Appendix:
Bureaucratic Representation and Gender Mainstreaming in International Organizations: Evidence from the World Bank

Table of Contents

1. Interview participants and topics

Catherine Weaver conducted interviews with key members of the World Bank staff in person in February 2007 in Washington, DC and via Zoom between February to March 2022. The interviews were semi-structured. Because the interview format was open-ended, the exact wording, order and content of the line of questioning varied in each interview. Below are the general objectives and "prompts" used for each meeting.

Interviews conducted in 2007 targeted key staff members who worked in the World Bank in the following capacities: as members of the Gender and Development Network, as Gender Focal Points in several of the Bank's sectoral (e.g.Agriculture, Education, Health) and regional operational units (e.g. Europe and Central Asia, Middle East and North Africa; the World Bank Staff Association, the research arm of the Bank (DECRG) country directors and one Vice President. All respondents provided consent with guarantee of indirect attribution (untraceable position title) only. Interviews were not recorded; handwritten notes were transcribed for later analysis. Interviews lasted on average about one hour.

Interviews conducted in 2022 were randomly selected from a list of respondents who completed our survey. They were conducted with World Bank task team leaders (and one DRG researcher) by Catherine Weaver via Zoom. All respondents provided consent with guarantee of indirect attribution (untraceable position title) only. We have indicated current or last known Global Practice affiliation only to ensure anonymity. Interviews were recorded and transcribed, enabling direct quotations where appropriate. Country placements not indicated to protect anonymity. The objectives of these interviews were broader, insofar as they were meant to solicit information regarding a large number of questions pertaining to staff hiring and promotion practices, the influence of organizational culture on staff operational behavior, and the influence of politics (from donor states and client governments) on staff decision-making. A list of the topics guiding the open-ended questions pertaining is provided below.

Table A1: Interview participants

Interview	Participant
2007A	Gender and Development unit within the Poverty Reduction and Economic Management
2007B	Gender and Development unit within the Poverty Reduction and Economic Management
2007C	Gender and Development unit within the Poverty Reduction and Economic Management
2007D	Gender and Development unit within the Poverty Reduction and Economic Management
2007E	Gender and Development unit within the Poverty Reduction and Economic Management
2007F	Gender and Development unit within the Poverty Reduction and Economic Management
2007G	Gender and Development unit within the Poverty Reduction and Economic Management
2007H	Gender and Development unit within the Poverty Reduction and Economic Management
2007I	Gender and Development unit within the Poverty Reduction and Economic Management
2007J	Gender and Development unit within the Poverty Reduction and Economic Management
2007K	Gender Focal Points in: Latin America and Caribbean, Europe and Central Asia, and Middle East and North Africa
2007L	Gender Focal Points in: Latin America and Caribbean, Europe and Central Asia, and Middle East and North Africa
2007M	Gender Focal Points in: Latin America and Caribbean, Europe and Central Asia, and Middle East and North Africa
2007N	Retired Bank staff
2007O	Retired Bank staff
2007P	Retired Bank staff
2007Q	Economist, Poverty Reduction and Economic Management
2007R	Economist, Poverty Reduction and Economic Management
2007S	NGO staff (external to World Bank)
2007T	NGO staff (external to World Bank)
2007U	World Bank Staff Association
2007V	World Bank Staff Association

2007W	World Bank Vice President
2022A	Office of the Chief Economist for Global Practices
2022B	Human Development Global Practice (retired)
2022C	Macroeconomics, Trade and Investment Global Practice
2022D	Macroeconomics, Trade and Investment Global Practice
2022E	Poverty and Equity Global Practie
2022F	Country Program Coordinator (retired)
2022G	Human Development Global Practice
2022H	Governance Global Practice
2022I	Social Protection Global Practice
2022J	Development Research Group

Interview topics

2007 Interviews

- Staff members' involvement in the creation, dissemination/training or implementation of gender mainstreaming within other organizations besides the World Bank;
- Staff members' involvement in the creation, dissemination/training or implementation of the World Bank's Gender Mainstreaming Strategy;
- Staff members' perceptions of factors facilitating or supporting gender mainstreaming, rooted in organizational culture, hierarchical structures, staff composition, management norms, organizational mandates and operating policies, Executive Board support for gender work, and client government attitudes about gender equity and gender mainstreaming;
- Tactics used by gender advocates within the Bank to gain traction in building internal support to draft a gender mainstreaming strategy and accompanying changes in operational policy;
- Procedures established to monitor and evaluation gender mainstreaming progress and to ensure accountability of Bank management and staff to gender mainstreaming goals;
- Interviewee's perspectives on the differences between the Bank's discourse and practice with respect to gender mainstreaming and the effects of these differences on the nature of the Bank's gender work in the field.

2022 Interviews

- Hiring and promotional norms and practices in the World Bank with respect to staff nationality, expertise, educational backgrounds, gender, and prior professional experience;
- Organizational culture norms and attitudes regarding what staff traits and experiences are needed to sustain good relationships with donor and borrower states;
- Perceptions of internal and external pressures to address representation across dimensions of gender, nationality, disciplinary training and expertise;
- Organizational culture norms and attitudes regarding what staff traits and experiences are needed to carry out the technical missions of the World Bank;
- Personal opinions on which staff traits with respect to gender, professional experience, nationality, educational training, learned skills or lived experiences are most important for securing and sustaining partner country trust and cooperation in operations;
- In-country experiences of doing work and pushing the World Bank's official agenda around sensitive topics, such as gender or health (HIV);
- Personal opinions on the World Bank's progress in reaching representational goals;

- Personal opinions on the opportunities and challenges in reaching representational goals in the World Bank.

2. Ethical practices concerning human participants

Interviews:
The Interviews were conducted in two waves. They were approved by the Institutional Review Boards of the the University of Kansas in 2007 and the University of Texas at Austin in 2021 (00001310). Interviews were conducted using a consent form that provided the respondent with a choice of (1) full attribution of comments with full name and title, (2) anonymized attribution using a position title that would not allow for easy identification of the respondent, and (3) no attribution/background information only. Interviews were conducted in person at the World Bank (2007) and online (2022). They averaged approximately 60 minutes each. Interviews were not recorded but one of the authors took detailed notes that were anonymized and saved in an encrypted file.

Survey experiment:
The survey experiment was approved by the Institutional Review Board of the University of Texas at Austin under study number: 00001310. In the following, we briefly discuss consent, deception, confidentiality, harm, impact, and compensation. No data from the survey was merged with any of the individual-level information collected for the observational analyses. The discussion is an abbreviated version of our IRB.

Consent:
The first page of the survey included the following consent form:

Welcome to our study on World Bank Task Teak Leaders and Project Management!

To better understand the important role of Task Team leaders in the World Bank, we are interested in grasping differences between Bank TTLs in their opinions on different elements of a project. This survey is structured to help us discern which factors you believe matter in getting projects approved and successfully implemented.

The study should take you around 15 minutes to complete. Your responses will be kept completely confidential, and your participation in this research is voluntary. You have the right to withdraw at any point during the study. The Principal Investigator of this study can be contacted at *email redacted for peer-review.*

By clicking the button below, you acknowledge:
Your participation in the study is voluntary.
You are 18 years of age.
You are aware that you may choose to terminate your participation at any time for any reason

Deception:
The survey included no deception. They were invited to participate in a survey on the "opinions of World Bank Task Team Leaders". We did not inform them directly of our interest in representation since we did not want to prime answers. Instead, we relied on the less controversial framing of "differences between TTLs" instead of bureaucratic representation. Specifically, we informed respondents that we were interested in understanding "how differences between TTLs shape their personal opinion on different World Bank projects". No deception was used in the experimental condition. Recipient knew they were rating hypothetical project profiles and that project elements were varied randomly. Specifically, they read the following statement:

> You will see two profiles of hypothetical World Bank Development Policy Loans (DPLs). These DPLs vary on five project-level design features and three country characteristics. Please share your opinion for both hypothetical projects:
>
> a) how likely a project will be approved in the executive board
> b) how likely a project will make a positive impact in the recipient country
>
> We ask you to do this for 7 sets of projects.

Confidentiality:
Data was coded numerically by assigning a unique ID to each respondent. Respondents cannot be personally identified.

Harm and Impact:
The research presents no more than minimal risk of harm to respondents. Respondents are not identifiable from the survey, they were explicitly asked to only give their personal opinion and they were not asked to disclose any information that could be seen as privileged information.

