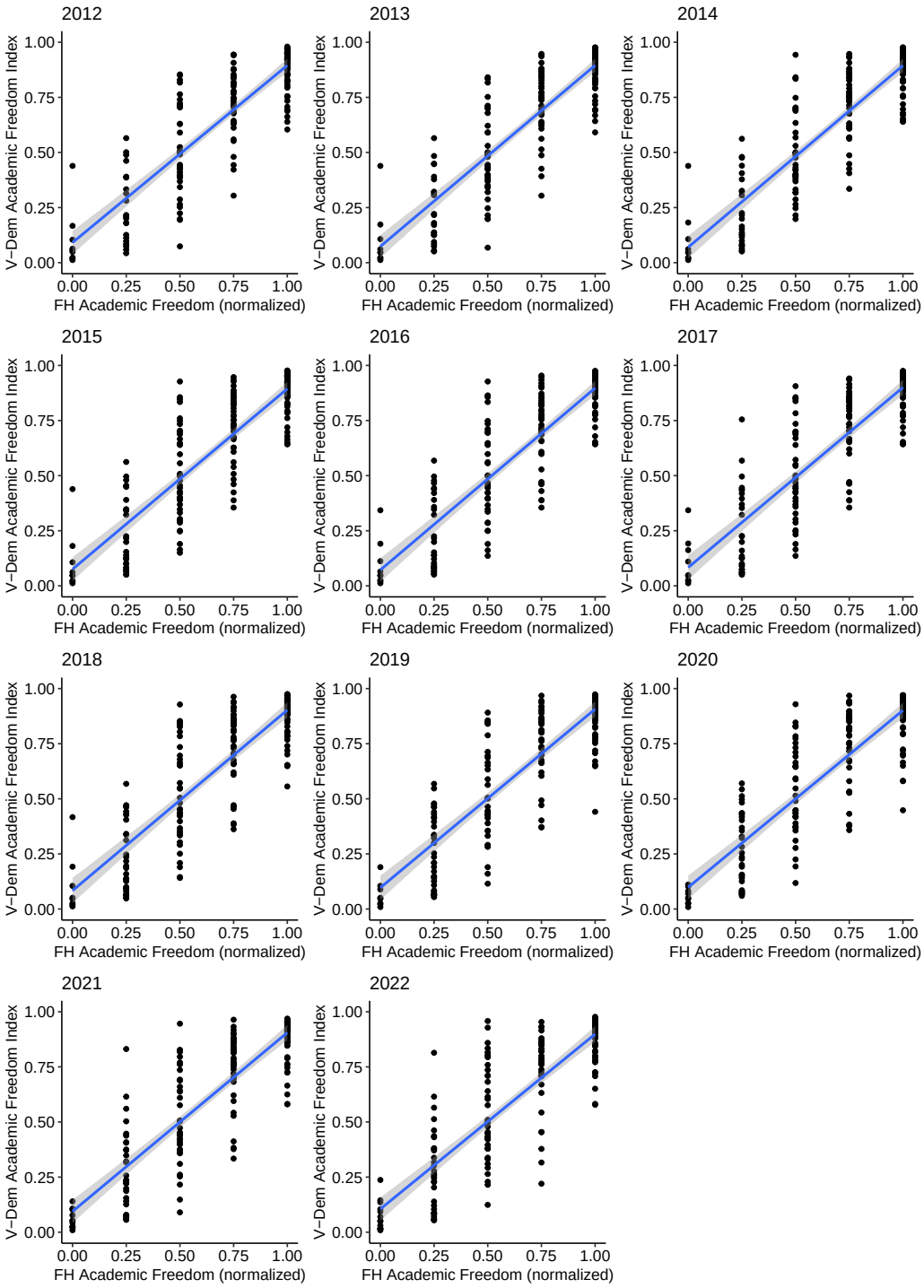


OLS regression with standard errors, clustered on countries. Measure-fixed effects, year-fixed effects are included in the model but omitted from the figure.

F Convergent Analysis Assessment

F.1 Traditional Convergent Assessment

Figure F1. Comparing the V-Dem Academic Freedom Index with Freedom House academic freedom measure for each year.



