Are International Monetary Fund (IMF) programs really (in)effective? Introducing a new tool to assess external validity of regression analysis

Tal Sadeh
Tel Aviv University
talsadeh@tauex.tau.ac.il

Gal Bitton
Tel Aviv University
galbitton@mail.tau.ac.il

Bernhard Reinsberg
University of Glasgow
Bernhard.Reinsberg@glasgow.ac.uk

Tags: Data Analysis; Developing Countries; Development; Empirical Analysis; Foreign Aid; International Finance; Institutions; Methodology; Political Economy; WTO/IMF/World Bank

Abstract

According to the literature, the IMF's track-record in averting financial crises and promoting economic growth is mixed, and evidence suggests that IMF programs may increase poverty and income inequality, and have adverse and even gendered effects on unemployment, and labour income and rights. IMF programs are also linked to deteriorating public health, educational outcomes, vaccination rates, child mortality, corruption, government instability and the likelihood of civil war. We replicate results from related articles in top journals, and find that many of them effectively base their conclusions on a small set of countries or years, even when their nominal samples are large. Indeed, researchers face a critical trade-off: the promise of enhanced internal validity comes at the cost of external validity. For this we develop indicators of effective samples, which tell us if a particular estimate is based on the entire data fed into the regression, or rather on an effectively narrower subset of observations (implying lower external validity). These indicators, which are comparable across models and datasets, can be applied to a range of regression analyses and methods. Our project has broader methodological implications. In the past decade, scholars of Political Science in general and International Political Economy (IPE) in particular have increasingly resorted to experimental designs in order to test their hypotheses. Experimental designs help identify causal effects but are criticized for suffering from lower external validity compared with big-data econometric models. The indicators of external validity that we develop can help scholars manage and optimize this trade-off.

Word count: 14,697 (11,938 exclusive of cover page and appendices)

Introduction

The International Monetary Fund (IMF) has attracted abundant research on its policies and practices. The Fund's mandate is to avert financial crises and promote economic growth. However, according to the literature the track-record on both outcomes is mixed. Some studies find that IMF programs harm economic growth (Dreher 2006; Marchesi and Sirtori 2011; Przeworski and Vreeland 2000), others contend that this effect is due to adverse selection (Bas and Stone 2014). IMF programs encourage current account liberalization and enhance credit ratings, but this may increase moral hazard and vulnerability to the volatility of transnational financial flows. IMF programs only partly catalyse private and bilateral foreign aid and direct investments. Conditionality catalyses more FDI, but may be counterproductive in some cases (Woo 2013). More broadly, the evidence suggests that IMF programs may increase poverty and income inequality, and have adverse and even gendered effects on unemployment, and labour income and rights. IMF programs are also linked to deteriorating public health, educational outcomes, vaccination rates, child mortality and corruption. There is some evidence that IMF programs also increase government instability and the likelihood of civil war.

This study extends the findings of the IMF effectiveness literature in a critical dimension. Over the past decades, researchers have been vexed by the challenge of identifying the treatment effect of IMF programs. The focus on internal validity has relegated external validity to a neglected second-order issue. Consequently, our knowledge about to what extent the findings from this research program generalize across different countries and different time periods has remained extremely limited. We begin to fill this gap by developing indicators of effective samples in econometric models that tell us if a particular estimate is based on the entire data fed into the regression, or rather on a narrower effective subset of observations (implying lower external validity). These are indicators of representative statistics, or in short *REPSTAT* indicators, are easy to produce, intuitive to understand, and computable for a range of models, including linear, multi-step or selection models, generalized linear models estimates (logit, probit, multinomial logit or probit, Poisson regression, and parametric or semiparametric duration models), random coefficient models and simultaneous equation estimators.

We find that the vast empirical work on the effectiveness of IMF programs includes many studies that effectively base their conclusions on a small set of countries or years, even when their nominal samples are large. In about one half of the IMF program treatments in the studies that we replicate, estimates effectively rely on fewer than one-quarter of the observations. No study had an effective sample size over 70 percent. In one study recently published in one of the most prominent journals in the field, the headline result is effectively based on only 9 observations, and overwhelmingly driven by the particular cases of one or two countries, or a single time period. Indeed, in contrast to experimentalists' arguments we find that researchers do face a critical trade-off: the promise of enhanced internal validity comes at the cost of external validity. The indicators of external validity that we develop can help scholars manage and optimize this trade-off.

Our project has broader methodological implications. In the past decade, scholars of Political Science in general and International Political Economy (IPE) in particular have increasingly resorted to experimental designs in order to test their hypotheses. Experimental designs create an artificial lab-like environment for testing hypotheses, which helps identify causal effects, but they are criticized for suffering from lower external validity compared with big-data econometric models that are based on observational data. In their defense, experimentalists retort that inference from regression analysis of observational data often relies on far fewer data points than implied by the entire dataset used (Aronow and Samii 2016; Samii

2016). Experimental designs are therefore preferable to observational studies on internal validity grounds and are no worse in terms of the generalizability of the findings. According to these arguments, no real tradeoff exists between internal and external validity in the choice between these two types of research designs. Our study demonstrates that this does not have to be true. As scholars optimize the tradeoff between internal and external validity, we suggest that *REPSTAT* indicators be reported as a norm.

The next section develops the conceptual framework of our method and suggests a range of applications. The third section reviews current literature on the effectiveness of IMF interventions, and the fourth section describes our data collection and replication project. As a side note, we are disappointed by the lower-than-expected standards of replication, even in very recent studies and sometimes even in the top journals in the field. The fifth section uses our indicators to highlight and characterize problems of external validity in this literature. In the concluding section we suggest some dos and don'ts to help scholars optimize their research designs.

Effective samples – conceptual framework and applications

In the past decade, scholars of Political Science in general and International Political Economy (IPE) in particular have increasingly resorted to experimental designs in order to test their hypotheses. Experimental designs create an artificial lab-like environment for testing hypotheses, which helps identify causal effects, but they are criticized for suffering from lower external validity compared with big-data econometric models that are based on observed data. Experimental designs base their conclusions on an abstraction of reality, which in the social sciences may or may not be relevant to other people in other times. Even if an experiment can be repeated with different people, scholars cannot turn the clock back and repeat the experiment in historical episodes. In their defense, Experimentalists retort that inference from regression analysis of observational data often relies on far fewer data points than implied by the entire dataset used (Aronow and Samii 2016; Samii 2016). Experimental designs are therefore preferable to observational studies on internal validity grounds and are no worse in terms of the generalizability of the findings. No real tradeoff exists between internal and external validity in the choice between these two types of research designs. We explore whether this is true. We develop several indicators of effective samples in econometric models, which tell us if a particular estimate is based on the entire data fed into the regression, or rather on a narrower subset of observations (implying lower external validity). These are indicators of representative statistics, or in short *REPSTAT* indicators.

Scholars using multiple regression design often attempt to base their results on a large n dataset. The larger the dataset, the lower the variance of the regression estimates can be, and presumably the more externally valid they are. The hope is, in other words, that such large dataset will produce findings generalizable to a population beyond the dataset. Even if a dataset covers all countries (or other panel units) and years, there is a super-population of possible realizations that could have materialized given that nature is stochastic (Aronow and Samii 2016, 252). This is one of the potential advantages of regression analysis over experimental research designs, which are typically limited to a very small subpopulation. However, the mere existence of an observation in a dataset does not mean that it necessarily contributes much to an effect's estimate. A data point adds value to the estimate only to the extent that it does not merely reflect information contained in other data points.