Compensation:
Respondents did not receive a direct monetary compensation. However, they were compensated in two ways for their engagement with the survey: First, all respondents could indicate that they would like to receive a short report on the findings from the survey. We sent the reports on the 23rd of May 2022 to all respondents that expressed interest in these reports. Second, we donated 10 USD on behalf of each respondent who filled out the full survey to the charity GiveDirectly on 6th of May 2022.

3. Pre-registration

The survey experiment was preregistered at https://doi.org/10.17605/OSF.IO/6J3M8 on the 8th of March 2022. None of the specification choices described in the preregistration report were altered.

Survey design
We conducted a conjoint survey experiment to grasp how different backgrounds of TTLs affect project decisions. Conjoint experiments originated from marketing research and were used to estimate the impact of product qualities on the decision to purchase a good (Green, Krieger, and Wind 2001). More recently, conjoint analysis has been used to understand the design of international treaties (Bechtel and Scheve 2013) and World Bank projects (Briggs 2021). In conjoint experiments,

respondents will be asked to evaluate profiles with randomized features. These features randomly take different levels for each profile. The conjoint experiment then enables quantifying the relative importance of these different features for the choices respondents make (Hainmueller, Hopkins, and Yamamoto 2014). There are three potential designs of conjoint analysis: first, the single profile conjoint experiment confronts respondents with one profile and asks them to evaluate this profile. Second, the paired conjoint experiment shows two profiles and requests evaluations on both profiles. Third, the paired conjoint experiment with forced-choice shows two profiles and asks respondents to choose between the two profiles (Hainmueller, Hangartner, and Yamamoto 2015).

We depart from extant approaches by using a paired conjoint experiment without a forced choice. Existing research on TTL decision-making has utilized paired conjoint experiments with a forced choice (Briggs 2021). The same approach is prominent in international relations research more generally (Bechtel and Scheve 2013; Ghassim, Koenig-Archibugi, and Cabrera 2022). In contrast, we employ the simple paired conjoint experiment because it fits more closely with the situation faced by TTLs when deciding on project design. The job of TTL rarely includes trade-offs between two projects. Instead, TTLs are intricately involved in project design and implementation (Heinzel and Liese 2021). Their decisions are closer to evaluating potential projects and deciding how project design would need to change to ensure approval and implementation. Therefore, we refrained from forcing respondents to make a choice. However, we still included two profiles because research on conjoint experiments shows that external validity is strongest when respondents are evaluating profiles side-by-side, because the paired choice increases respondent engagement. Therefore, the likelihood of nondifferentiation (giving the same answer to various similar questions) and acquiescence response bias (tendency to agree regardless of content) decrease (Hainmueller, Hangartner, and Yamamoto 2015).

The two profiles that respondents see were described as two hypothetical World Bank DPLs. They included eight features with two levels each. The eight features focused on key project design and country environment factors that may impact decisions. The levels were independently randomized. Table 1 lists the features and their levels.

Table A2: features and levels of project profiles in the conjoint experiment

Feature	**Level 1**	**Level 2**
Project design:		
Number of prior action conditions	16	4
Project amount	Above average	Below average
Environmental and social risk	High	Low
Macroeconomic risk	High	Low
Gender-disaggregated targets	**Yes**	**No**
Country environment:		
Recipient country CPIA score	4	2
US has opposed last project of recipient country	Yes	No
Recipient country income level	MIC	LIC

The features and levels are designed as close to the World Bank's language as possible to ensure that respondents interpret the features and levels in the same way as we do.

First, we use data on the design of DPLs from the World Bank's Development Policy Action Database to identify levels corresponding to tight or loose conditionality (World Bank 2021). We define tight

conditionality as two standard deviations above the mean (rounded 16 conditions) and loose conditionality as two standard deviations below the mean (rounded 4 conditions). Second, we refrain from using exact monetary values for the levels of project amounts because TTLs working in different sectors work with considerably different budgets. For example, infrastructure projects tend to be larger than education projects (Zeitz 2021). Therefore, we only ask for above or below average project amounts (Briggs 2021). Third, for environmental and social risk we draw on the risk matrices included in DPLs. The project approval documents specify risk on nine dimensions in a categorical coding scheme as High, Substantial, Moderate, and Low. Fourth, the same is true for macroeconomic risk, which is another component of risk matrices. Fifth, we ask for gender-disaggregated project targets, which is a key project design component that TTLs are asked to include based on gender mainstreaming norms (Winters et al. 2018). Sixth, TTLs typically discuss the CPIA score of the recipient country as part of the project approval documents. CPIA scores are only publicly available for IDA countries. We draw on the IDAs resource allocation index, which is an aggregation of the sub-component of the CPIA scores used to calculate how much money a given IDA country should attain. We choose CPIA scores aligning with the highest group (4) and lowest group (2) of countries as levels for the variable. Seventh, the feature focusing on US opposition highlights whether the US has opposed the last loan by the recipient country. It is a simple binary indicating US opposition. Finally, we use World Bank income classification to indicate whether a country is a middle-income-country (MIC) or low-income-country (LIC).

After seeing the two profiles, respondents were asked to rate each project on two dimensions: a) how likely it is to get board approval b) how likely it is to have a positive impact on development outcomes in the recipient country. The DVs are measured on a scale from 1 (least likely) to 10 (most likely).

Hypotheses

The paper includes tests of Hypotheses 5 and 14 in the preregistration document. The other hypotheses do not focus on questions of gender mainstreaming or gender representation and are, therefore, not included in this paper. The paper included one hypothesis for each of the eight conjoint profiles (H1-H8) and six conditional hypotheses focusing on different elements of bureaucratic representation (gender, education, nationality). The results on the hypotheses not focusing on gender will be reported elsewhere. The full list of hypotheses was:

H1: More conditionality increases the likelihood of approval and development impact
H2: Higher project amounts decrease the likelihood of approval and increase the likelihood of development impact
H3: Higher environmental and social risk decreases the likelihood of approval and development impact
H4: Higher Macroeconomic risk decreases the likelihood of approval and development impact
H5: Having gender-disaggregated targets increase the likelihood of approval and development impact
H6: Higher CPIA scores increase the likelihood of approval and development impact
H7: Past US opposition decreases the likelihood of approval and does not affect development impact
H8: Being a Middle-income country decreases the likelihood of approval and development impact
H9: The relationship between US opposition and board approval becomes weaker when the project includes more conditions
H10: US nationals evaluate the likelihood of approval and development impact lower than non-US nationals if the US has opposed recipients' projects in the past
H11: Economists evaluate the likelihood of development impact higher than non-Economists if the CPIA rating is high

H12: Economists evaluate the likelihood of development impact lower than non-Economists if there is more macroeconomic risk
H13: Non-Economists evaluate the likelihood of development impact higher than Economists if there is more environmental and social risk
H14: Women evaluate the likelihood of development impact higher than Men if there are gender-disaggregated targets

We pre-registered the following analysis:

"Our sample will include all observations that we received in the questionnaire. Our main analysis will be based on calculating the marginal mean for each level. We estimate the marginal mean to mitigate concerns about differences in reference categories when comparing results of conjoint analyses between sub-samples (Leeper, Hobolt, and Tilley 2020). We will calculate 95% confidence interval for each marginal mean and differences in marginal means with 95% confidence intervals to grasp sub-group differences. Standard errors will be clustered by respondent. We will apply post-stratification weights to weight responses by the region and sector that respondents indicated as their most recent area of work.
H1-H8 will be tested by presenting marginal means for the entire sample. We further compare the following groups of respondents to understand how passive representation of US nationals, economists and women in World Bank staff shapes differences in DPL design:
- US nationals vs. other nationals (H10)
- Economists vs. other educational backgrounds (H11-H13)
- Women vs. Men (H14)
The sub-group analysis allows us to systematically assess experimental conditions in the sub-group. Nevertheless, it remains observational in one key respect because we cannot manipulate respondents' backgrounds experimentally."

4. Data sources and descriptive statistics

In this section, we display several descriptive statistics to give more information on our main variables. Table A3 contains the descriptive statistics on the main variables used in the article. Table A4 displays the variable labels, explains what these variables are measuring, and indicates the sources of the data used.