F.2 Statistical Analysis of Measure Convergence

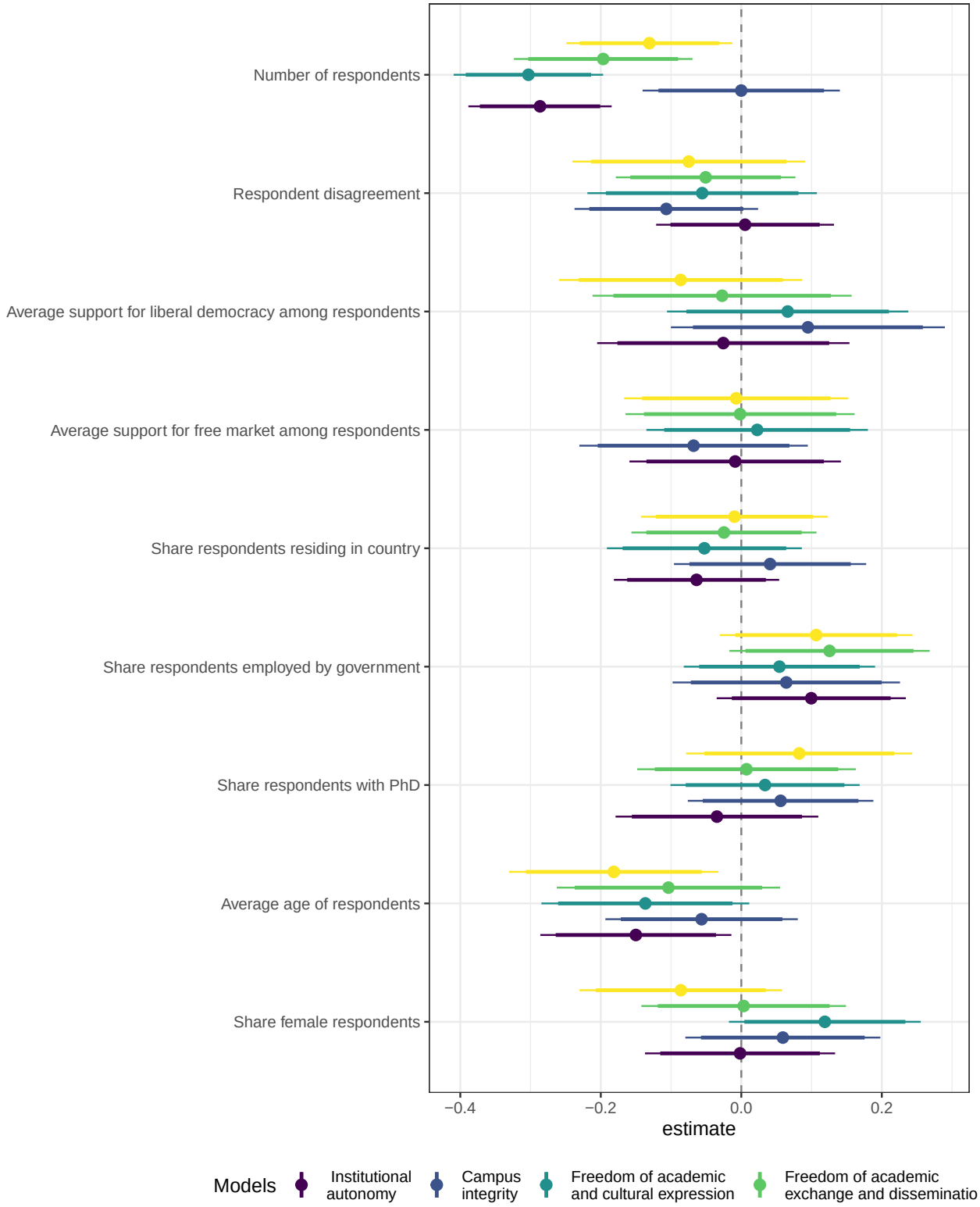
	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	0.331	0.181	0.585	−0.548	−0.377	0.209
	[−0.908; 1.570]	[−1.216; 1.579]	[−0.703; 1.872]	[−1.827; 0.730]	[−1.711; 0.957]	[−0.875; 1.292]
Respondent's Gender	−0.128*	−0.099	−0.080	−0.080	−0.070	−0.069
	[−0.216; −0.040]	[−0.210; 0.011]	[−0.189; 0.029]	[−0.181; 0.020]	[−0.162; 0.021]	[−0.157; 0.018]
Share female respondents	−0.274	0.011	−0.006	0.187	0.377	0.079
	[−0.741; 0.192]	[−0.459; 0.481]	[−0.441; 0.430]	[−0.259; 0.634]	[−0.063; 0.818]	[−0.312; 0.470]
Respondent's Age	0.050	0.038	0.075	0.038	0.127*	0.069*
	[−0.028; 0.127]	[−0.031; 0.106]	[−0.001; 0.151]	[−0.047; 0.123]	[0.050; 0.203]	[0.003; 0.136]
Average Age of respondents	−0.409*	−0.236	−0.342*	−0.129	−0.311	−0.304*
	[−0.750; −0.067]	[−0.605; 0.132]	[−0.658; −0.027]	[−0.447; 0.189]	[−0.655; 0.032]	[−0.598; −0.010]
Respondent: PhD education	−0.057	0.045	−0.016	−0.038	−0.045	−0.013
	[−0.146; 0.032]	[−0.061; 0.151]	[−0.122; 0.090]	[−0.155; 0.079]	[−0.155; 0.065]	[−0.095; 0.070]
Share of respondents with PhD	0.232	0.021	−0.097	0.156	0.094	0.124
	[−0.231; 0.695]	[−0.422; 0.463]	[−0.507; 0.313]	[−0.219; 0.532]	[−0.289; 0.477]	[−0.249; 0.497]
Respondent employed by government	0.120	0.208	0.068	0.211	−0.036	0.077
	[−0.072; 0.312]	[−0.063; 0.478]	[−0.197; 0.334]	[−0.087; 0.509]	[−0.307; 0.235]	[−0.142; 0.297]
Share respondents employed by government	0.814	0.950	0.753	0.484	0.411	0.674
	[−0.259; 1.887]	[−0.154; 2.054]	[−0.290; 1.796]	[−0.770; 1.739]	[−0.644; 1.465]	[−0.296; 1.645]
Respondent: Reside in country	0.134*	0.191*	0.183*	0.088	0.062	0.146*
	[0.043; 0.225]	[0.057; 0.324]	[0.058; 0.308]	[−0.054; 0.231]	[−0.069; 0.194]	[0.038; 0.253]
Share respondent reside in country	0.123*	0.171*	0.174*	0.091	0.049	0.133*
	[0.031; 0.215]	[0.058; 0.285]	[0.061; 0.286]	[−0.022; 0.203]	[−0.066; 0.164]	[0.041; 0.225]
Respondent supports free markets	0.041*	0.074*	0.055*	0.086*	0.057*	0.062*
	[0.002; 0.080]	[0.036; 0.112]	[0.016; 0.094]	[0.048; 0.125]	[0.018; 0.096]	[0.032; 0.092]
Average support for free market among respondents	−0.010	−0.002	−0.012	−0.090	0.030	−0.028
	[−0.226; 0.207]	[−0.223; 0.218]	[−0.215; 0.192]	[−0.310; 0.129]	[−0.183; 0.243]	[−0.205; 0.150]
Respondent supports electoral democracy	0.003	0.011	0.010	−0.040	−0.024	−0.012
	[−0.041; 0.047]	[−0.040; 0.062]	[−0.040; 0.061]	[−0.095; 0.015]	[−0.076; 0.027]	[−0.049; 0.024]
Average support for electoral democracy among respondents	0.304	0.121	0.081	0.112	0.175	0.195
	[−0.028; 0.636]	[−0.238; 0.481]	[−0.248; 0.411]	[−0.222; 0.445]	[−0.157; 0.507]	[−0.093; 0.482]
Respondent supports liberal democracy	0.005	−0.011	0.024	0.010	0.005	−0.001
	[−0.041; 0.051]	[−0.067; 0.046]	[−0.032; 0.081]	[−0.048; 0.068]	[−0.052; 0.061]	[−0.046; 0.044]
Average support for liberal democracy among respondents	−0.135	−0.043	−0.040	0.148	0.103	0.002
	[−0.413; 0.143]	[−0.337; 0.252]	[−0.326; 0.247]	[−0.163; 0.459]	[−0.171; 0.377]	[−0.258; 0.263]
Respondent Disagreement	−0.206*	−0.329*	−0.456*	−0.000	−0.472*	−0.341*
	[−0.394; −0.017]	[−0.547; −0.112]	[−0.622; −0.291]	[−0.279; 0.278]	[−0.641; −0.304]	[−0.470; −0.211]
Number of respondents	−0.012	−0.011	0.001	−0.024	−0.012	−0.003
	[−0.039; 0.015]	[−0.041; 0.019]	[−0.029; 0.031]	[−0.054; 0.007]	[−0.051; 0.026]	[−0.009; 0.003]
Dummy Freedom to research and teach						−0.122*