Regression analysis estimates the coefficient of each independent variable, calculating it as the average effect of the variable across all of the observations in the dataset used (the nominal sample). Each

observation has a different weight in this average. The weight of each observation in turn is given by the amount of information it adds over the other covariates. This added information is calculated in linear regression as the squared difference between the variable's actual value in that observation, and its fitted value when regressed on the other covariates (making sure to exclude the same observations that are excluded from the outcome regression, if any).[1]

Consider Ordinary Least Squares (OLS) Regression (1) with an outcome *Y*, a treatment (causal effect that may be dichotomous or continuous) T and its coefficient β_T, and a vector of confounding variables *X* and their coefficients β, for n observations indexed by *i*.

$$(1)\ Y_i = \alpha + \beta_T T_i + \beta X_i + \varepsilon_i$$

w_{Ti} – the weight of each observation *i* in calculating the average coefficient β_T is based on the fitted value $\hat{T}_i$ from Regression (2):

$$(2)\ T_i = \alpha_T + \beta X_i + \mu_i$$

Specifically:

$$(3)\ w_{Ti} = (T - \hat{T}_i)^2$$

As Samii (2016, 945) suggests, the size of the effective sample, in contrast to the nominal sample, is the number observations with a large weight, "…the transformation of the nominal sample after reweighting by the multiple regression weights." The size of the effective sample is important because it indicates the extent to which the result of the regression can be generalized to a wider population. If the result is in fact supported by a small number of observations it may not be generalizable, and may even reflect the existence of some latent omitted variable.

To count the number of such meaningful observations per treatment, one may want to establish a threshold weight for them. This raises three concerns. First, any such threshold might be arbitrary, unless some objective way is found to determine it. Second, it would be helpful if the threshold is defined in universal terms, that would allow comparison of effective sample size for different variables, and even different models and datasets. Finally, setting an arbitrary threshold is an inefficient way of calculating the effective sample size, because it overplays the importance of observations that are just above the threshold, while discarding all of the observations that are just below the threshold.

We propose to overcome these problems by applying a concentration/fragmentation index, similar to the Herfindahl-Hirschman Index (HHI) that scholars use to calculate industrial concentration or parliamentary fragmentation. The Effective Sample Concentration (ESC) index measures how concentrated are the observations' weights in calculating the treatment's coefficient β_T. Specifically, we calculate the relative weight of each observation (Rw_{Ti}), which is the share of its weight in the sum of weights of all observations (i.e. its share in the sum of squared residuals when the variable is regressed on the other covariates):

[1] Results are not much different for generalized linear models estimates (logit, probit, multinomial logit or probit, Poisson regression, and parametric or semiparametric duration models) and random coefficient models (Aronow and Samii 2016, 256–57).

$$(4)\ Rw_{Ti} = \frac{w_{Ti}}{\sum_{i=1}^{n} w_{Ti}} = \frac{(T - \hat{T}_i)^2}{\sum_{i=1}^{n}(T - \hat{T}_i)^2}$$

Next, the relative weights of all of the observations are squared and these squares are summed to produce the concentration index:

$$(5)\ ESC_T = \sum_{i=1}^{n} Rw_{Ti}^2$$

The resulting index ranges from the inverse of n – the nominal sample's size (i.e. all observations have identical weights), to 1 (all but a single observation have zero weights). The inverse of this index returns the effective sample size in terms of the number of equally-weighted observations that would return the same fragmentation index, which ranges from 1 to n.

$$(6)\ EffSamp_T = \frac{1}{ESC_T}$$

The effective sample size ($EffSamp_T$) is an indicator of external validity of the findings. Sorting observations in the nominal sample by their weights (w_{Ti}) can highlight the observations with the highest impact on the estimated coefficient.

Both the ESC index and the size of the effective sample are sensitive to the specification of the model – any change in the vector of confounding variables, adding or removing fixed effects, and dropping or adding observations, requires re-calculation of the *REPSTAT* indicators. We thus also calculate the Relative Effective Sample (RES) as a universal measure, that would allow comparison of effective sample size across different models and different datasets:

$$(7)\ RES_T = \frac{EffSamp_T}{n}$$

RES ranges from $1/\ n$ to 1. We interpret a value of say, 0.43 to mean that the estimate β_T is effectively based on 43 percent of the observations in the nominal sample.

If a multiple regression includes more than one treatment then these indicators can all be calculated separately for each of the treatments, and the effective sample size for each treatment indicates the ability to generalize its coefficient's estimate to other populations or super-populations. In other words, the effective sample size is a feature of the treatment variable, not a feature of the entire model. Of course, all this is only relevant for observational data. In experimental designs, regression is used to analyze the drivers of change in an outcome variable, but since the variables are randomized, observations are by definition equally weighted.

Applications

The *REPSTAT* indicators have a range of practical applications.

1. Test of tests: Note that the size of the effective sample behind the treatment's estimated coefficient is independent of the coefficient's statistical significance. Statistically significant results may be derived from a relatively small effective number of observations, which would call into question the

importance of such results. Alternatively, significant results based on a relatively large effective number of observations should be taken more seriously in support of theoretical claims. Similarly, insignificant results, or significant results in the unexpected direction may not be taken as a serious refutation of hypotheses if based on a relatively small effective number of observations. In such instances, perhaps it is better to conclude that the test itself has failed, and more information is needed to come to conclusions.

2. Model specification: Social scientists sometimes face a tradeoff between the specification of a comprehensive vector of confounders to properly control for all conceivable intervening effects, and the size of the nominal sample. As available observed data may not perfectly overlap across confounders, the more confounders are specified in the model the more observations are dropped and the smaller is the resulting nominal sample. This poses a dilemma if the drop in the size of the nominal sample is interpreted as an erosion of external validity of the findings. However, it is possible that adding a confounder to the model results in only a small decline, or even a rise in the treatment's effective sample size even if the nominal sample declines. This can happen if the dropped observations contributed little to the estimate of the treatment's coefficient. In such cases no tradeoff exists between internal and external validity.

3. Tangibility and latency: To make the concept of the effective sample more tangible, a group of the m heaviest observations can be singled out, where $m = EffSamp_T$, and its features studied.[2] This exercise, which is more helpful the smaller is the effective sample, can identify which countries and/or years mostly drive the estimate; *t*-test can be applied to find out whether these m observations are distinguishable from the other observations in any meaningful way. The heaviest observations may be significantly larger or smaller even in the model's dependent variable, treatment or confounders. Scholars may want to reflect on this latency and what it means to their findings. Alternatively, arbitrary thresholds can determine m, such that the included observations make up the top decile of the nominal sample by weight, or the heaviest 5 percent.

4. Multi-step and selection models: *REPSTAT* indicators can be calculated for instrumented treatments as well, and/or include instrumented confounders in the calculation, since the instruments are not included in the outcome (final step) regression. There is no need to calculate *REPSTAT* indicators for the early stage regressions, but if the multi-step estimator is executed automatically the instrumented variables must be replicated and specified manually in Regression (2) above, as either the dependent variable (if it is the treatment variable) or one of the confounders. This is true for manual instrumentation of variables that are then specified in the outcome regression,[3] Two-Step Least Squares (2SLS) regressions, such as *ivreg* or *ivprobit* commands in *Stata*, as well as Generalized Method-of-Moments (GMM) estimators (*xtabond2*) (Roodman 2009). Similarly, in selection models (such as a manual Heckman procedure or *etregress*) the hazard or the inverse Mills ratio must be replicated and specified as one of the confounders in Regression (2).

---

[2] Of course, these m observations are merely an illustration of the effective sample, as in reality they are not equally weighted, and together they account for less than 100 percent of the estimate.

[3] In the literature reviewed for this study, we find such practice in OLS regressions, but also in some Conditional Mixed Process (CMP) estimators.

5. Fixed effects and categorical variables: *REPSTAT* indicators treat fixed effects of any form (such as country fixed effect or year fixed effects) as well as factor (or categorical) variables as sets of confounders. If the outcome regression includes such terms they must be specified in Regression (2); for users of Stata statistical package, if the option *fe* is selected in the outcome regression it should be selected in Regression (2) as well.