Table A3: Descriptive statistics on variables

Variable name	N	Mean	SD	Min	Max
Gender mainstreaming (Any MS)	2076	0.828	0.377	0.000	1.000
Gender mainstreaming (Deep MS)	2076	1.986	1.141	0.000	3.000
TTL Women	2076	0.539	0.499	0.000	1.000
TTL Gender expertise	2076	0.478	1.159	0.000	13.000
CD women	2076	0.307	0.461	0.000	1.000
PM women	2076	0.366	0.482	0.000	1.000
Gender project	2076	0.632	3.460	0.000	50.000
IDA	2076	0.627	0.484	0.000	1.000
Amount (log)	2076	17.965	1.314	0.000	21.416

Post-conflict country	2076	0.175	0.380	0.000	1.000
Women ministers	2076	0.158	0.101	0.000	0.538
Principals gender lending	2076	0.408	0.158	0.002	0.906
Women economic rights	2076	0.884	0.617	0.000	3.000
Women infant mortality	2076	32.263	20.600	2.500	104.300
Women vulnerable employment	2076	63.513	25.181	1.570	98.570
GDP per capita	2076	3093.851	3193.320	219.961	14741.192
Population (log)	2076	17.268	1.846	12.810	21.044
Women appointed in sector (within 3 years)	2076	0.310	0.095	0.000	0.505
Gender mainstreaming in sector (within 3 years)	2076	1.816	0.682	0.000	3.000
Gender mainstreaming in country (within 3 years)	2029	1.851	0.723	0.000	3.000

Table A4: Data sources and explanations

Variable	Explanation	Source
Gender mainstreaming (Any MS)	Rating of gender mainstreaming in analysis, actions, monitoring or evaluation	World Bank (2018)
Gender mainstreaming (Deep MS)	Rating of gender mainstreaming in analysis, actions, monitoring and evaluation	World Bank (2018)
TTL Women	Binary variable indicating whether one of the recorded TTLs was a woman	Authors calculation
TTL Gender expertise	Count of the number of gender-themed projects TTL has supervised in the past	Authors calculation
CD women	Binary variable indicating whether country director was a woman	Authors calculation
PM women	Binary variable indicating whether practice manager was a woman	Authors calculation
Gender project	Percentage of project focused on Gender theme	World Bank (2020a)
IDA	Binary variable indicating whether IDA contributed funds to project	World Bank (2020a)
Amount (log)	Log (+1) of amount of contributed funds to project by IDA and IBRD	World Bank (2020a)
Post-conflict country project	Binary variable indicating whether country was on World Bank list of post-conflict countries (at approval)	World Bank (2018)
Women ministers	Share of women ministers in cabinet of recipient country (at approval)	Nyrup and Bramwell (2020)
Women economic rights	Economic rights of women in recipient country (at approval)	Cingranelli and Richards (2010)
Women infant mortality	Infant mortality of girls in recipient country (at approval)	World Bank (2020b)
Women vulnerable employment	Percent of women in vulnerable employment in recipient country (at approval)	World Bank (2020b)
Principals gender lending	Mean share of funds allocated by top 5 DAC shareholders (USA, UK, Germany, France, and Japan) to recipient country that had a gender marker (at approval)	OECD (2024)
GDP per capita	GDP per capita of recipient country (at approval)	World Bank (2020b)
Population (log)	Log of recipient countries population (at approval)	World Bank (2020b)

TTL count (project)	Count of the number of TTLs listed on project in database (at approval, used as instrument for women TTLs)	Authors calculation
Women appointed in sector (within 3 years)	The share of women TTLs appointed in each sector in all project approved within 3 years of the project of interest	Authors calculation
Gender mainstreaming in sector (within 3 years)	The average gender mainstreaming rating in each sector in all project approved within 3 years of the project of interest	Authors calculation based on World Bank (2018)
Gender mainstreaming in country (within 3 years)	The average gender mainstreaming rating in each country in all project approved within 3 years of the project of interest	Authors calculation based on World Bank (2018)

Information on the dependent variable

We also discuss the composition of our main dependent variable, the GMI, in more detail. The GMI is an additive index of three components: analysis, actions, as well as monitoring and evaluation. We display the temporal differences in the inclusion of the three components in Figure A1 and disaggregate the different levels of dependent variable in Table A5. Figure A1 shows that inclusion of the three GMI components is somewhat similar and increases over the period under investigation.

Figure A1: Different GMI components over time

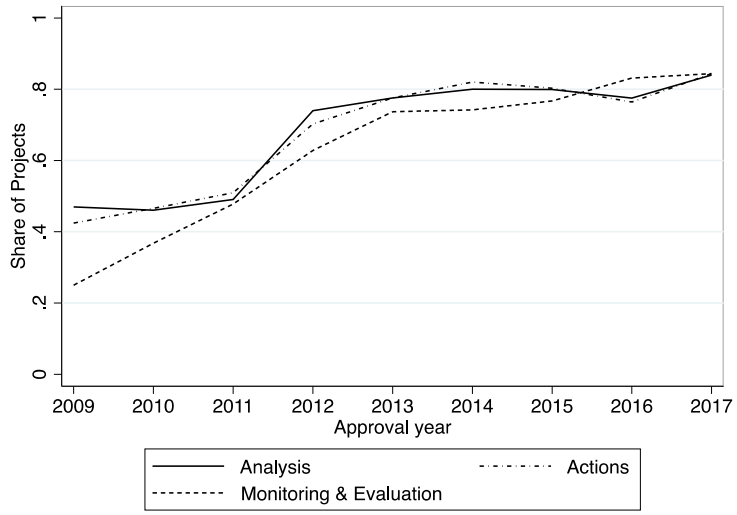


Table A1 shows that differences between components appear in projects with shallow mainstreaming. 43.1% of these projects include analysis components, 25.8% include gender mainstreaming in actions and 31.2% include gender mainstreaming in monitoring and evaluation.

Table A5: Descriptive statistics on dependent variable

	GMI = 1		**GMI = 2**		**GMI = 3**	
	Percentage	**Observations**	**Percentage**	**Observations**	**Percentage**	**Observations**
Analysis	43.1	152	68.0	363	100	1,216
Actions	25.8	91	79.0	422	100	1,216
Monitoring & Evaluation	31.2	110	53.0	283	100	1,216
Total		353		534		1,216

We further show differences in GMI by sector (Figure A2), region (Figure A3), lending instrument (Figure A4), and project volume (Figure A5). Figure A2 shows that gender mainstreaming is deepest in Agriculture, Education, Health, and Social Protection and most shallow in the Energy, Financial and Public Administration sectors. Figure A3 reveals that the mean GMI is lowest in Europe and Central Asia and highest in Middle East and North Africa. Figure A4 illustrates that Program for Results shows the deepest gender mainstreaming, while Development Policy Loans include the shallowest gender mainstreaming. Finally, Figure A5 shows that projects with a GMI have somewhat smaller loan volumes, on average.