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dummy Institutional Autonomy						[−0.152; −0.093] −0.473* [−0.529; −0.417]
Dummy Campus Integrity						−0.169* [−0.236; −0.101]
Freedom of academic and cultural expression						−0.262* [−0.323; −0.200]
R^2	0.039	0.051	0.062	0.032	0.066	0.069
Adj. R^2	0.038	0.049	0.060	0.030	0.064	0.069
Num. obs.	16228	13817	13811	13800	13787	75021
RMSE	0.914	0.853	0.831	0.878	0.857	0.878
N Clusters	177	177	177	177	177	177

* Null hypothesis value outside the confidence interval.

Table F1. Explaining deviations from FH academic freedom indicator (D3) with aggregate respondent characteristics

Figure F2. Explaining deviations from FH academic freedom indicator (D3) with aggregate respondent characteristics, sub-indicators



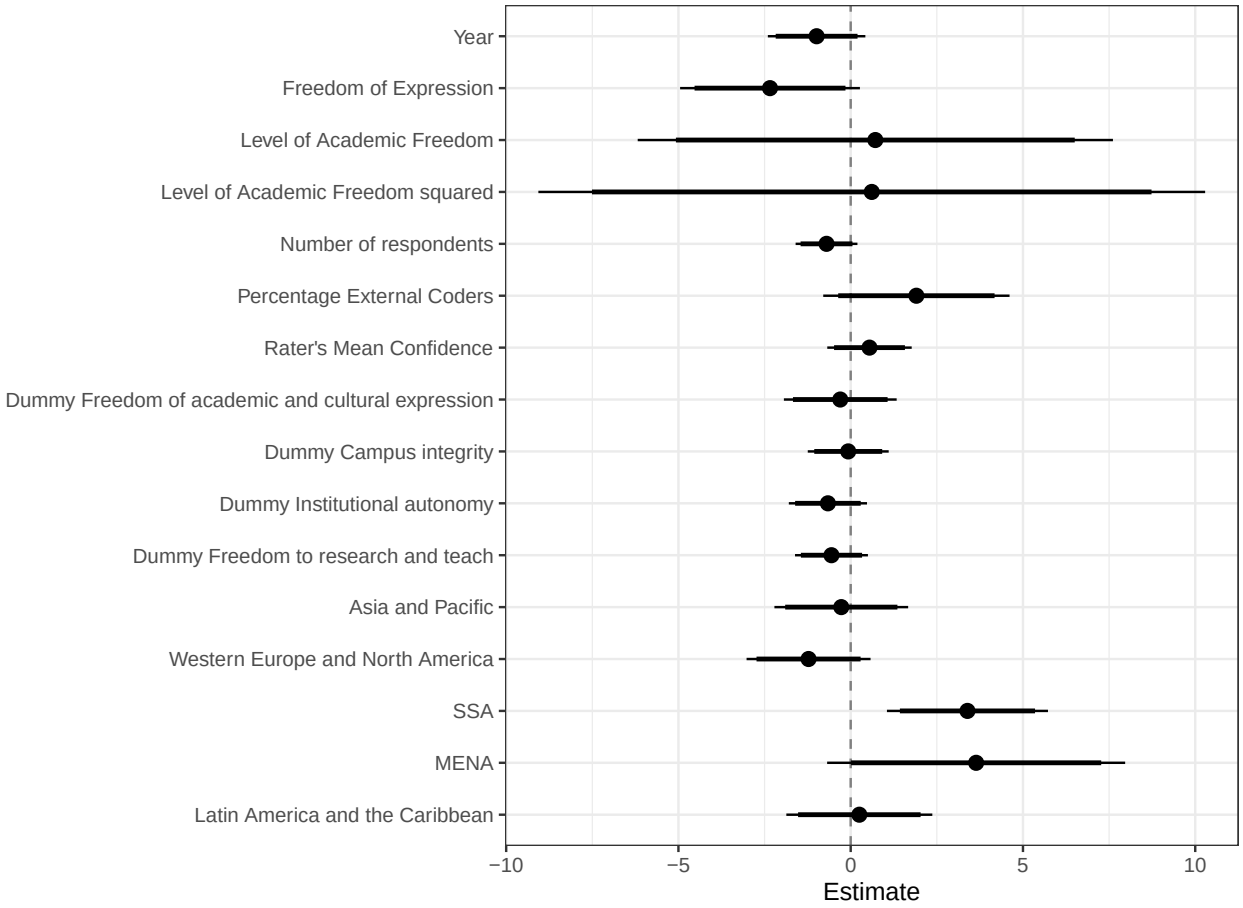
OLS regression with standard errors, clustered on countries. The dependent variable is the absolute residuals from regressing each V-Dem measure on Freedom House's D3 indicator on academic freedom and educational system. Year-fixed effects, and respondent characteristics are included in the model but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