6. Interaction terms: *REPSTAT* indicators can be calculated for interacted treatments, and/or include interactions among the confounders. Of course, all constitutive terms must be specified as confounders in Regression (2) even if the interaction term is the treatment (and is specified as the dependent variable there). The indicators' values for the interaction term are bounded by those that would be calculated for its constitutive terms. This means that in marginal effects analysis, the effective sample of a particular point is some weighted average of the effective sample of the interaction term and the effective sample of the relevant constitutive term. For accurate calculation specify the particular linear combination of the two terms as dependent variable in Regression (2).

7. Matching models: *REPSTAT* indicators are calculated per a specified outcome equation. They thus cannot be calculated in the absence of a full observational outcome equation. For example, *REPSTAT* indicators cannot be calculated when the Propensity Score Matching (PSM) method is used, in which a treatment effect is computed by taking the average of the difference between the outcome for the treated subject (country) and the outcome for its matched/paired subject (the control country).

8. Simultaneous equation estimators: Studies using simultaneous equations often seek to address inferential threats such as selection bias and endogenous treatments by explicitly modeling these processes through auxiliary equations. The related estimates are obtained by maximizing a log-likelihood function. While one of the equation is typically regarded as the outcome equation, all equations are estimated jointly in one stage, with a common error structure.

   Joint estimation poses a challenge for *REPSTAT* indicators, which are calculated based on an independently estimated outcome equation. In estimation of simultaneous equations, the added value of each observation in generating a particular coefficient in a particular equation should be calculated considering all of the variables in all of the equations. Here it matters how variables are grouped together in different equations, and only their residual matters for subsequent estimation stages after partialling out the effect of any predetermined covariates. Variables may also appear in more than one equation, thus constraining the search for the optimal set of coefficients more than other variables do.

   While we provide a more technical explanation in Appendix A, we lay out the key steps for how to obtain *REPSTAT* indicators for the outcome equation in a simultaneous-equation framework here. To fix ideas, assume there is one auxiliary equation for a potentially endogenous treatment variable, including covariates from the outcome equation and an excluded instrument. First, we need to generate the residuals of all covariates that appear in both equations, obtained from a regression of all covariates appearing in the auxiliary equation. This isolates the added value of a regressor on the outcome after adjusting for the underlying explanatory value of all these other variables. Second, where regressors only appear in the outcome equation, they enter without adjustments. Third, when there are multiple auxiliary equations, we must calculate cascading residuals of outcome-equation regressors based on all of the equations in which they appear. This calculation of cascading residuals continues until all auxiliary equations are exhausted, and we use the last calculation of residuals in the outcome equation for any variable that appeared in any other equation. Fourth, further adjustments

are necessary when we do not observe all equations over the entire dataset.[4] In essence, for observations that are defined for all equations, we perform the calculation of cascading residuals as described above. For observations that only pertain to the outcome equation, we can use the original regressors' values in the outcome equation. Where we observe only the auxiliary equations but not the outcome equation, the residuals of the covariates are zero, because they make no contribution to explaining variation in the outcome.

A review of the literature

The International Monetary Fund (IMF)—as the global lender of last resort for countries in economic trouble—is one of the most powerful international organizations. To uphold global financial stability, the IMF conducts regular assessments of the macroeconomic policies of its 190 member states and provides technical assistance on fiscal issues and macroeconomic policies to its lower-income members. However, the IMF has the most prominent role as a provider of emergency loans to countries in economic trouble. As the IMF often attaches far-reaching policy conditions to its loans, it is discussed controversially, especially in the developing countries (Easterly 2005; Stone and Steinwand 2008; Vreeland 2003).

Reflecting its prominent role in global financial governance, the IMF has attracted abundant research on its policies and practices. One branch of the literature seeks to understand the determinants of IMF policies (Dreher, Sturm, and Vreeland 2009; Lang and Presbitero 2018; Lombardi and Woods 2008). Key actors who can influence IMF policies include powerful member states (Stone 2004), the IMF staff (Lang and Presbitero 2018), and political elites in borrowing countries (Chwieroth 2014). A prominent finding in the political economy literature is that powerful states can often exert disproportionate influence on IMF policy decisions. Consequently, borrowing countries that are geopolitically aligned with those powerful states can expect to obtain more generous lending packages under more favourable terms compared to those who lack such political clout (Barro and Lee 2005; Dreher and Jensen 2007; Dreher, Sturm, and Vreeland 2009). While powerful states sometimes exert influence over lending decisions in high-profile cases, it is typically the institutional rules and organizational self-interests that shape these decisions. Recent research shows that the Fund offers more generous lending terms when the default of a borrower could threaten systemic financial stability (Brown 2023; Kaplan and Shim 2021). Looking beyond lending operations, the IMF staff also uses its 'room for discretion' in the preparation of debt sustainability ratings (Lang and Presbitero 2018).

Another branch of the 'IMF literature'—which is the focus of our inquiry—is concerned with the effectiveness of IMF interventions. Some have considered effectiveness in terms of the degree to which the Fund affects economic policies, linked to promoting economic freedom (Boockmann and Dreher 2003), reducing the risk of nationalization (Biglaiser, Lee, and Staats 2016), and effective resource governance (Goes 2023). Yet, effectiveness can also be understood as the degree to which the Fund accomplishes its mandate, which is to avert financial crises and promote economic growth. The IMF's track-record on both outcomes is mixed. Some find that IMF programs have no effect on exports (Demir

[4] Such complication is for example typical of Conditional Mixed Process (CMP) estimators, which by default use the union of the different auxiliary equations' sets of observations. Other simultaneous estimators, such as Seemingly Unrelated Bivariate Probit, take only data that are observed across all equations, so are less complicated for *REPSTAT* calculations.

2022) and even harm economic growth (Dreher 2006; Marchesi and Sirtori 2011; Przeworski and Vreeland 2000), others contend that this effect is due to adverse selection (Bas and Stone 2014). As regards crisis prevention, IMF programs do encourage current account liberalization (Pinheiro, Chwieroth, and Hicks 2015). This may make countries more vulnerable to the volatility of transnational financial flows, implying elevated risks for financial crises (Dreher and Walter 2010), especially among borrowers with 'moral hazard' (Lipscy and Lee 2019). Such moral hazard arises from close ties to powerful shareholders, which lessen the incentive to self-insure against crises and increase incentives for reckless macroeconomic policies. Another reason for why IMF programs do not prevent future crises is that they come with unrealistic expectations about 'catalysis' of external resources. Specifically, IMF programs catalyse aid on average, but only for sectors like budget support and debt relief that relate to IMF activities and more strongly for the most powerful bilateral donors in terms of IMF vote shares (Stubbs, Kentikelenis, and King 2016). Regarding foreign direct investment (FDI), findings are mixed, and catalysis effects are contingent. IMF borrowers tend to be more attractive to U.S. investors but not all IMF programs have the same effect (Biglaiser and DeRouen 2010). Similarly, programs with stricter conditions appear to catalyse more FDI (Woo 2013). Other research finds that IMF programs decrease FDI inflows (Bird and Rowlands 2002), especially in sectors that are highly dependent on external capital and have low sunk costs in the host country (Breen and Egan 2019). A final explanation for limited effectiveness of IMF programs is that countries are overburdened with policy conditions, leading to program interruptions, loss of investor confidence, and the return to the Fund as lender of last resort. Analysis at program level shows that programs with more binding conditions increase the likelihood of program interruptions (Reinsberg, Stubbs, and Kentikelenis 2022). These program interruptions trigger a loss in investor confidence, which increases the cost of financing (Chapman *et al*. 2015; Edwards 2006; Reinsberg, Stubbs, and Kentikelenis 2021). Notwithstanding the adverse effects of interruptions, there is evidence that IMF programs can enhance credit ratings due to 'signalling effect' (Gehring and Lang 2020). This may be especially true for left-wing governments (Cho 2014), when the legislature approves the program (David, Guajardo, and Yépez 2022), or when the recipient government is popular among voters, if investors associate higher government popularity with better implementation of the program (Shim 2022).