Figure A2: GMI by sector

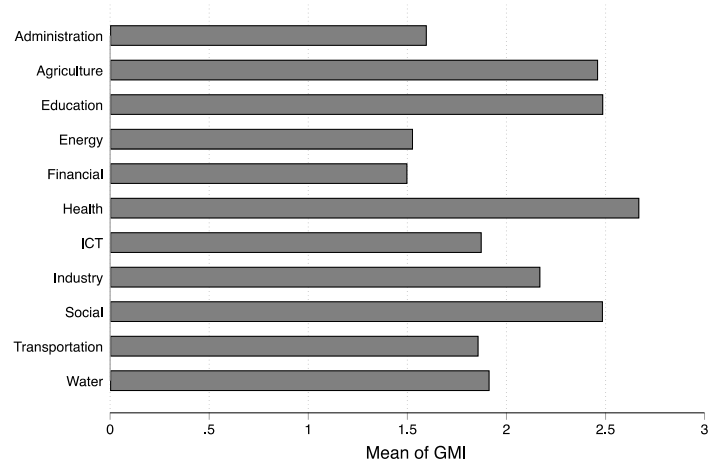


Figure A3: GMI by region

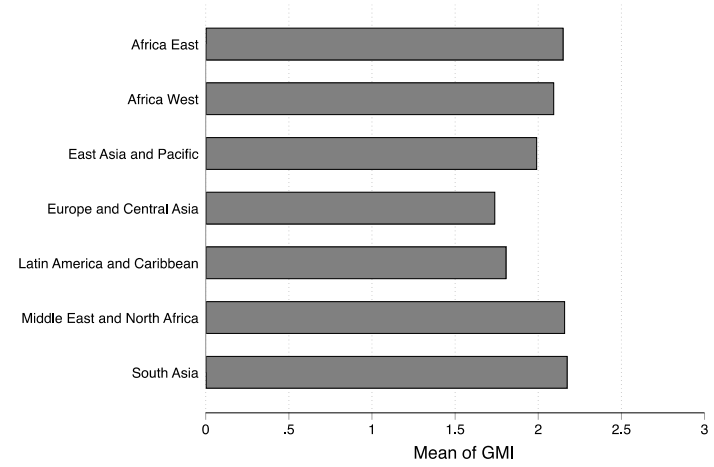


Figure A4: GMI by lending instrument

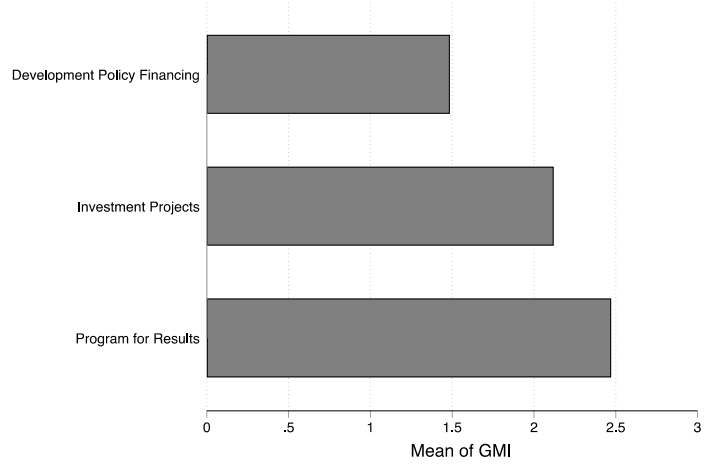
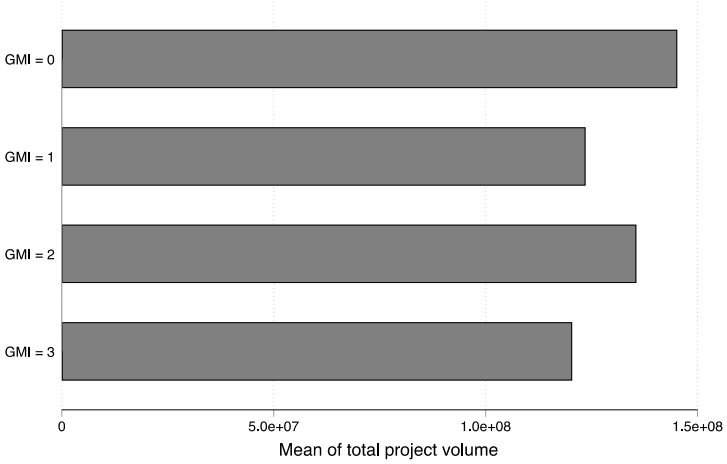


Figure A5: Loan volume by GMI



5. Robustness checks for observational analysis

First-stage regression models

We display the first-stage OLS regression for the instrumental variable models presented in the main article. These models predict the likelihood that at least one TTL is a woman in a given project. The model shows that our instrument—the count of TTLs appointed to the project—is a strong instrument for the presence of women in a project. Each additional appointed TTL increases the likelihood that at least one women TTL is appointed by around 15% ($p<0.001$).

Table A6: First stage models

	(1) DV: TTL Women
TTL count	0.1537***
	(0.0146)
TTL Gender expertise	0.0056
	(0.0121)
CD Women	0.0141
	(0.0356)
PM Women	-0.0378
	(0.0236)
Gender project	0.0025
	(0.0028)
IDA	0.0375
	(0.0483)
Amount (log)	-0.0026
	(0.0131)
Post-conflict country	-0.0613
	(0.0775)
Women ministers	0.2397
	(0.2105)
Principals gender lending	-0.0914
	(0.1080)
Women economic rights	-0.0130
	(0.0362)
Women infant mortality	0.0067
	(0.0038)
Women vulnerable employment	-0.0022
	(0.0058)
GDP per capita	0.0000
	(0.0000)
Population (log)	0.9838
	(0.5268)
Country fixed effects	Yes
Sector-year fixed effects	Yes
Observations	2076
R^2	0.237

Country-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Alternative estimation approaches

We opted for Ordinary-Least-Squares regressions in the main models presented in the article. However, the Deep MS variables are a count of the number of gender mainstreaming features in a given project. Hence, we re-estimate the models using an approach more tailored to count data. Specifically, we employ Poisson-Pseudo-Maximum-Likelihood models that perform better than traditional negative binomial or Poisson regressions when including a substantial number of fixed effects. The results are consistent with the OLS results presented in the main body of the article.

Table A7: Poisson-pseudo-likelihood models

	(2) Deep MS	(3) Deep MS
TTL Women	0.0570*	0.0593*
	(0.0249)	(0.0245)
Women in sector (last 3 years)		0.5469*
		(0.2614)
TTL Gender expertise	0.0227**	0.0260***
	(0.0080)	(0.0075)
CD Women	0.0890*	0.1105**
	(0.0381)	(0.0397)
PM Women	-0.0332	-0.0209
	(0.0248)	(0.0257)
Gender project	0.0107***	0.0118***
	(0.0027)	(0.0025)
IDA	0.1594*	0.1854**
	(0.0635)	(0.0591)
Amount (log)	0.0278	0.0266
	(0.0172)	(0.0172)
Post-conflict country	-0.2303*	-0.2291*
	(0.1056)	(0.1047)
Women ministers	0.0359	-0.2271
	(0.2433)	(0.2354)
Principals gender lending	0.1460	0.1059
	(0.1460)	(0.1390)
Women economic rights	0.0471	0.0635
	(0.0342)	(0.0332)
Women infant mortality	0.0099	0.0096
	(0.0088)	(0.0080)
Women vulnerable employment	-0.0007	-0.0010
	(0.0075)	(0.0078)
GDP per capita	0.0002*	0.0002*
	(0.0001)	(0.0001)
Population (log)	-1.5859	-1.9654*
	(0.8851)	(0.9635)
Gender mainstreaming in sector (last 3 years)		0.7602***
		(0.1045)
Gender mainstreaming in country (last 3 years)		0.3533**
		(0.1079)
Interaction (last 3 years)		-0.2041***
		(0.0470)
Country fixed effects	Yes	Yes
Sector fixed effects	No	Yes
Year fixed effects	No	Yes
Sector-year fixed effects	Yes	No
Observations	2074	2027
Pseudo R^2	0.105	0.094

Country-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Similarly, we re-estimate models using maximum likelihood. Table A7 displays the results we attain when employing logistic regression (Any MS) and ordered logit regression (Deep MS). Again, our results are substantively similar. Women TTLs do not appear to design projects with more mainstreaming in general (Model 4), but gender differences appear when differentiating between shallower and deeper mainstreaming (Model 5).

Table A8: Non-linear models

	(4) Any MS	(5) Deep MS
Estimation approach	Logit	Ologit
TTL Women	0.0061	0.2654*
	(0.2173)	(0.1206)
TTL Gender expertise	0.3718	0.3045***
	(0.2569)	(0.0867)
CD Women	0.7208	0.3703*
	(0.3755)	(0.1839)
PM Women	-0.2886	-0.1381
	(0.2096)	(0.1177)
Gender project	0.0000	0.1905***
	(.)	(0.0517)
IDA	1.6467***	0.4957
	(0.4170)	(0.2737)
Amount (log)	0.1452	0.1385
	(0.1209)	(0.0785)
Post-conflict country	-2.0318*	-0.9793*
	(0.7890)	(0.4666)
Women ministers	-1.1819	-0.3547
	(1.9981)	(1.0237)
Principals gender lending	0.8597	0.7570
	(0.9951)	(0.6537)
Women economic rights	0.4825	0.2156
	(0.4388)	(0.1839)
Women infant mortality	0.0000	0.0055
	(0.0243)	(0.0298)
Women vulnerable employment	-0.0203	-0.0081
	(0.0462)	(0.0272)
GDP per capita	-0.0001	0.0004
	(0.0004)	(0.0002)
Population (log)	-6.3987	-6.7532*
	(4.5862)	(3.2409)
Country fixed effects	Yes	Yes
Sector-year fixed effects	Yes	Yes
Observations	1261	2081
Pseudo R^2	0.302	0.242

Country-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

So far, we always estimated separate models for the any MS and the deep MS variables. However, the choice whether to include any gender mainstreaming at all affect whether staff can include deep mainstreaming. Therefore, we expect that errors would be correlated between the two models. To ensure that such correlated errors do not affect the conclusions we can draw, we re-estimate them using the conditional mixed process estimator, a special variant of seemingly unrelated regression (Model 6). We also test robustness to unobserved country-level confounders by including more stringent country-year fixed effects in the regressions (Model 7). The results are similar as the ones reported in the main body of the article.