G Respondent Disagreement and Biases - Perceptions

G.1 Respondent Disagreement - Perceptions

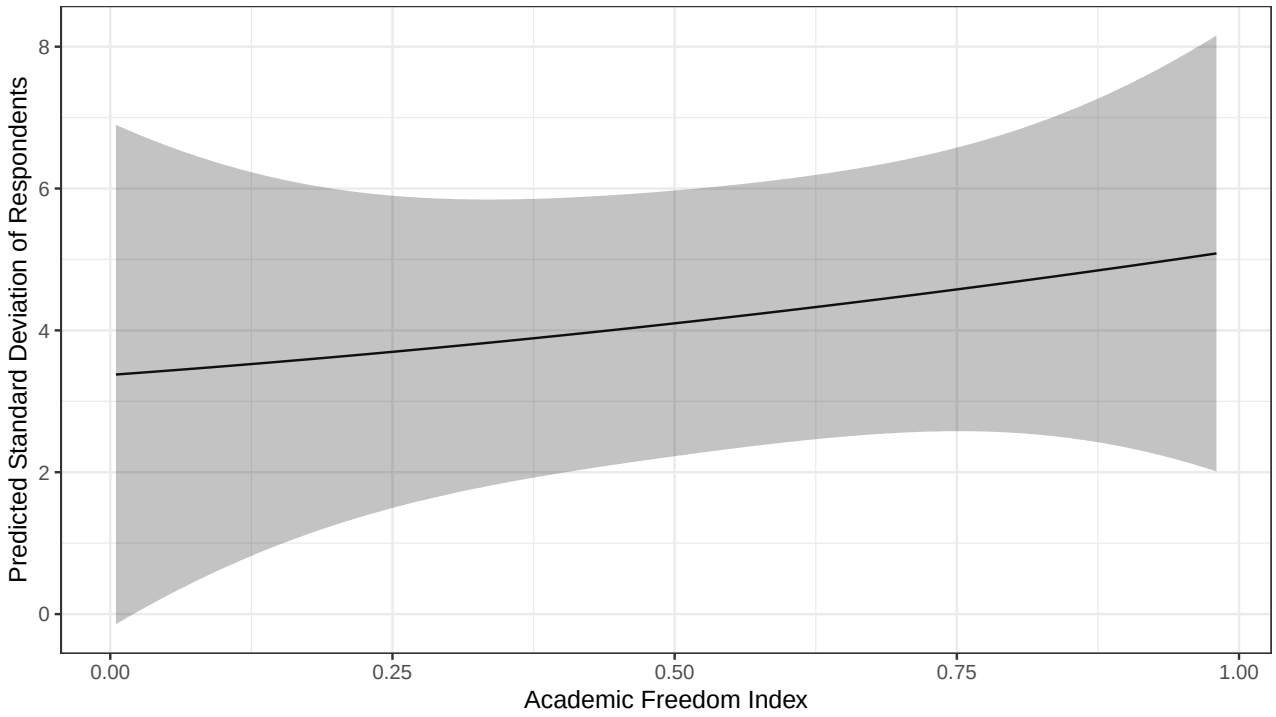
In addition, Table G1 and Figure G1 estimate coder disagreement using the standard deviation of measurement model-adjusted ratings among respondents for each country and year instead of raw ratings among respondents that do not correct for DIF. Thus, when the results in Table G1 and G1 show no systematic correlates as in the main paper, we can conclude that there is little evidence for systematic biases also when accounting for DIF.

Figure G1. 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.

Figure G2. Predicted respondent disagreement by AFI



OLS regression with standard errors, clustered on countries.

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	64.959*	43.421	39.501	45.952	40.809	32.141
	[4.835; 125.082]	[−17.751; 104.594]	[−9.511; 88.513]	[−18.356; 110.261]	[−10.436; 92.055]	[−7.846; 72.128]
Year	−0.037*	−0.024	−0.017	−0.022	−0.020	−0.015
	[−0.070; −0.004]	[−0.058; 0.009]	[−0.045; 0.011]	[−0.057; 0.013]	[−0.049; 0.010]	[−0.035; 0.006]
Freedom of Expression	−3.555	−2.516	−9.763*	−4.018	−3.524	−3.702
	[−8.001; 0.891]	[−8.373; 3.342]	[−16.626; −2.901]	[−8.660; 0.624]	[−13.454; 6.406]	[−7.889; 0.484]
Percentage of Extern Coders	8.496	5.056	2.519	10.905	0.231	6.572
	[−3.072; 20.065]	[−6.768; 16.880]	[−9.817; 14.854]	[−3.219; 25.029]	[−10.720; 11.182]	[−2.923; 16.068]
Rater's Mean Confidence	8.331*	6.580	−1.542	0.139	2.429	2.541
	[0.143; 16.520]	[−1.330; 14.489]	[−11.677; 8.593]	[−9.789; 10.066]	[−12.811; 17.668]	[−3.272; 8.354]
Freedom to research and teach	0.472					−0.558
	[−0.813; 1.756]					[−1.625; 0.509]
Number of Coders Freedom to research and teach	0.777*					
	[0.095; 1.458]					
Freedom of academic exchange and dissemination		0.796				
		[−0.466; 2.059]				
NoC Freedom of academic exchange and dissemination		0.588*				
		[0.012; 1.165]				
Institutional autonomy			1.480*			−0.662
			[0.281; 2.679]			[−1.805; 0.482]
NoC Institutional autonomy			0.401			
			[−0.079; 0.881]			
Campus integrity				0.613		−0.072
				[−0.773; 1.999]		[−1.256; 1.112]
NoCs Campus integrity				0.566		
				[−0.039; 1.171]		
Freedom of academic and cultural expression					0.111	−0.303
					[−1.924; 2.145]	[−1.953; 1.348]
NoC Freedom of academic and cultural expression					0.183	
					[−0.016; 0.382]	
Academic Freedom Index						1.149
						[−10.075; 12.373]
Academic Freedom Index sqr.						0.611
						[−9.168; 10.390]
Number of Coders Academic Freedom Index						−0.036
						[−0.083; 0.012]
Latin America and the Caribbean	2.472	1.964	2.686	−1.915	0.144	0.251
	[−1.359; 6.304]	[−1.798; 5.727]	[−0.515; 5.886]	[−6.660; 2.830]	[−2.407; 2.696]	[−1.931; 2.432]
MENA	4.808	7.049*	2.058	4.564	2.420	3.641
	[−0.930; 10.546]	[0.496; 13.602]	[−2.456; 6.572]	[−3.922; 13.049]	[−1.258; 6.099]	[−0.816; 8.098]
SSA	4.171	3.447	3.665*	0.143	7.852*	3.389*
	[−0.259; 8.601]	[−0.241; 7.135]	[0.465; 6.864]	[−5.953; 6.239]	[3.480; 12.225]	[0.995; 5.783]