The effectiveness of IMF programs can be understood more broadly, beyond the macroeconomic outcomes that the programs seek to directly control. A large body of research examines the effects of IMF programs for poverty, inequality, and human development more broadly. Evidence suggests that IMF programs increase poverty, especially where they contain far-reaching structural reforms that cut back state provision, restructure tax systems, and raise unemployment (Biglaiser and McGauvran 2022). Probing these mechanisms further, studies find adverse effects of IMF programs on unemployment rates and labour income (Chletsos and Sintos 2022; Vreeland 2002) while informal employment in the shadow economy increases (Blanton, Early, and Peksen 2018; Chletsos and Sintos 2021). Through structural reforms, IMF programs also undermine labour rights, unless there are strong institutions that can counterbalance these pressures (Blanton, Blanton, and Peksen 2015; Caraway, Rickard, and Anner 2012; Lee and Woo 2021; Reinsberg *et al*. 2019b). Abundant research finds that IMF programs increase income inequality (Forster *et al*. 2019; Lang 2020; Oberdabernig 2013; Vreeland 2002). This effect is due to policy conditions requiring fiscal restraint, external sector reforms stipulating trade and capital account liberalization, financial sector reforms entailing inflation-control measures, and reforms that restrict external debt, which operate beyond the effects of economic crises that necessitated IMF intervention in

the first place (Forster *et al*. 2019). IMF-mandated reforms have distributional effects that most affect the poorest groups in society. For example, several studies find that IMF programs disproportionally hurt the economic rights of women, with adverse effects on their wellbeing (Detraz and Peksen 2016; Kern, Reinsberg, and Lee 2024; Metinsoy 2022). These gendered effects of IMF programs are mitigated when the political leadership includes women (Reinsberg *et al*. 2023).

As concerns human development more broadly, IMF programs are linked to deteriorating public health and educational outcomes. An analysis of 16 countries in West Africa found that IMF policy reforms reduce government health spending, limiting staff expansion of doctors and nurses and making investments in health systems more difficult (Stubbs *et al*. 2017). In a global sample, the relationship between IMF programs and government health expenditure is negative in all world regions except Sub-Saharan Africa (Kentikelenis, Stubbs, and King 2015). IMF programs adversely affect vaccination rates and child mortality, due to policy conditions that mandate cuts in public-sector health systems (Daoud and Reinsberg 2019; Forster et al. 2020). Moving beyond health, IMF programs have also been associated with decreased public expenditure for education (Stubbs *et al*. 2020). While the Fund has increasingly opted to include pro-poor spending floors, these measures appear to be insufficient to reverse the general contractionary nature of its adjustment loans (Kentikelenis, Stubbs, and King 2016). Moreover, civil society organizations have lamented the strong IMF emphasis on austerity in the wake of the Covid-19 pandemic (Stubbs *et al*. 2022). Nevertheless, Andrijic and Barbic (2021) find that IMF programmes increase people's evaluation of their satisfaction with life in 154 countries between 2005 and 2018, both in the short-term and the long-term.

The pernicious effects of IMF conditions on the provision of public services suggest that these programs could have long-lasting effects on state capacity—the ability of states to deliver public goods throughout their territory. Panel analysis for developing countries confirms this expectation, even when addressing reverse causality (Reinsberg *et al*. 2019). A perverse effect of IMF programs can be to induce corruption, rather than to curb it, for example through creating rent-seeking opportunities in large-scale privatization (Reinsberg *et al*. 2019a). The cutback of state capacities also affects the effective control of corruption (Reinsberg, Kentikelenis, and Stubbs 2021). These studies remind us that the on-the-ground effects of IMF programs may differ from their intended effects. Powerful elites will use these programs to advance their interests. For example, incumbents will implement conditions selectively and with a view to harm the opposition while protecting their own supporters (Reinsberg and Abouharb 2023).

A sizeable literature studies the effects of IMF programs on conflict. There is some evidence that IMF programs increase government instability (Dreher and Gassebner 2012). IMF conditions also increase the likelihood of coups d'etat, as the elites that are not closely linked to the incumbent leader anticipate their economic fortunes to deteriorate. The (re)distributive struggles induced by IMF programs may also affect the likelihood of civil war (Hartzell, Hoddie, and Bauer 2010), though findings are not robust to removing outliers and alternative definitions of war (Midtgaard, Vadlamannati, and de Soysa 2014). Indeed, Vadlamannati, Østmoe, and Soysa (2014) show that IMF interventions reduce ethnic enmity, especially in countries that are highly fractionalized. Focusing on human security, studies find a deterioration of related outcomes in the wake of IMF programs (Abouharb and Cingranelli 2009; Nelson and Wallace 2017; Reinsberg, Shaw, and Bujnoch 2022). Yet, IMF programs appear to enhance procedural democracy: Birchler, Limpach, and Michaelowa (2016) find that when the IMF and World Bank focus their conditions on participative processes and government accountability—as with their poverty reduction programs—they reduce aid fungibility for recipients, and positively affect democratization. A

key challenge of the literature examining the implications of IMF programs on leader survival is selection bias. As rational leaders should undergo IMF programs only if they expect them to prolong their tenure, the interpretation of effects is not straightforward (Kern, Reinsberg, and Shea 2024; Smith and Vreeland 2006; Williams 2012). Emphasizing the endogeneity of IMF programs, recent research suggests that the availability of a lender of last resort and global financial integration create perverse incentives for country elites, who can plunder the wealth of their nations and deposit their fortunes in offshore tax havens before steering their countries into financial disaster and socializing the cost of adjustment upon the wider population *(Kern et al*. 2023).

In sum, the track record of IMF programs is relatively poor for a range of outcomes from economic growth, income inequality, human development, and peace and security. While many of these outcomes are attributable to the policy content of these programs, local elites are by no means innocent bystanders in the adjustment process. A commitment to pro-poor policies and good governance can go a long way to mitigate the adverse (yet unavoidable) adjustment effects of IMF programs. However, as we show below, the vast empirical work on the effectiveness of IMF programs includes many studies that effectively base their conclusions on a small set of countries or years, even when their nominal samples are large.

Case selection and general overview of the data

Our task is to probe the external validity of the findings relating to the effects of IMF programs. We obtained our sample of relevant studies in three steps. First, we created a long list using a keyword search in the Web of Science with terms related to IMF programs. This yielded 613 studies published during 2008 and 2022. Substantively, this period coincides with the renewed interest in the Fund as a lender of last resort in the Global Financial Crisis. Furthermore, we had practical concerns about finding adequate replication material for earlier studies, given that standards of replicability and replication software are commonly expected to have evolved over the past decade. Second, we focused on observational studies using global country-year panel data, because this type of studies is supposed to prioritize external validity of findings. Third, we only kept the papers that had an IMF variable as the treatment in a specified outcome equation. Studies most commonly use a binary variable indicating participation in an IMF program, although some more recent studies also examine the impact of IMF conditions and other aspects of participation in IMF programs. Where models analyse more than one IMF treatment together, we report *REPSTAT* indicators separately for each treatment. These requirements reduced the pool significantly, leaving us with 70 studies. Fourth, we dropped six studies that were published as book chapters or that appeared in non-ranked journals. We decided to do so after it became apparent that replication material was generally unavailable for these studies. We were also concerned that replication of book projects would involve work of an unnecessary magnitude for the scope of this study.