Table A9: Conditional mixed process models (two-stage model)

	(6)	(7)
Any MS (probit)		
TTL women	0.0082	-0.0403
	(0.1213)	(0.1634)
TTL Gender expertise	0.1779	0.1916
	(0.1192)	(0.2132)
CD Women	0.3875*	5.1927***
	(0.1930)	(1.0374)
PM Women	-0.1874	-0.1480
	(0.1176)	(0.2290)
Gender project	0.0000	0.0000
	(.)	(.)
IDA	0.9741***	1.9989***
	(0.2336)	(0.4356)
Amount (log)	0.0796	0.1976
	(0.0694)	(0.1119)
Post-conflict country	-1.2071**	-10.8663***
	(0.4479)	(0.3659)
Women ministers	-0.5180	
	(1.1187)	
Principals gender lending	0.4915	
	(0.5343)	
Women economic rights	0.2510	
	(0.2380)	
Women infant mortality	-0.0002	
	(0.0151)	
Women vulnerable employment	-0.0080	
	(0.0263)	
GDP per capita	-0.0000	
	(0.0002)	
Population (log)	-4.2811	
	(2.6692)	
Deep MS (OLS)		
TTL women	0.1207***	0.1155**
	(0.0360)	(0.0402)
TTL Gender expertise	0.0406***	0.0410**
	(0.0123)	(0.0139)
CD Women	0.0378	1.1627**
	(0.0449)	(0.3743)
PM Women	-0.0279	-0.0447
	(0.0363)	(0.0375)
Gender project	0.0205***	0.0138***
	(0.0035)	(0.0037)
IDA	-0.0445	-0.0020
	(0.0819)	(0.1100)
Amount (log)	0.0470*	0.0856***
	(0.0236)	(0.0250)
Post-conflict country	-0.0591	0.0364
	(0.0980)	(0.1593)
Women ministers	-0.1295	
	(0.3849)	
Principals gender lending	0.1053	
	(0.1981)	
Women economic rights	0.1079*	
	(0.0527)	
Women infant mortality	0.0068	
	(0.0107)	
Women vulnerable employment	-0.0016	
	(0.0091)	
GDP per capita	0.0001	
	(0.0001)	
Population (log)	-1.4081	
	(1.0900)	
Project-level controls	Yes	Yes
Country-level controls	Yes	No
Country fixed effects	Yes	No
Country-year fixed effects	No	Yes
Sector-year fixed effects	Yes	Yes
Observations	2073	2335

Country-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Alternative clustering of standard errors

Additionally, we test the sensitivity of our results to alternative clustering approaches. In the main models, we clustered by country. Now, we cluster standard errors at the country-year level to account for correlated errors in the same country and the same year. Models 8 and 9 are OLS regressions, Models 10 and 11 instrumental variable models. The results remain robust to alternative clustering.

Table A10: country-year clustered standard errors

	(8) Any MS	(9) Deep MS	(10) Any MS	(11) Deep MS
TTL Women	0.0075	0.1074*	0.0913	0.3811**
	(0.0159)	(0.0434)	(0.0561)	(0.1446)
TTL Gender expertise	0.0100	0.0622***	0.0084	0.0571***
	(0.0051)	(0.0175)	(0.0051)	(0.0172)
CD Women	0.0539*	0.1737**	0.0533*	0.1716**
	(0.0228)	(0.0641)	(0.0229)	(0.0644)
PM Women	-0.0212	-0.0628	-0.0178	-0.0516
	(0.0164)	(0.0505)	(0.0166)	(0.0511)
Gender project	0.0038*	0.0297***	0.0036*	0.0292***
	(0.0016)	(0.0057)	(0.0016)	(0.0057)
IDA	0.1468**	0.2799*	0.1438**	0.2701*
	(0.0485)	(0.1249)	(0.0483)	(0.1242)
Amount (log)	0.0084	0.0537	0.0068	0.0486
	(0.0091)	(0.0287)	(0.0091)	(0.0277)
Post-conflict country	-0.1850***	-0.4221***	-0.1817***	-0.4114***
	(0.0392)	(0.1237)	(0.0382)	(0.1231)
Women ministers	0.0537	-0.0015	0.0351	-0.0621
	(0.1421)	(0.4147)	(0.1438)	(0.4163)
Principals gender lending	0.0873	0.1947	0.0958	0.2224
	(0.0697)	(0.2067)	(0.0690)	(0.2053)
Women economic rights	0.0072	0.0934	0.0078	0.0955
	(0.0212)	(0.0662)	(0.0212)	(0.0662)
Women infant mortality	0.0024	0.0078	0.0017	0.0056
	(0.0046)	(0.0118)	(0.0045)	(0.0115)
Women vulnerable employment	0.0001	-0.0029	0.0004	-0.0019
	(0.0042)	(0.0115)	(0.0042)	(0.0114)
GDP per capita	0.0001**	0.0001	0.0001*	0.0001
	(0.0000)	(0.0001)	(0.0000)	(0.0001)
Population (log)	-0.8305*	-3.0326**	-0.9012*	-3.2636**
	(0.3980)	(1.1142)	(0.4018)	(1.1251)
Country fixed effects	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes
Observations	2076	2076	2076	2076
R^2	0.368	0.456		

Country-year-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Staff categories

In our main models, we included all three staff categories (TTL, CD, PM). To ensure that our results are not driven by this choice, we now present separate estimates for each staff category. The results are very similar. Women TTLs are associated with deeper mainstreaming (Models 12 and 13). Country directors are associated with an increased likelihood of any mainstreaming as well as deeper mainstreaming (Models 14 and 15). Finally, the coefficients for practice managers fail to attain statistical significance at conventional thresholds (Models 16 and 17).

Table A11: Estimating models for each staff category separately

	(12) Any MS	(13) Deep MS	(14) Any MS	(15) Deep MS	(16) Any MS	(17) Deep MS
TTL Women	0.0068	0.1067*				
	(0.0174)	(0.0485)				
CD Women			0.0585*	0.1889**		
			(0.0265)	(0.0717)		
PM Women					-0.0218	-0.0730
					(0.0150)	(0.0476)
TTL Gender expertise	0.0112*	0.0655**	0.0100	0.0648**	0.0099	0.0650**
	(0.0052)	(0.0195)	(0.0052)	(0.0199)	(0.0052)	(0.0199)
Gender project	0.0036*	0.0298***	0.0035*	0.0296***	0.0039*	0.0301***
	(0.0017)	(0.0062)	(0.0016)	(0.0061)	(0.0019)	(0.0067)
IDA	0.1411***	0.2526*	0.1507***	0.2958*	0.1408***	0.2705*
	(0.0338)	(0.1152)	(0.0340)	(0.1178)	(0.0331)	(0.1177)
Amount (log)	0.0086	0.0540	0.0071	0.0537	0.0062	0.0520
	(0.0106)	(0.0302)	(0.0105)	(0.0311)	(0.0106)	(0.0311)
Post-conflict country	-0.1856**	-0.4292*	-0.1795**	-0.4083*	-0.1898**	-0.4363*
	(0.0606)	(0.1950)	(0.0584)	(0.1924)	(0.0604)	(0.2000)
Women ministers	0.1247	0.2488	0.0401	0.0298	0.1099	0.2417
	(0.1428)	(0.4553)	(0.1393)	(0.4581)	(0.1394)	(0.4490)
Principals gender lending	0.0962	0.2250	0.0831	0.2035	0.1026	0.2554
	(0.0800)	(0.2619)	(0.0786)	(0.2607)	(0.0805)	(0.2667)
Women economic rights	0.0160	0.1080	0.0065	0.0773	0.0120	0.1074
	(0.0239)	(0.0700)	(0.0246)	(0.0700)	(0.0243)	(0.0723)
Women infant mortality	0.0014	0.0058	0.0018	0.0030	0.0022	0.0036
	(0.0052)	(0.0126)	(0.0048)	(0.0136)	(0.0049)	(0.0136)
Women vulnerable employment	0.0002	-0.0045	0.0005	-0.0036	0.0002	-0.0031
	(0.0043)	(0.0120)	(0.0044)	(0.0123)	(0.0044)	(0.0124)
GDP per capita	0.0001*	0.0002	0.0001*	0.0002	0.0001*	0.0002
	(0.0000)	(0.0001)	(0.0000)	(0.0001)	(0.0000)	(0.0001)
Population (log)	-0.8398	-2.7903*	-0.9950	-3.3644**	-0.7592	-2.8318*
	(0.5436)	(1.3181)	(0.5157)	(1.2801)	(0.5103)	(1.3432)
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2087	2087	2103	2103	2102	2102
R^2	0.365	0.453	0.368	0.453	0.367	0.452

Country-clustered standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Controlling for project financiers

Moreover, we expand our approach to controlling for the preferences of political principals. In the main models, we accounted for the preferences of the largest five shareholders of the Bank. However, recent changes in the funding of IOs led to a substantial increase in earmarked funding allocated to IOs (Graham 2023; Heinzel, Cormier, and Reinsberg 2023; Patz and Goetz 2019). These funding modalities allow donors to directly affect the projects they fund. To account for this argument, we collected new data on the funding sources of projects. Specifically, we scraped the World Bank projects & operations website to identify the third parties that co-fund projects. We then include financier dummies for each funder that appears in the data more than once to account for differences in the preferences of the additional funding sources for individual projects. Model 18 and 19 show results using OLS models and Models 20 and 21 focus on instrumental variable regressions. While some of the significance levels change, the results remain consistent with those reported in the main body of the article.