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Western Europe and North America	−0.555	−1.381	0.326	−2.239	−1.048	−1.223
	[−3.395; 2.286]	[−5.156; 2.393]	[−1.859; 2.512]	[−6.820; 2.341]	[−3.509; 1.413]	[−3.069; 0.623]
Asia and Pacific	0.589	1.164	0.729	−1.615	−0.531	−0.273
	[−2.537; 3.716]	[−2.092; 4.420]	[−1.532; 2.990]	[−7.235; 4.006]	[−2.885; 1.822]	[−2.267; 1.721]
R^2	0.110	0.092	0.100	0.097	0.147	0.077
Adj. R^2	0.110	0.092	0.099	0.096	0.146	0.077
Num. obs.	14682	14674	14663	14666	18095	73814
RMSE	9.155	9.060	8.554	10.166	10.144	9.428
N Clusters	180	180	180	180	183	180

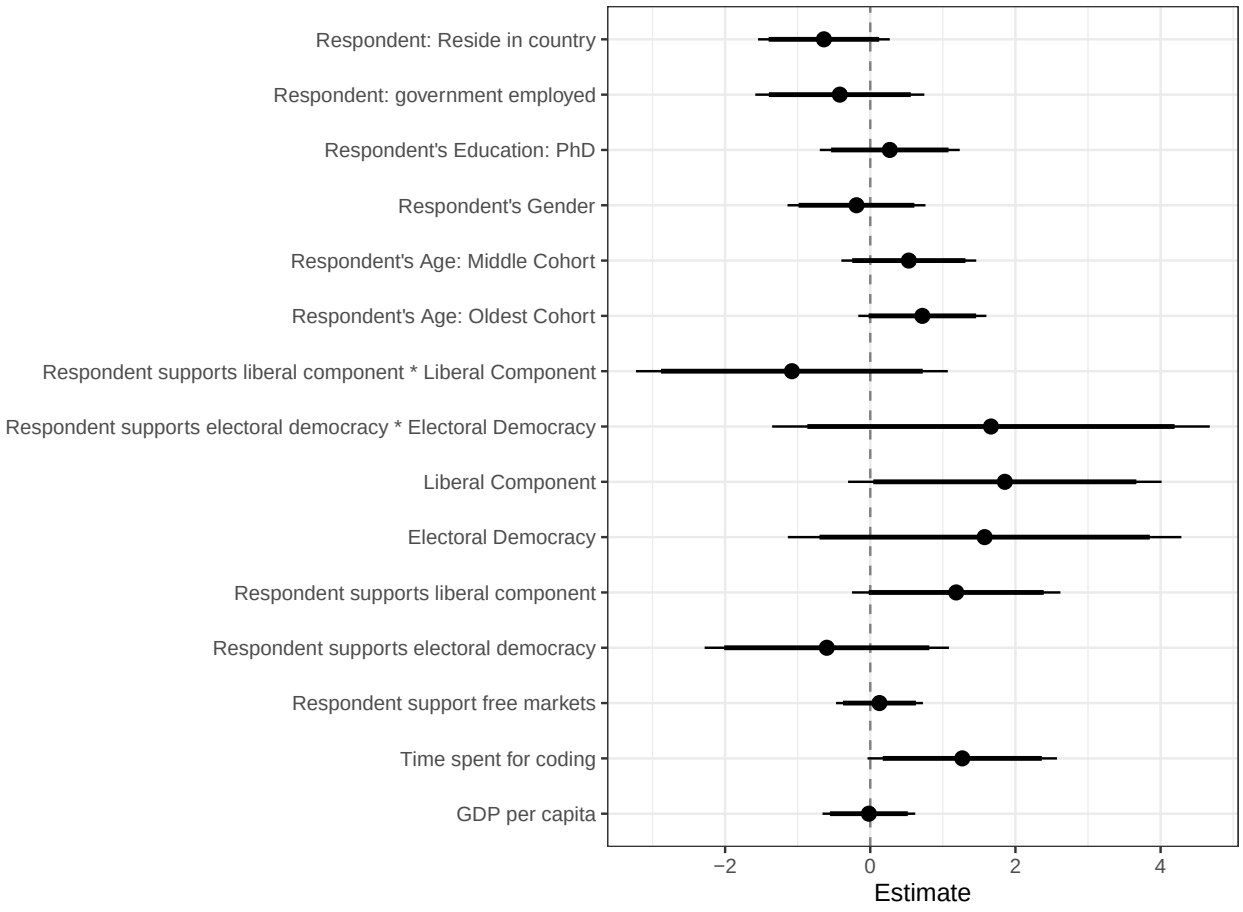
* Null hypothesis value outside the confidence interval.