This left us with 64 studies. Out of these studies, we located full replication materials for 14 studies online (21.9%). For the remaining 50 papers, we contacted the authors to request the replication material. We provided authors with information about the aim of our study and how we intend to use the results from the replication exercise. As of the writing of this draft, we obtained responses from the authors of 32 articles (6.0% of the 50), while the remainder 18 did not respond to our request at all. Among those who responded, authors of 15 papers ultimately delivered the full replication files that allowed us to successfully replicate the published findings (bringing the total of replicated studies to 29). Partial replication material was delivered for two more studies. In two cases the authors admitted that no

replicable material was available. Authors of 13 additional studies responded to our requests, but ultimately did not deliver the material. Sometimes an incomplete replication package was due to key variables being subject to proprietary agreements with commercial providers. In other cases, some files got lost during the transition to another IT system. Overall, the outcome of our attempts to merely obtain replication files is disappointing at this preliminary stage but we hope more replication material will be eventually accessible.

In the 29 studies for which full replication material was obtained, 21 of which are ranked within Q1 in their field (overwhelmingly Economics, International Relations and/or Political Science) and 4 in Q2, we have prioritized the models that are reported in the papers, including appendices at their end if any, but in some of them we have also managed to replicate models reported in online supplementary material. While most models have a single treatment variable, some may have more than one treatment, leaving us with 977 treatments in 508 models. The preferred unit of analysis of our study is the individual treatment because we can compute effective samples for each individual treatment variable. Treatments nest in models, which nest in studies. Using this set of studies, we provide results on the external validity of IMF research.

Figure 1 shows that the common treatments in these models are a dummy for participation in any IMF program (402 treatments), the number of IMF conditions that the country must comply with per particular program type (286 treatments), a dummy for participation in a particular IMF program (121 treatments), the total number of IMF conditions that the country must comply with (78 treatments), and total IMF disbursements (50 treatments). Other treatment types include an index for a country's influence on the IMF (16 treatments), the number of years of participation in an IMF program (16 treatments), and amounts of specific IMF lending facilities (8 treatments).[5] Figure 2 shows the frequencies (by number of models) of methods used in the replicated studies. These methods are of curse not mutually exclusive – models may use combinations of them.

[5] Replicated studies that employed specific rather than general treatments commonly distinguished between structural and stabilization conditions, poverty-reduction schemes, or conditions aimed at different sectors in the economy. While our case-selection procedure focused specifically on IMF programs, one of the studies also specified World Bank conditionality and funding as treatments, and we include these treatments in our data.

Figure 1: Types of IMF treatments

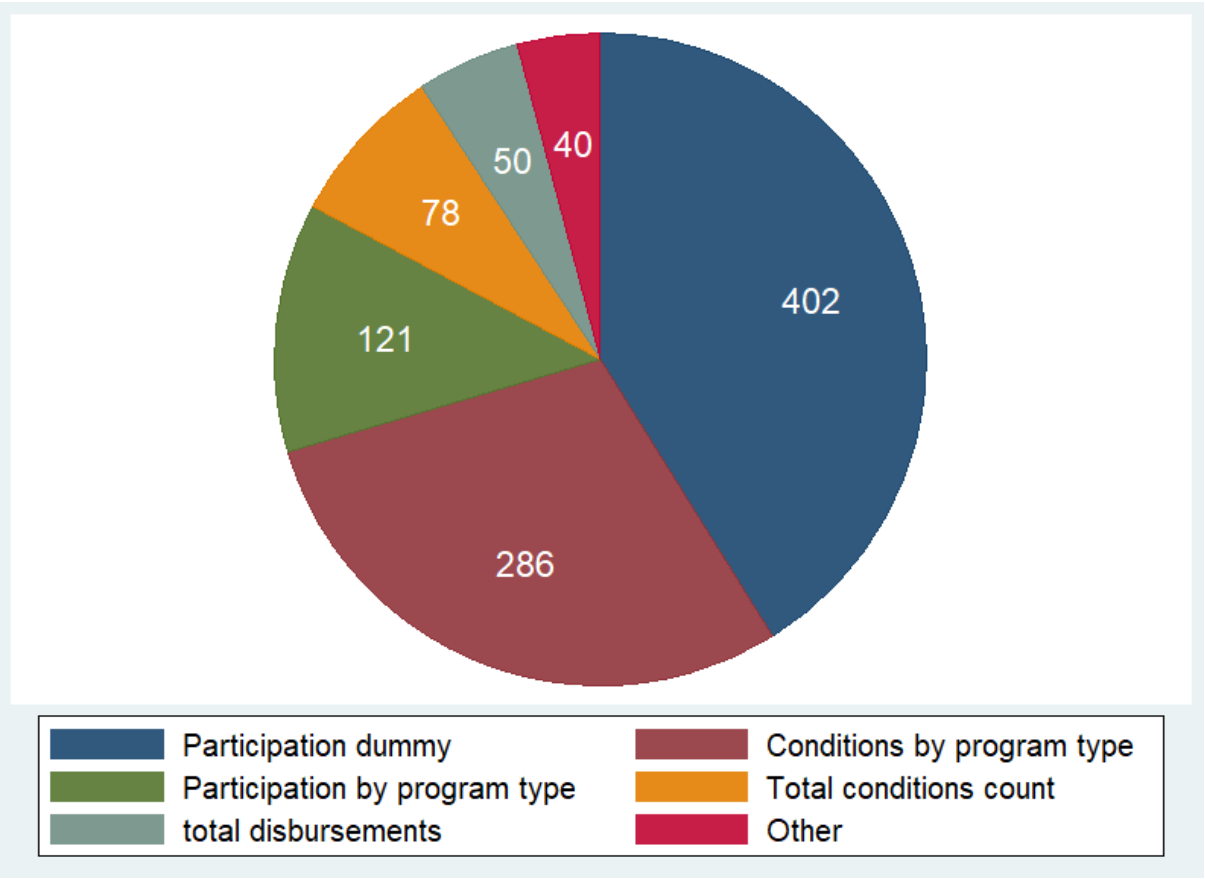
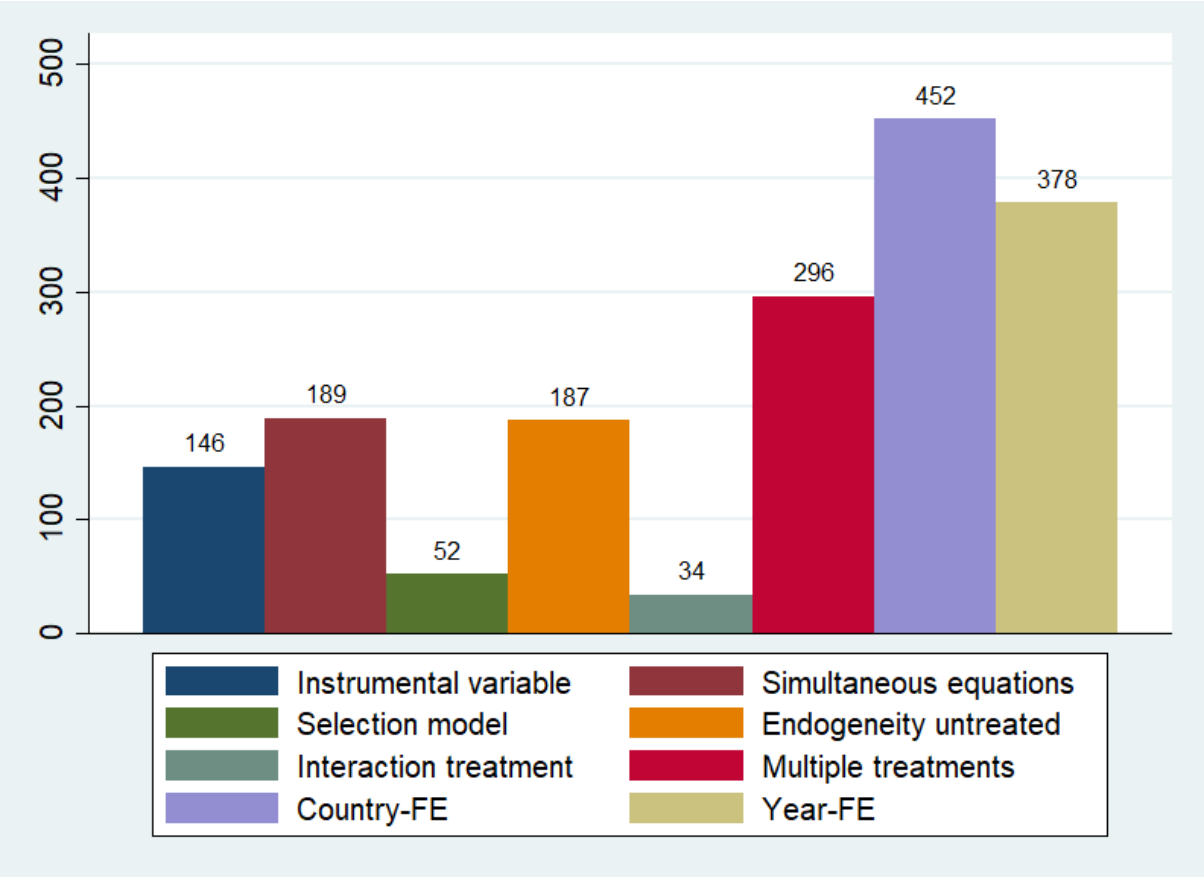