Table A12: Controlling for financier dummies

	(18) Any MS	(19) Deep MS	(20) Any MS	(21) Deep MS
TTL Women	0.0074	0.0919^{+}	0.0960^{+}	0.3391*
	(0.0179)	(0.0486)	(0.0506)	(0.1324)
TTL Gender expertise	0.0102^{+}	0.0607**	0.0087	0.0566**
	(0.0055)	(0.0200)	(0.0054)	(0.0192)
CD Women	0.0601*	0.2103**	0.0592*	0.2077**
	(0.0270)	(0.0743)	(0.0269)	(0.0736)
PM Women	-0.0203	-0.0622	-0.0176	-0.0545
	(0.0164)	(0.0513)	(0.0165)	(0.0508)
Gender project	0.0047*	0.0333***	0.0044*	0.0325***
	(0.0022)	(0.0070)	(0.0020)	(0.0066)
IDA	0.1447***	0.2802*	0.1414***	0.2708*
	(0.0350)	(0.1237)	(0.0351)	(0.1203)
Amount (log)	0.0085	0.0483	0.0068	0.0435
	(0.0117)	(0.0319)	(0.0115)	(0.0312)
Post-conflict country	-0.1833**	-0.4151*	-0.1817**	-0.4105*
	(0.0579)	(0.1813)	(0.0542)	(0.1707)
Women ministers	0.0285	-0.1612	0.0025	-0.2339
	(0.1497)	(0.4749)	(0.1526)	(0.4775)
Principals gender lending	0.0922	0.2360	0.1021	0.2636
	(0.0849)	(0.2749)	(0.0834)	(0.2701)
Women economic rights	0.0081	0.0733	0.0086	0.0746
	(0.0242)	(0.0690)	(0.0244)	(0.0692)
Women infant mortality	0.0026	0.0060	0.0018	0.0037
	(0.0054)	(0.0117)	(0.0054)	(0.0115)
Women vulnerable employment	-0.0008	-0.0052	-0.0005	-0.0043
	(0.0045)	(0.0126)	(0.0045)	(0.0125)
GDP per capita	0.0001^{+}	0.0001	0.0001^{+}	0.0001
	(0.0000)	(0.0001)	(0.0000)	(0.0001)
Population (log)	-0.9091^{+}	-3.4212*	-0.9838^{+}	-3.6298**
	(0.5222)	(1.3056)	(0.5226)	(1.3194)
Financier dummies	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes
Observations	2076	2076	2076	2076
R^2	0.389	0.478		

Country-clustered standard errors in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Rating biases

An additional concern are rating biases. There is a possibility that the post-hoc evaluation of gender mainstreaming was partly driven by gender stereotypes. Evaluators had access to project documents that list the TTLs that were in charge of these projects when assigning GMI scores. Hence, it would be possible that these evaluators implicitly rated the GMI differently when women staff were in charge of projects if their implicit biases shaped ratings. We adapt an approach used by Malik and Stone (2018) to test rating biases in the World Bank for our purposes. Their study probes the degree to which World Bank project evaluations, conducted by the Independent Evaluation Group, are biased. The Independent Evaluation Group publishes project-level evaluations that rate the outcomes of projects on an 1-6 ordinal scale (from Highly Unsatisfactory to Highly Satisfactory). Like in the case of the GMI, evaluators have substantial discretion in determining these ratings. To understand whether this discretion leads to biases, Malik and Stone (2018) collect data from evaluation reports on the achievement of individual numerical objectives. They then calculate a more objective evaluation rating based on the share of project objectives that were achieved. To adapt their approach to our purposes, we scraped the World Bank projects and operations website to collect data on thousands of individual objectives. We then used keyword search to identify objectives that directly mentioned the following terms: "gender, women, men, girls, boys". Based on this key-word search, we created a new variable that indicates whether a project included gender-disaggregated targets—a key part of the monitoring and evaluation score in the GMI. We then test whether our main explanatory variables

explain differences in the M&E rating, holding our more objective score constant (Models 22-24). Our variables do not attain statistical significance at conventional thresholds. Our approach has some limitations since we lack the ability to conduct similar tests for the other GMI dimensions. Nevertheless, we believe that these tests provide evidence that rating bias does not drive our results—at least for the monitoring and evaluation sub-component of the GMI.

Table A13: Testing for rating bias

	(22) M&E rating	(23) M&E rating	(24) M&E rating
Gender-disaggregated objectives	0.1300***	0.1324***	0.1332***
	(0.0321)	(0.0325)	(0.0327)
TTL Women	0.0334		
	(0.0208)		
CD Women		0.0441	
		(0.0311)	
PM Women			-0.0092
			(0.0258)
Country fixed effects	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes
Observations	1456	1456	1455
R^2	0.454	0.453	0.452

Country-clustered standard errors in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Performance models

Some authors have also demonstrated gender differences in bureaucratic or legislative performance (Anzia and Berry 2011; Park 2013). One possibility arising from this literature would be that women staff run higher performing projects in general. In that case, our results would not be driven by anything specific to gender mainstreaming but an overall greater adherence of women staff to organizational guidelines. We conduct additional tests for this argument by employing the World Bank's Independent Evaluation Group (IEG) ratings (World Bank 2015). As discussed, these ratings measure the performance of the World Bank in projects on a scale of Highly Unsatisfactory (1) to Highly Satisfactory (6). IEG have been widely used in the literature to understand differences in project performance (Denizer, Kaufmann, and Kraay 2013; Heinzel 2022; Honig 2019; Honig, Lall, and Parks 2022; Kilby 2000). We estimate two OLS models and two instrumental variable regressions to probe whether Women staff supervise higher performing projects in general. The coefficients for women staff fail to attain statistical significance at conventional thresholds. We interpret this as some initial evidence that our results are not driven by gender differences in adherence to organizational guidelines in general.

Table A14: Performance models

	(25) IEG Performance rating	(26) IEG Performance rating	(27) IEG Performance rating	(28) IEG Performance rating
TTL Women	0.0267	0.0258	0.1794	0.0791
	(0.0636)	(0.0696)	(0.2767)	(0.2960)
TTL Gender expertise		0.0450		0.0419
		(0.0384)		(0.0394)
CD Women		-0.0685		-0.0702
		(0.1074)		(0.1066)
PM Women		-0.0082		-0.0056
		(0.0799)		(0.0821)
Gender project		0.0017		0.0016
		(0.0124)		(0.0123)
IDA		0.2457		0.2418
		(0.1552)		(0.1473)
Amount (log)		0.0943		0.0946
		(0.0506)		(0.0513)
Post-conflict country		0.0019		0.0168
		(0.2512)		(0.2566)
Women ministers		0.9841		0.9709
		(0.6894)		(0.6708)
Principals gender lending		0.2819		0.2988
		(0.2853)		(0.3067)
Women economic rights		-0.0792		-0.0782
		(0.0971)		(0.0973)
Women infant mortality		0.0085		0.0089
		(0.0154)		(0.0155)
Women vulnerable employment		0.0015		0.0008
		(0.0205)		(0.0205)
GDP per capita		0.0001		0.0001
		(0.0001)		(0.0001)
Population (log)		-0.6832		-0.6806
		(1.6811)		(1.6884)
Country fixed effects	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes
Observations	886	792	886	792
R^2	0.305	0.304		
F Statistic			27.739	28.377

Country-clustered standard errors in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Gender themes

We controlled for gender themes in the main part of the manuscript. The gender theme is conceptually distinct from gender mainstreaming. Gender themes measure whether projects have gender equality as a main goal, while gender mainstreaming implies that all projects should ensure that people are not withheld from the benefits of operations due to their gender. Nevertheless, there is some overlap between the two measures empirically. To ensure that our results are not driven by our choice to control for or include gender themed projects in the analysis we conduct additional tests. Specifically, we re-estimate models excluding the gender theme variable (Model 29 and 30) and excluding all projects with gender themes from the analysis (Model 31 and 32). While some of the significance levels change, the results are consistent with the conclusions we drew in the main body of the article.