Table G1. Linear Models predicting respondents' disagreement (coder perceptions)

G.2 Respondent Biases -Perceptions

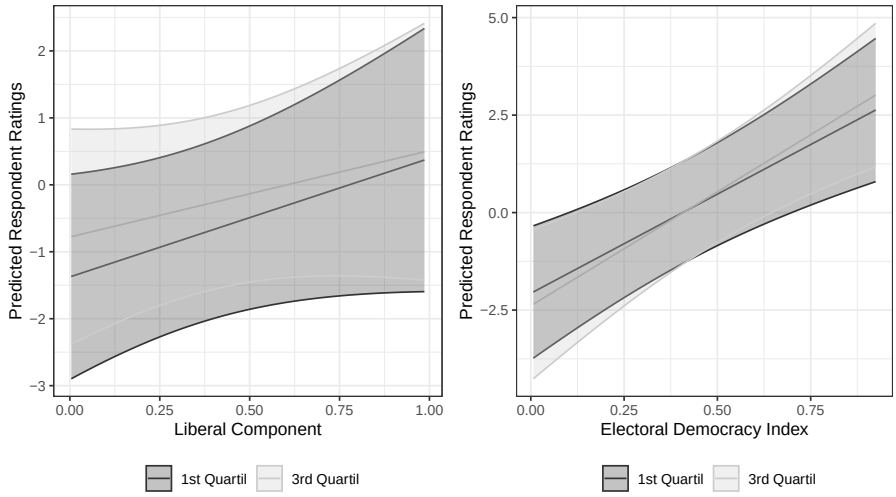
In addition, Table G2 and Figure G3 predict respondent ratings with respondent and country characteristics using measurement model-adjusted ratings from country experts instead of raw ratings that do not correct for DIF. Thus, when the results in Table G2 and G3 show few systematic correlates as in the main paper, we can conclude that there is little evidence for systematic biases also when accounting for DIF.

Figure G3. Predicting respondent perceptions with respondent and country characteristics (Pooled Model)



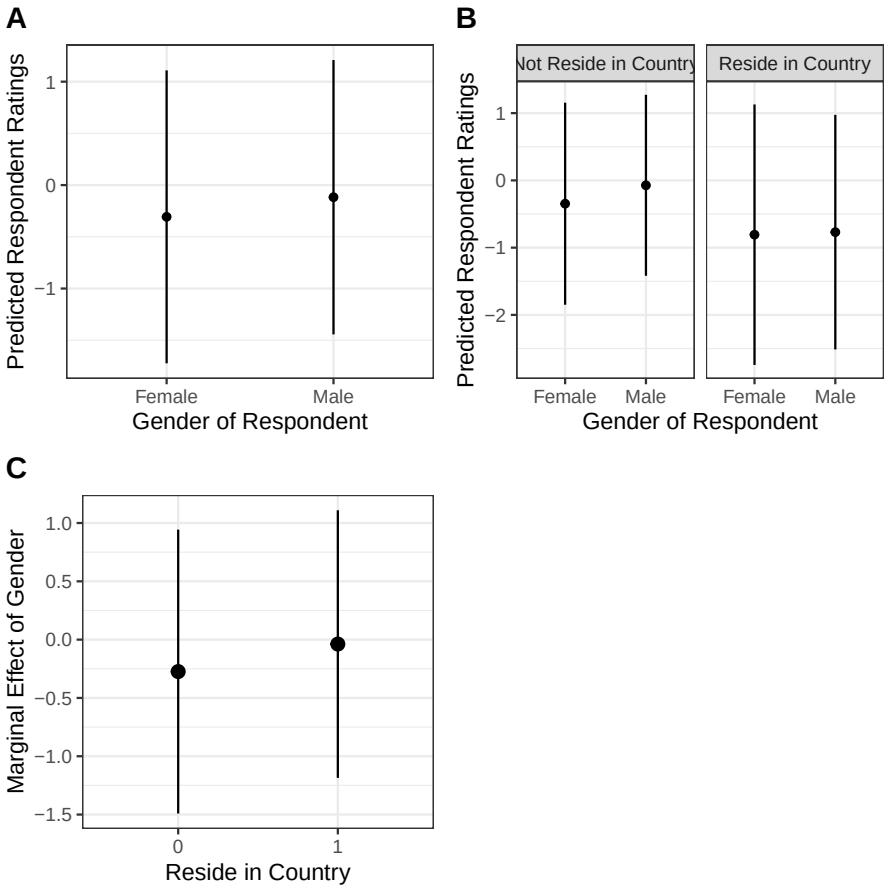
OLS regression with standard errors, clustered on countries. Measure-fixed effects, year-fixed effects are included in the model but omitted from the figure.

Figure G4. Predicted respondent ratings by Democratic Quality and First and Third Quartile 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.

Figure G5. Predicted respondent ratings by Respondent's Gender and Respondent's Reside/Born in Country