Figure 2: Methods of IMF treatments (model frequencies)



The replicated studies span a variety of outcomes in the IMF literature (See Table B1 in Appendix B). Key issue areas include democracy and the rule of law, economic development, income inequality, health, education, financial crises, civil war, and other social development indictors.

What the data say: external validity in IMF research

Figures 3-4 illustrate the size of effective samples (in absolute number of equally weighted observations - $EffSamp_T$), how they differ from nominal samples, and how studies of IMF effectiveness fair on this matter. Figure 3 shows that nominal and effective samples are only noisily related across all of the replicated studies and all of the treatments and methods that they employ. We next focus on the binary indicator for IMF program participation as the most common treatment in the replicated models. Figure 4 sorts the IMF dummy treatments by the nominal sample of the model in which they were specified, and shows that as the nominal sample declines, sometimes the effective sample does not fall much, or even increases. To ensure that the change in the effective sample results only from changing model specification and dataset, the graphs disregard interactive treatments and distinguish models that account for the endogeneity of treatment (with instrumental variables, simultaneous equations and selection models) from those that do not.

Figure 3: Nominal and effective samples across all replicated studies.

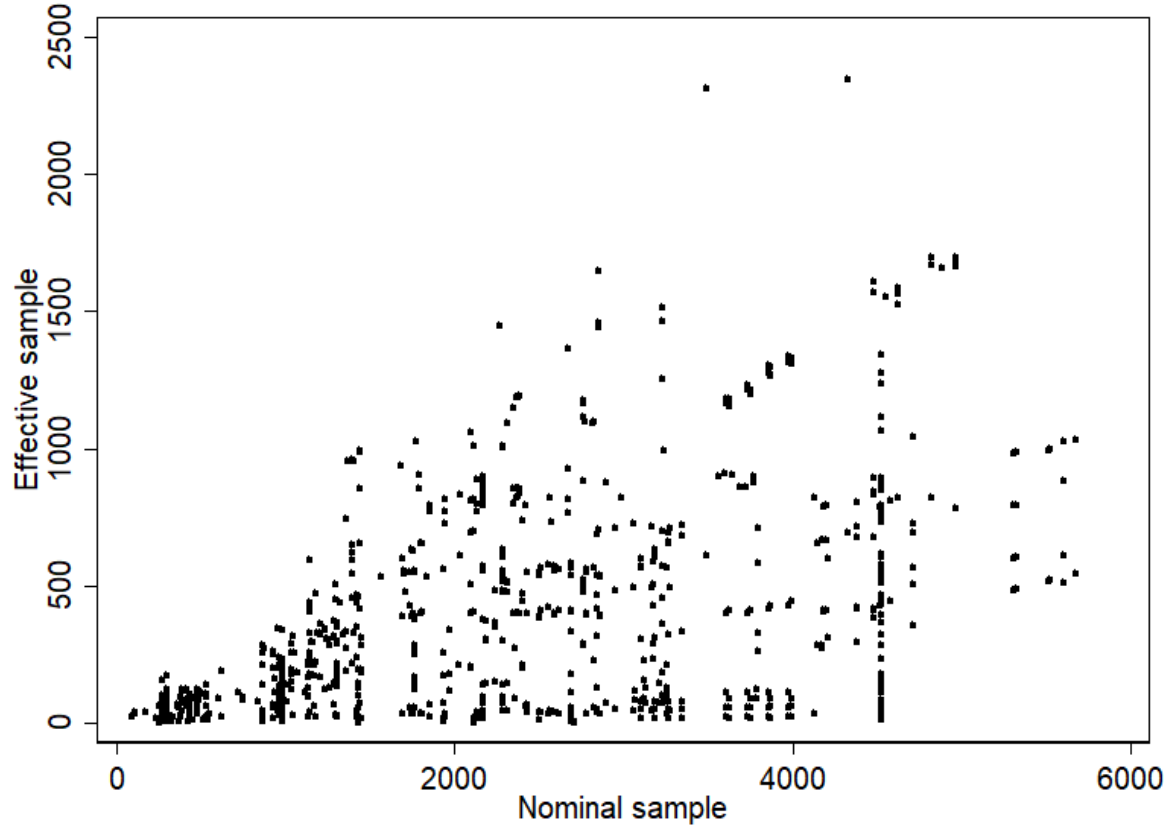


Figure 4: Effective samples sometimes rise when nominal samples decline

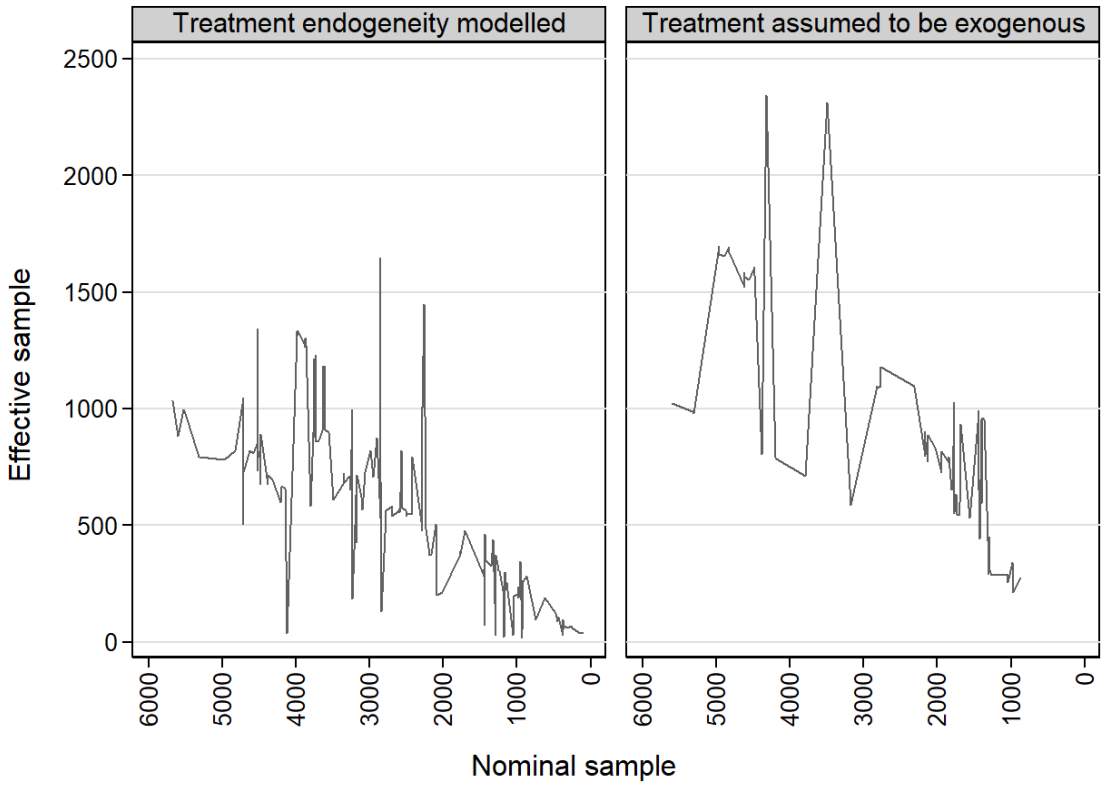
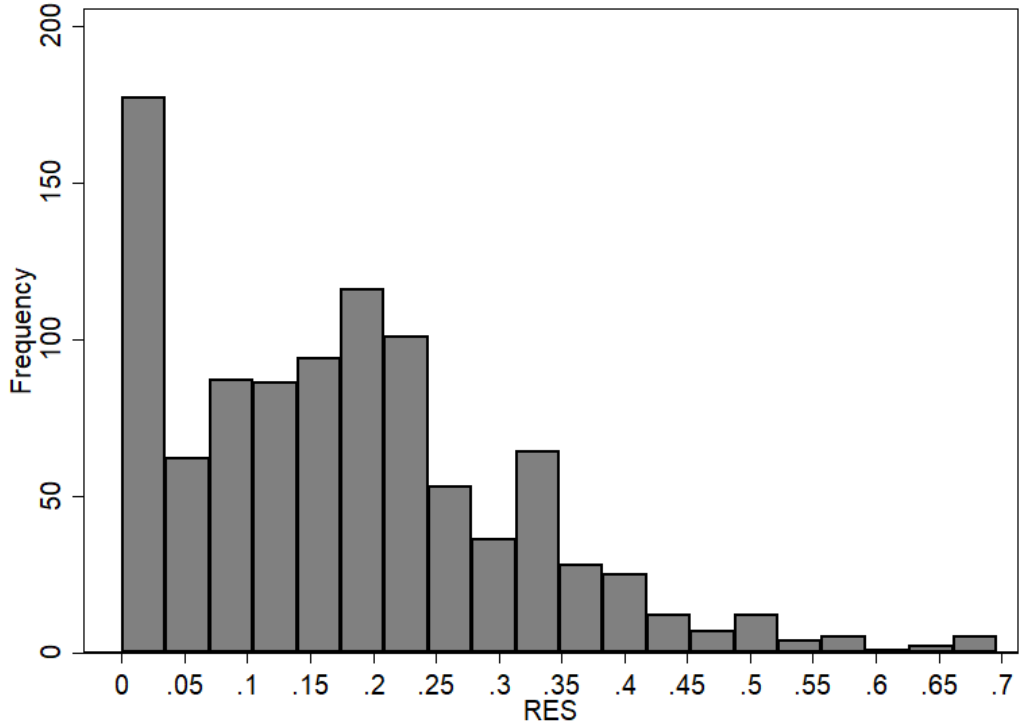
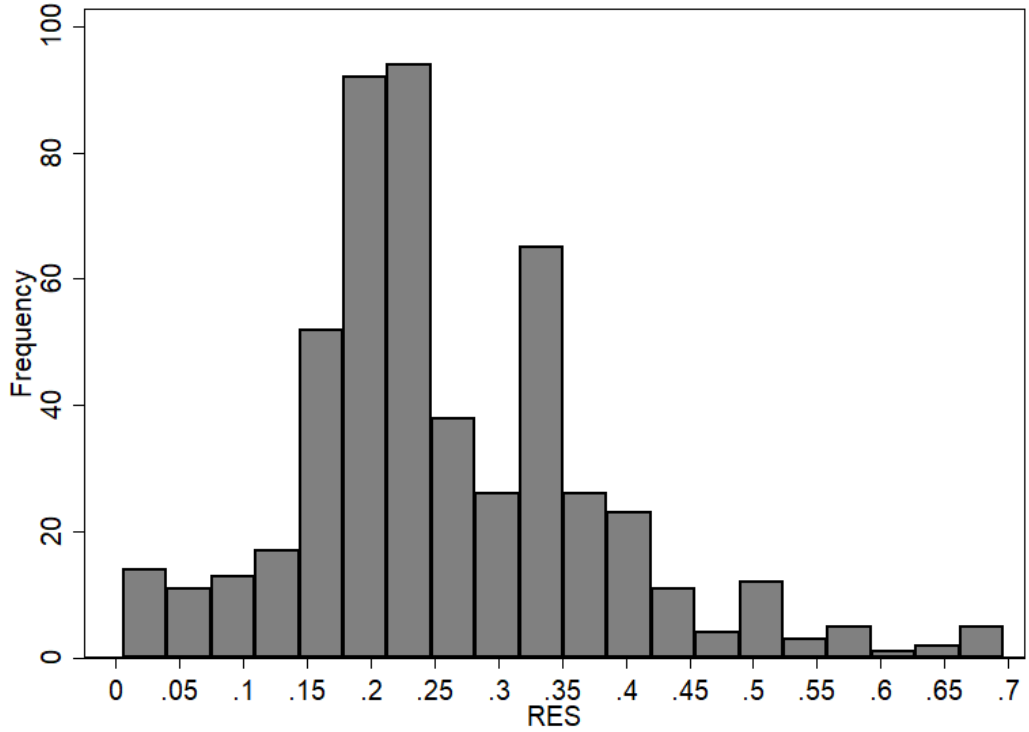