Table A15: without gender theme control variable

	(29) Any MS	(30) Deep MS	(31) Any MS	(32) Deep MS
TTL Women	0.0078	0.1096*	0.0055	0.0976+
	(0.0177)	(0.0493)	(0.0186)	(0.0499)
TTL Gender expertise	0.0124*	0.0818***	0.0135*	0.0740**
	(0.0051)	(0.0184)	(0.0067)	(0.0238)
CD Women	0.0548*	0.1810*	0.0519+	0.1560*
	(0.0265)	(0.0724)	(0.0281)	(0.0752)
PM Women	-0.0229	-0.0763	-0.0234	-0.0674
	(0.0154)	(0.0492)	(0.0165)	(0.0510)
IDA	0.1471***	0.2820*	0.1511***	0.2840*
	(0.0340)	(0.1173)	(0.0351)	(0.1189)
Amount (log)	0.0082	0.0524+	0.0091	0.0536+
	(0.0106)	(0.0298)	(0.0111)	(0.0303)
Post-conflict country	-0.1840**	-0.4147*	-0.1937***	-0.4515*
	(0.0576)	(0.1859)	(0.0560)	(0.1815)
Women ministers	0.0630	0.0723	0.0598	-0.0293
	(0.1417)	(0.4504)	(0.1451)	(0.4621)
Principals gender lending	0.0849	0.1759	0.1060	0.2596
	(0.0795)	(0.2632)	(0.0827)	(0.2637)
Women economic rights	0.0075	0.0958	0.0079	0.0928
	(0.0247)	(0.0718)	(0.0251)	(0.0741)
Women infant mortality	0.0025	0.0086	0.0025	0.0054
	(0.0054)	(0.0137)	(0.0052)	(0.0122)
Women vulnerable employment	-0.0001	-0.0040	-0.0008	-0.0070
	(0.0044)	(0.0123)	(0.0044)	(0.0122)
GDP per capita	0.0001*	0.0001	0.0001*	0.0001
	(0.0000)	(0.0001)	(0.0000)	(0.0001)
Population (log)	-0.8348	-3.0662*	-0.8145	-3.1312*
	(0.5134)	(1.3180)	(0.5150)	(1.3084)
Country fixed effects	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes
Observations	2076	2076	1971	1971
R^2	0.367	0.450	0.370	0.453

Country-clustered standard errors in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Interaction models

We also include interaction terms between some of our key variables to nuance our results and understand some of the scope conditions. Specifically, we interact our Women TTL and Women CD variables to probe how differences in hierarchy affect GMI implementation (Models 33 and 34). The interactions show that men staff are substantially more likely to implement gender mainstreaming when they answer to women as CDs compared to men CDs. Additionally, we interact our gender variables with the gender expertise variables (Models 35 and 36). To this end, we create a binary gender expertise indicator to ease the interpretation of the interaction. Our results show that gender expertise makes a substantial difference: Men TTLs with gender expertise also design projects with strong gender mainstreaming components and their projects appear to even include deeper mainstreaming on average than those of Women TTLs without gender expertise—although the difference between these two groups is not statistically significant.

Table A16: Interaction models

	(33) Any MS	(34) Deep MS	(35) Any MS	(36) Deep MS
TTL Women	-0.0003	0.1063^{+}	0.0025	0.1169*
	(0.0214)	(0.0594)	(0.0209)	(0.0537)
CD Women	0.0386	0.1716^{+}	0.0538*	0.1759*
	(0.0376)	(0.0991)	(0.0264)	(0.0710)
Interaction TTL Women * CD Women	0.0259	0.0036		
	(0.0339)	(0.0926)		
TTL Gender expertise (count)	0.0102^{+}	0.0622**		
	(0.0053)	(0.0195)		
TTL Gender expertise (binary)			0.0475*	0.3122***
			(0.0213)	(0.0636)
Interaction TTL Women * TTL Gender expertise (binary)			0.0125	-0.0736
			(0.0279)	(0.0720)
PM Women	-0.0218	-0.0629	-0.0203	-0.0598
	(0.0156)	(0.0497)	(0.0150)	(0.0469)
Gender project	0.0037^{+}	0.0297***	0.0035^{+}	0.0292***
	(0.0019)	(0.0063)	(0.0019)	(0.0061)
IDA	0.1465***	0.2799*	0.1447***	0.2709*
	(0.0342)	(0.1144)	(0.0343)	(0.1187)
Amount (log)	0.0087	0.0538^{+}	0.0082	0.0534^{+}
	(0.0105)	(0.0298)	(0.0105)	(0.0295)
Post-conflict country	-0.1859**	-0.4222*	-0.1850**	-0.4170*
	(0.0578)	(0.1866)	(0.0582)	(0.1879)
Women ministers	0.0556	-0.0013	0.0512	-0.0201
	(0.1414)	(0.4500)	(0.1425)	(0.4501)
Principals gender lending	0.0877	0.1948	0.0897	0.2197
	(0.0801)	(0.2623)	(0.0797)	(0.2650)
Women economic rights	0.0069	0.0934	0.0056	0.0848
	(0.0246)	(0.0704)	(0.0246)	(0.0706)
Women infant mortality	0.0024	0.0078	0.0022	0.0073
	(0.0054)	(0.0133)	(0.0053)	(0.0130)
Women vulnerable employment	0.0002	-0.0029	0.0001	-0.0030
	(0.0044)	(0.0123)	(0.0044)	(0.0124)
GDP per capita	0.0001*	0.0001	0.0001*	0.0001
	(0.0000)	(0.0001)	(0.0000)	(0.0001)
Population (log)	-0.8338	-3.0331*	-0.8569^{+}	-3.1549*
	(0.5116)	(1.2952)	(0.5111)	(1.2942)
Country fixed effects	Yes	Yes	Yes	Yes
Sector-year fixed effects	Yes	Yes	Yes	Yes
Observations	2076	2076	2076	2076
R^2	0.368	0.456	0.461	0.461

Country-clustered standard errors in parentheses; $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. Robustness checks for experimental analysis

We display the full results for the conjoint experiment displayed in Figure 3 in the main article in Table A17.

Table A17: Full results of conjoint experiment (AMCE)

	(37) Approval	(38) Impact
Gender-disaggregated targets: Yes	0.5099***	0.3540***
	(0.1107)	(0.0933)
Conditionality: High	-0.2009	-0.0624
	(0.1219)	(0.1264)
Amount: High	0.2226*	0.2608**
	(0.1089)	(0.0815)
Environmental and social risk: Yes	-0.4218***	-0.2720*
	(0.1238)	(0.1092)
Macroeconomic risk: High	-0.2765**	-0.1574
	(0.0961)	(0.0974)
CPIA: High	0.3638***	0.4348***
	(0.1020)	(0.1130)
US board voted against: Yes	-1.2684***	-0.0235
	(0.1417)	(0.1065)
Country: MIC	-0.1160	-0.1289
	(0.1086)	(0.1047)
Constant	6.8622***	5.7835***
	(0.1840)	(0.1725)
Observations	2492	2492
R^2	0.108	0.028

Respondent-clustered standard errors in parentheses; * $p < 0.05$ (pre-registered confidence level); ** $p < 0.01$, *** $p < 0.001$

Differences between sectors

We also conducted additional exploratory robustness checks on the survey results. Initially, we include fixed effects for the primary sector of work as self-reported by World Bank staff to ensure that imbalance across sectoral staff in our sample does not drive the results. The results are substantively similar.

Table A18: controlling for recipients' primary sector of work

	(39) Approval	(40) Impact
Women vs. Men (Gender mainstreaming: No)	0.230	0.511
	(0.317)	(0.296)
Women vs. Men (Gender mainstreaming: Yes)	0.415	0.676*
	(0.283)	(0.288)
Sector fixed effects	Yes	Yes
Observations	2422	2422

Respondent-clustered standard errors in parentheses; * $p < 0.05$ (pre-registered confidence level)

Gender expertise

We further probe whether gender expertise drives the results. To this end, we asked respondents to indicate their perceived expertise on gender issues on a scale from 1 (very low) to 5 (very high). We report simple linear regressions with a host of fixed effects (sector, region of work, degree and nationality) to show that women, on average, report higher gender expertise than men.