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

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	−5.444*	−3.325	−5.488*	0.475	−9.855*	−4.526*
	[−9.234; −1.653]	[−6.703; 0.053]	[−10.274; −0.702]	[−3.448; 4.399]	[−16.912; −2.797]	[−7.360; −1.692]
time spent for coding	0.021	0.030	0.019	0.027	0.016	0.022
	[−0.008; 0.050]	[−0.000; 0.059]	[−0.011; 0.050]	[−0.019; 0.072]	[−0.013; 0.044]	[−0.001; 0.045]
Respondent's Gender	−0.238	−0.221	0.168	0.009	−0.679	−0.190
	[−1.428; 0.953]	[−1.757; 1.316]	[−1.539; 1.875]	[−1.734; 1.753]	[−1.592; 0.234]	[−1.151; 0.771]
Respondent's Age: Middle Cohort	1.005	0.417	−0.051	0.223	1.405	0.531
	[−0.182; 2.192]	[−1.006; 1.840]	[−0.971; 0.870]	[−2.091; 2.537]	[−0.349; 3.158]	[−0.420; 1.481]
Respondent's Age: Oldest Cohort	1.187	0.629	0.992	0.444	0.407	0.718
	[−0.033; 2.407]	[−0.855; 2.112]	[−0.261; 2.245]	[−1.690; 2.578]	[−1.109; 1.923]	[−0.184; 1.620]
Respondent: PhD education	1.190	0.630	−0.325	0.522	−1.089	0.269
	[−0.135; 2.515]	[−0.942; 2.202]	[−1.745; 1.096]	[−1.057; 2.101]	[−2.498; 0.320]	[−0.705; 1.243]
Respondent: Government employed	−0.870	−0.096	0.333	−1.301	−0.476	−0.420
	[−2.254; 0.514]	[−2.196; 2.004]	[−2.491; 3.156]	[−2.663; 0.061]	[−2.096; 1.144]	[−1.626; 0.786]
Respondent: Reside in country	0.128	−1.308*	−0.862	−1.004	0.055	−0.639
	[−1.143; 1.399]	[−2.566; −0.050]	[−2.039; 0.314]	[−2.454; 0.445]	[−0.996; 1.106]	[−1.556; 0.279]
Respondent supports free markets	0.088	0.019	0.085	0.025	−0.051	0.052
	[−0.258; 0.434]	[−0.455; 0.493]	[−0.284; 0.455]	[−0.549; 0.598]	[−0.634; 0.532]	[−0.195; 0.298]
Respondent supports electoral democracy	−0.911	−0.932	−0.931	−1.234	2.193*	−0.319
	[−2.478; 0.655]	[−2.060; 0.196]	[−2.455; 0.593]	[−2.587; 0.120]	[0.326; 4.059]	[−1.229; 0.592]
Respondent supports liberal democracy	1.006	0.728*	1.316	0.141	−0.215	0.596
	[−0.642; 2.655]	[0.030; 1.427]	[−0.456; 3.089]	[−0.897; 1.179]	[−1.209; 0.779]	[−0.139; 1.331]
Country liberal component	4.175	3.039	8.677	−0.694	1.065	3.207
	[−0.703; 9.054]	[−0.620; 6.698]	[−0.763; 18.118]	[−7.944; 6.555]	[−5.670; 7.801]	[−0.580; 6.994]
Respondent supports liberal democracy * LibDem	1.088	1.047	−0.597	−0.156	−0.363	0.232
	[−0.196; 2.371]	[−0.179; 2.274]	[−2.308; 1.114]	[−1.900; 1.588]	[−1.378; 0.653]	[−0.803; 1.268]
Respondent supports electoral democracy * EDI	−1.105	−0.479	−1.413	0.275	0.353	−0.479
	[−2.979; 0.770]	[−1.388; 0.431]	[−3.737; 0.912]	[−1.092; 1.641]	[−0.926; 1.631]	[−1.447; 0.489]
Country electoral democracy level	−0.811	0.943	−0.538	−1.808	13.056*	2.807
	[−7.442; 5.820]	[−5.672; 7.557]	[−8.036; 6.960]	[−10.891; 7.276]	[0.392; 25.721]	[−2.112; 7.726]
Latin America and the Caribbean	0.091	0.749	−0.958	−1.018	−0.167	−0.277
	[−0.902; 1.085]	[−0.875; 2.373]	[−3.258; 1.342]	[−3.246; 1.210]	[−0.948; 0.614]	[−1.052; 0.499]
MENA	3.110	3.898*	1.465	2.389	−0.691	2.173
	[−0.291; 6.510]	[0.476; 7.321]	[−1.026; 3.957]	[−1.591; 6.369]	[−2.257; 0.875]	[−0.270; 4.616]
SSA	1.603	1.448	0.330	0.224	0.108	0.818
	[−0.819; 4.025]	[−0.218; 3.114]	[−1.052; 1.713]	[−2.511; 2.959]	[−1.851; 2.067]	[−0.256; 1.892]
Western Europe and North America	−0.361	−0.020	−1.067	−0.138	−0.251	0.542
	[−0.496; 1.581]	[−2.490; 1.768]	[−1.238; 1.198]	[−3.592; 1.458]	[−1.027; 0.751]	[−1.169; 0.666]
Asia and Pacific	0.171	0.675	−0.712	−0.312	−0.516	−0.043
	[−1.323; 1.664]	[−0.948; 2.299]	[−1.772; 0.348]	[−2.669; 2.045]	[−1.450; 0.419]	[−0.853; 0.766]
GDP pc	−0.018	0.007	−0.022	0.014	0.020	−0.001
	[−0.054; 0.018]	[−0.041; 0.055]	[−0.055; 0.011]	[−0.028; 0.057]	[−0.010; 0.050]	[−0.029; 0.027]
Dummy Freedom to research and teach						−0.220

	Freedom to research and teach	Freedom of academic exchange and dissemination	Institutional autonomy	Campus integrity	Freedom of academic and cultural expression	Pooled Model
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Dummy Institutional Autonomy						[−0.805; 0.365]
						−1.201*
						[−1.804; −0.598]
Dummy Campus Integrity						−0.145
						[−0.769; 0.479]
Freedom of academic and cultural expression						−1.263*
						[−2.025; −0.500]
R^2	0.027	0.039	0.046	0.025	0.059	0.030
Adj. R^2	0.025	0.038	0.044	0.023	0.057	0.030
Num. obs.	80668	80162	80049	79457	56956	377292
RMSE	11.986	11.629	10.680	12.979	10.643	11.712
N Clusters	175	175	175	175	177	178

* Null hypothesis value outside the confidence interval.

Table G2. Linear Models predicting respondents rating with country characteristics - Perceptions

H Analyzing Anchor Vignettes

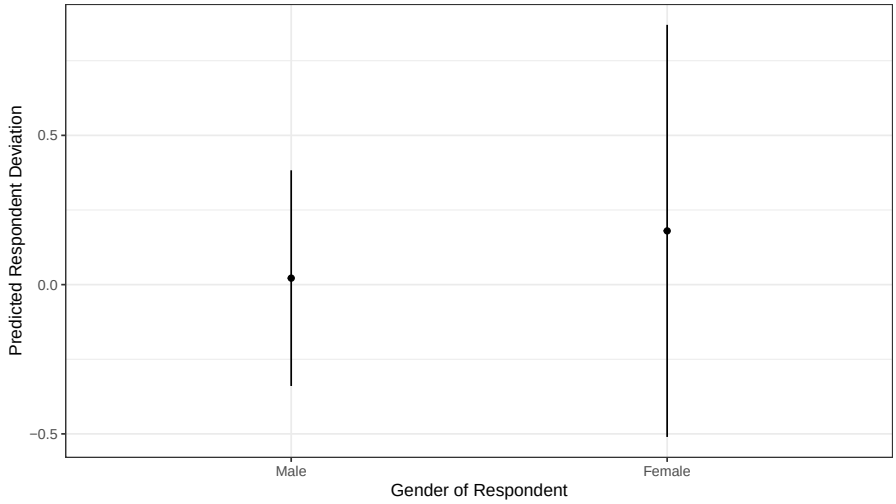
In this section, we analyze the deviations of respondents' measurement model-adjusted ratings from the mean rating among anchor vignettes that are description of hypothetical cases that were rated by experts. By analyzing these deviations in a pooled models, we are able to identify systematic predictors of respondent's coding behavior independently from errors that come from different country background characteristics.