Figure 5: Relative Effective Sample (RES) for the entire dataset.



Note: The figure is based on all treatments using dichotomous variables relating to IMF programs, including dummies for participation in any IMF program and specific program facilities.

Figure 6: Relative Effective Sample (RES) for IMF program dummy variables



Absolute numbers can only tell part of the story. To be sure, the total number of observations is limited by the fixed number of countries and the small number of years during which IMF adjustment programs have been a relevant phenomenon. For some outcomes, data availability may be limited, whereas in other cases, the periodicity of the panel observations—such as the use of five-year periods—reduces the sample size. In these cases, it would be important that the analysis is based on a large share of the available observations, given that their absolute number will be low anyway. Figure 7 shows RES_T – the share of effective sample observations in the nominal (estimation) samples across all models and treatments. Values range between a minimum of 0.0007 and 0.70; the median and average values are both around 17%. However, for almost a quarter of treatments RES<0.07, and only in about 10% of treatments are estimates based on RES>0.35. For example, in one study recently published in one of the most prominent journals in the field, in almost all treatments RES=0.02 and the effective sample size is equivalent to 11 or even only 9 equally-weighted observations. Closer inspection reveals that even though the dataset in that study spans well over 100 countries since the 1980s, the estimates of interest are overwhelmingly driven by the particular cases of Britain (34 percent), Germany (11 percent) and the Netherlands (6 percent), or the particular period of 2005-2010 (44 percent).