Table A19: differences between men and women in their self-assessments of expertise on gender issues

	(41)	(42)	(43)	(44)	(45)
TTL women	0.4316**	0.4250**	0.3862*	0.3526*	0.5612*
	(0.1597)	(0.1544)	(0.1542)	(0.1700)	(0.2666)
Sector fixed effects	No	Yes	Yes	Yes	Yes
Region fixed effects	No	No	Yes	Yes	Yes
Degree fixed effects	No	No	No	Yes	Yes
Nationality fixed effects	No	No	No	No	Yes
Observations	169	169	168	150	94
R^2	0.042	0.241	0.275	0.367	0.546

Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$

We then include gender expertise fixed effects and re-estimate our models. The main coefficient of interest is only marginally significant. These results indicate to us that differential gender expertise between women and men staff is partly but not wholly responsible for the gender differences we observe in the conjoint experiment.

Table A20: controlling for self-assessments of expertise on gender issues

	(46) Approval	(47) Impact
Women vs. Men (Gender mainstreaming: No)	0.229	0.275
	(0.326)	(0.304)
Women vs. Men (Gender mainstreaming: Yes)	0.483	0.514+
	(0.303)	(0.278)
Gender expertise fixed effects	Yes	Yes
Observations	2422	2422

Respondent-clustered standard errors in parentheses; + $p < 0.10$, * $p < 0.05$ (pre-registered confidence level)

Furthermore, we use self-reported gender expertise as an alternative sub-group indicator and estimate the marginal means difference between those with high gender expertise and those with low gender expertise. The results show that those with high gender expertise perceive likelihood of approval and impact higher for projects including gender-disaggregated targets. The coefficients are marginally significant. They also generally perceive the likelihood of approval higher than staff with low gender expertise.

Table A21: self-assessments of expertise on gender issues as treatment (High and very high expertise)

	(48) Approval	(49) Impact
High gender expertise vs. low gender expertise (Gender mainstreaming: No)	0.482	0.577+
	(0.338)	(0.317)
High gender expertise vs. low gender expertise (Gender mainstreaming: Yes)	0.473+	0.582+
	(0.256)	(0.306)
Observations	2422	2422

Respondent-clustered standard errors in parentheses; + $p < 0.10$, * $p < 0.05$ (pre-registered confidence level)

Weighting

Finally, we re-estimate models using alternative weighting approaches. Specifically, we re-weight by gender (Tabel A22), by gender and education (Table A23) as well as gender and nationality (Table A24). The results are generally in line with the results reported in the main body of the article. The Women vs Men (Gender Mainstreaming: Yes) coefficient is statistically significant ($p<0.05$) in all three models. However, we also see a statistically significant difference between men and women in their assessment of the likely impact of a project without gender mainstreaming in Model 51. Furthermore, we observe statistically significant gender differences between women and men in their assessments of the likelihood of approval in Models 52 and 53. Overall, the robustness checks substantiate the conclusions drawn in the main models presented in the article.

Table A22: Alternative weighting by gender

	(50) Approval	(51) Impact
Women vs. Men (Gender mainstreaming: No)	0.227	0.564+
	(0.301)	(0.287)
Women vs. Men (Gender mainstreaming: Yes)	0.456+	0.621*
	(0.257)	(0.272)
Observations	2422	2422

Respondent-clustered standard errors in parentheses; + $p < 0.10$, * $p < 0.05$ (pre-registered confidence level)

Table A23: Alternative weighting by gender and education

	(52) Approval	(53) Impact
Women vs. Men (Gender mainstreaming: No)	0.140	0.638*
	(0.280)	(0.278)
Women vs. Men (Gender mainstreaming: Yes)	0.358	0.743*
	(0.235)	(0.270)
Observations	2408	2408

Respondent-clustered standard errors in parentheses; + $p < 0.10$, * $p < 0.05$ (pre-registered confidence level)

Table A24: Alternative weighting by gender and nationality

	(54) Approval	(55) Impact
Women vs. Men (Gender mainstreaming: No)	0.785*	0.583
	(0.466)	(0.383)
Women vs. Men (Gender mainstreaming: Yes)	0.777*	0.841*
	(0.329)	(0.370)
Observations	1568	1568

Respondent-clustered standard errors in parentheses; + $p < 0.10$, * $p < 0.05$ (pre-registered confidence level)

7. References

Anzia, Sarah F, and Christopher R Berry. 2011. "The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55(3): 478–93.

Bechtel, Michael M., and Kenneth F. Scheve. 2013. "Mass Support for Global Climate Agreements Depends on Institutional Design." *Proceedings of the National Academy of Sciences of the United States of America* 110(34): 13763–68.

Briggs, Ryan C. 2021. "Why Does Aid Not Target the Poorest?" *International Studies Quarterly* 65(3): 739–52.

Cingranelli, David L., and David L. Richards. 2010. "The Cingranelli and Richards (CIRI) Human Rights Data Project." *Human Rights Quarterly* 32(2): 401–24.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance." *Journal of Development Economics* 105: 288–302.

Ghassim, Farsan, Mathias Koenig-Archibugi, and Luis Cabrera. 2022. "Public Opinion on Institutional Designs for the United Nations: An International Survey Experiment." *International Studies Quarterly* 66(3): sqac027.

Graham, Erin R. 2023. *Transforming International Institutions: How Money Quietly Sidelined Multilateralism at The United Nations*. Oxford: Oxford University Press.

Green, Paul E., Abba M. Krieger, and Yoram Wind. 2001. "Thirty Years of Conjoint Analysis: Reflections and Prospects." *Interfaces* 31(3-Supplement): 56–73.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 112(8): 2395–2400.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1): 1–30.

Heinzel, Mirko. 2022. "International Bureaucrats and Organizational Performance. Country-Specific Knowledge and Sectoral Knowledge in World Bank Projects." *International Studies Quarterly* 66(2): sqac013.
Heinzel, Mirko, Ben Cormier, and Bernhard Reinsberg. 2023. "Earmarked Funding and the Control-Performance Trade-Off in International Development Organizations." *International Organization* 77(2): 475–95.
Heinzel, Mirko, and Andrea Liese. 2021. "Managing Performance and Winning Trust: How World Bank Staff Shape Recipient Performance." *Review of International Organizations* 16: 625–53.
Honig, Dan. 2019. "When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation." *International Organization* 73(1): 171–201.
Honig, Dan, Ranjit Lall, and Bradley C. Parks. 2022. "When Does Transparency Improve Institutional Performance? Evidence from 20,000 Projects in 183 Countries." *American Journal of Political Science*: 1–21.
Kilby, Christopher. 2000. "Supervision and Performance: The Case of World Bank Projects." *Journal of Development Economics* 62: 233–59.
Leeper, Thomas J., Sara B. Hobolt, and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28(2): 207–21.
Malik, Rabia, and Randall W Stone. 2018. "Corporate Influence in World Bank Lending." *The Journal of Politics* 80(1): 103–18.
Nyrup, Jacob, and Stuart Bramwell. 2020. "Who Governs? A New Global Dataset on Members of Cabinets." *American Political Science Review* 114(4): 1366–74.
OECD. 2024. "Creditor Reporting System." https://stats.oecd.org/index.aspx?DataSetCode=CRS1# (January 5, 2024).
Park, Sanghee. 2013. "Does Gender Matter? The Effect of Gender Representation of Public Bureaucracy on Governmental Performance." *American Review of Public Administration* 43(2): 221–42.
Patz, Ronny, and Klaus H. Goetz. 2019. *Managing Money and Discord in the UN*. Oxford: Oxford University Press.
Winters, Janelle, Genevie Fernandes, Lauren McGivern, and Devi Sridhar. 2018. "Mainstreaming as Rhetoric or Reality? Gender and Global Health at the World Bank." *Wellcome Open Research* 3(18).
World Bank. 2015. *Independent Evaluation Group. World Bank Project Performance Ratings -Codebook*. Washington, DC. https://ieg.worldbankgroup.org/sites/default/files/Data/reports/ieg-wb-project-performance-ratings-codebook_092015.pdf.
———. 2018. *Monitoring Gender Mainstreaming in the World Bank Lending Operations Database*. Washington, D.C: World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/207551599142919624/uzbekistan-health-system-improvement-project.
———. 2020a. "World Bank Projects & Operations." https://projects.worldbank.org/ (March 2, 2020).
———. 2020b. "World Development Indicators." http://datatopics.worldbank.org/world-development-indicators/ (January 1, 2020).
———. 2021. "Development Policy Actions." pubdocs.worldbank.org/en/861551551301960896/DPF-dev-policy-action-database.xlsx (February 9, 2021).
Zeitz, Alexandra O. 2021. "Emulate or Differentiate? Chinese Development Finance, Competition, and World Bank Infrastructure Funding." *Review of International Organizations* 16(2): 265–92.