	Model 1	Model 2
Intercept	0.136	0.895
	[−0.237; 0.509]	[−0.440; 2.231]
Women	0.158	0.090
	[−0.413; 0.730]	[−0.509; 0.689]
Age Block 2	−0.044	−0.045
	[−0.581; 0.493]	[−0.613; 0.523]
Age Block 3	0.111	0.166
	[−0.486; 0.707]	[−0.447; 0.778]
PhD degree	−0.230	−0.395
	[−0.671; 0.212]	[−0.831; 0.041]
Goverment employee	0.313	0.345
	[−0.913; 1.540]	[−0.967; 1.657]
v2cafres	−0.007	−0.011
	[−0.055; 0.040]	[−0.059; 0.036]
v2cainsaut	−0.032*	−0.033*
	[−0.056; −0.008]	[−0.063; −0.002]
v2casurv	−0.002	−0.008
	[−0.066; 0.063]	[−0.076; 0.059]
v2clacfree	−0.115*	0.029
	[−0.183; −0.046]	[−0.030; 0.087]
Time spent for coding		−0.003
		[−0.015; 0.008]
Satisfaction with coding experience		−0.156
		[−0.457; 0.145]
R^2	0.000	0.001
Adj. R^2	−0.001	−0.001
Num. obs.	5711	5691
RMSE	8.964	8.910
N Clusters	25	25

* Null hypothesis value outside the confidence interval.

Table H1. Linear Models predicting respondents deviations from mean

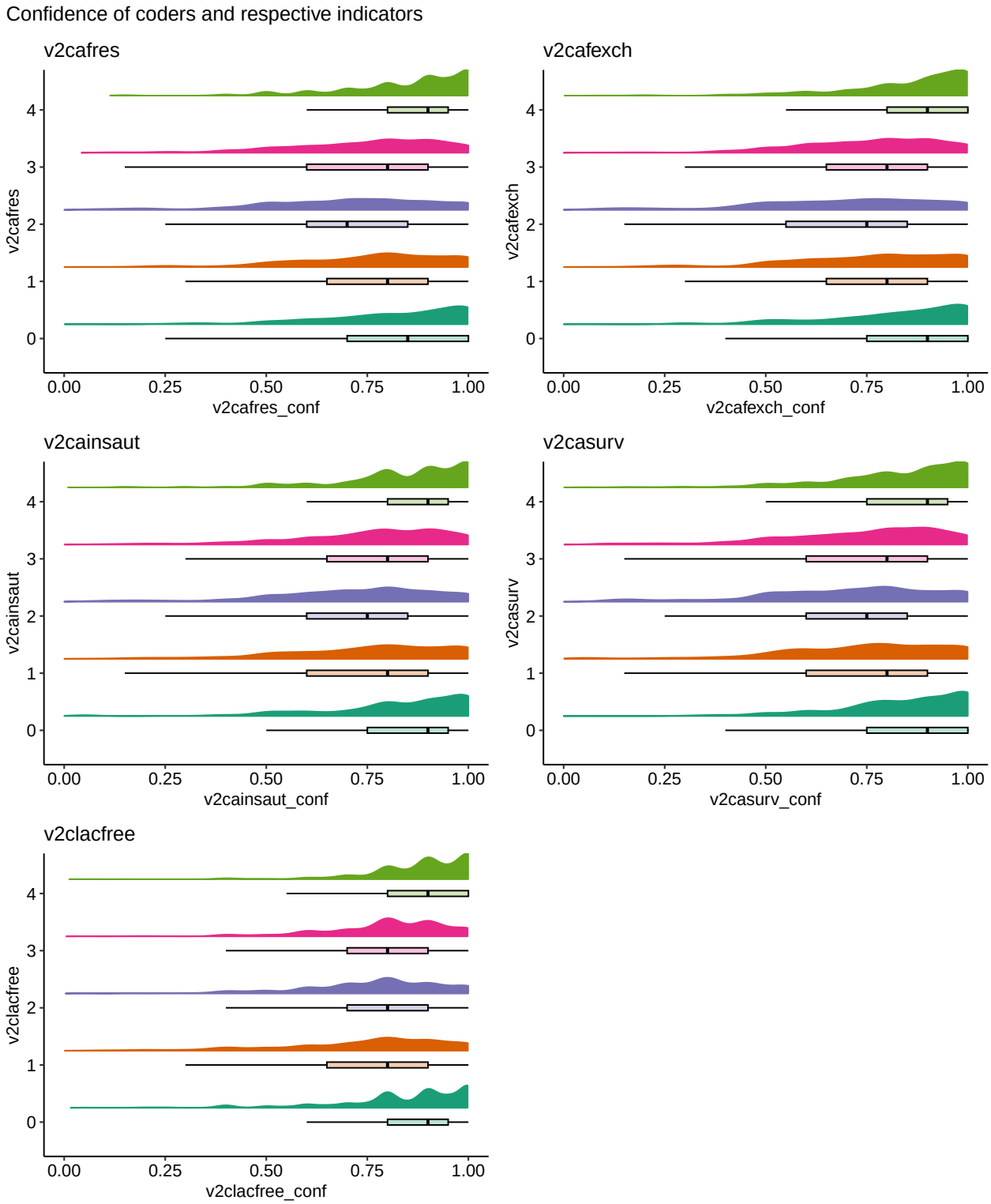
Figure H1. Predicting respondent deviations with respondent characteristics (Pooled Model)



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

I Distribution of Coder Confidence

Figure I1. Confidence of Coders across AFI indicators



For each indicator and its respective ordinal categories, there is a density plot and a boxplot to show the distribution.