Figure 7: Histogram of Relative Effective Sample (RES) for IMF conditionality treatments

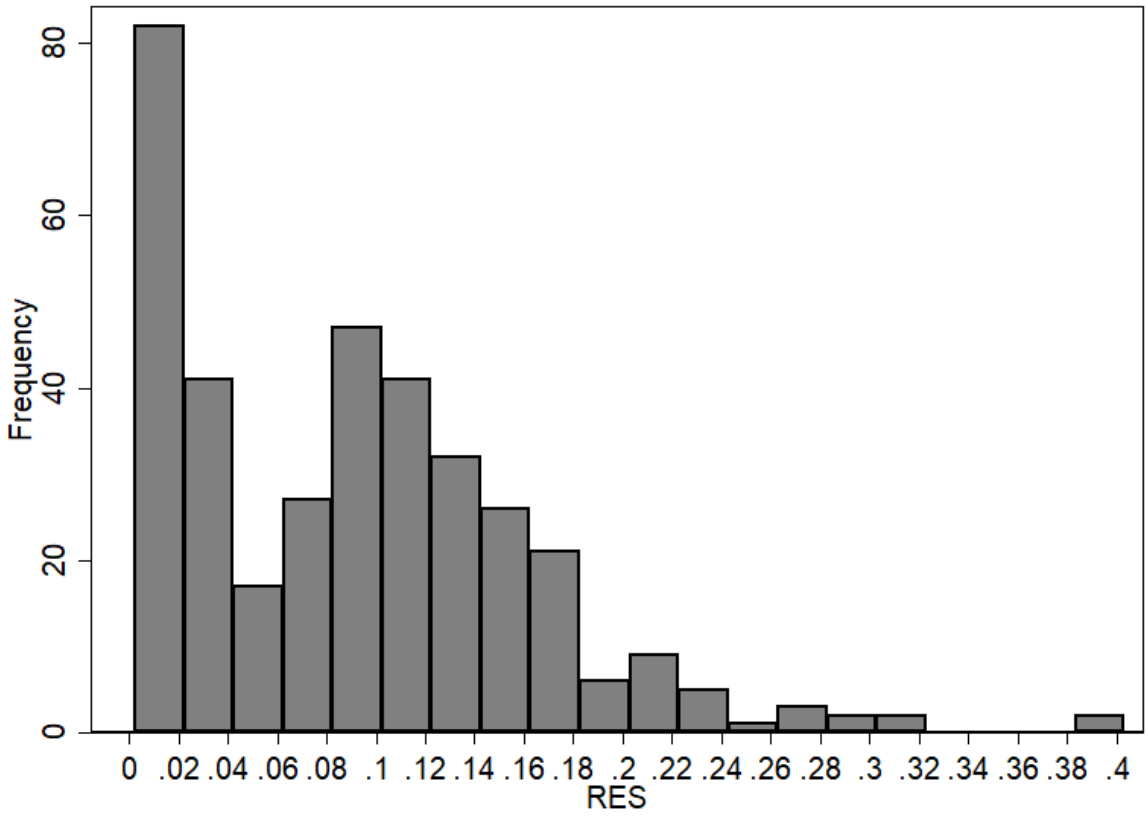
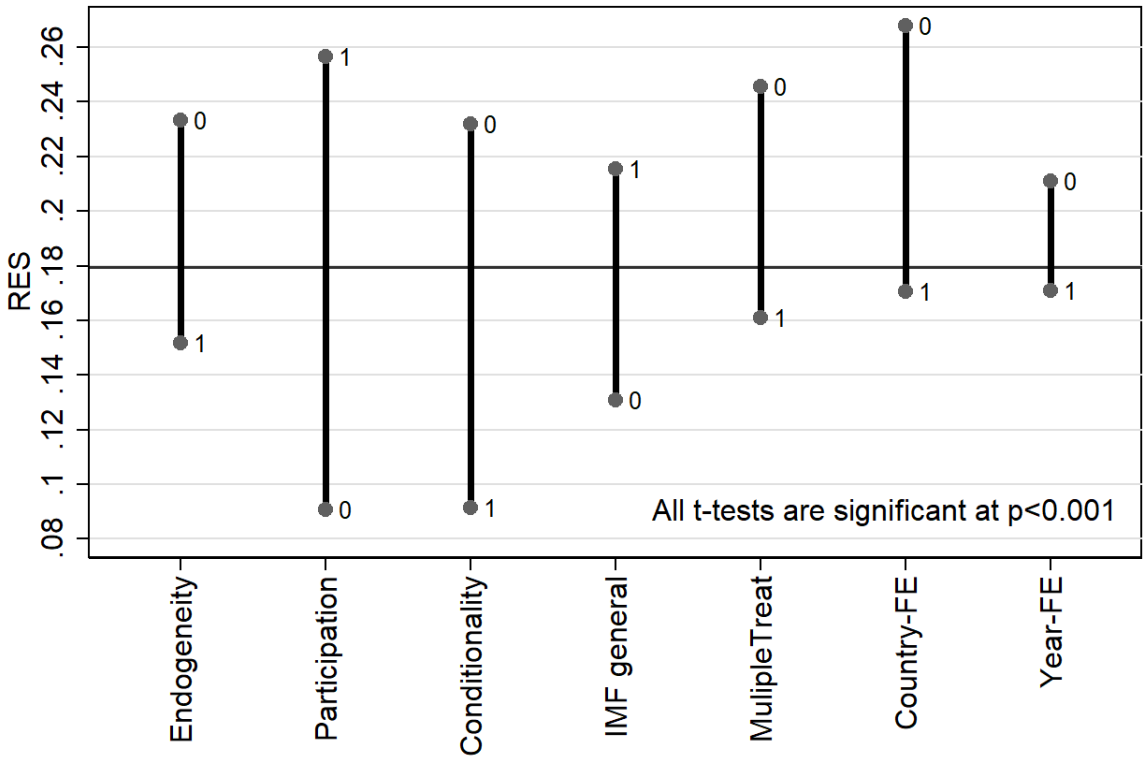


Figure 6 focuses on IMF dummy variables treatments and shows that only in a small percentage of cases the RES drops below 5%, but that estimates for a plurality of treatments still rest on effectively fewer than one-quarter of the observations. 5% of the treatments attain effective samples above 45% of the observations. These few cases mark the best-case external validity scenario in IMF effectiveness research.

Figure 7 scrutinizes the effective sample shares for IMF conditionality treatments. The bulk of the distribution—with a portion of 90%—hovers below the average RES level for the entire data—18%. This clear difference in RES between different types of treatments lead us to explore which model characteristics can potentially explain variation in the RES. Figure 8 illustrates a set of t-tests for differences in means, splitting the data alternatively by whether treatments are modelled as endogenous or assumed to be exogenous, whether they measure participation in IMF program or not, IMF conditionality; the effect of general IMF involvement or specific program-types; whether they are specified with additional treatments in a single model; and whether country and year fixed effects are used. The horizontal line represents the average RES in the entire data. The figure shows that estimates obtained through methods that seek to maximize internal validity—including instrumental-variable analysis, simultaneous-equation models, and selection models—have lower relative effective samples than other models—almost half the relative effective sample.

This is an interesting finding, as it confirms that researchers may face a trade-off between internal validity and external validity. For example, instrumental variable models may help with causal identification but often effectively rest on few observations—a common critique of these models as those are known to identify local average treatment effects. Note however that this is a statistical tendency: some endogenously-modelled IMF dummy treatments reach a RES of 0.64, while some treatments that are regarded as exogenous achieve only RES=0.001. Figure 8 also shows that simple participation treatments enjoy greater RES than conditionality and other types of treatments, and that coding for specific types of IMF programs or specifying multiple treatments or fixed effects of either type are costly in terms of RES. All of these tests are statistically significant at $p<0.001$. As these differences are unadjusted for confounding influences, we turn below to multivariate regression analysis.

Figure 8: Average Relative Effective Sample (RES) for different model features

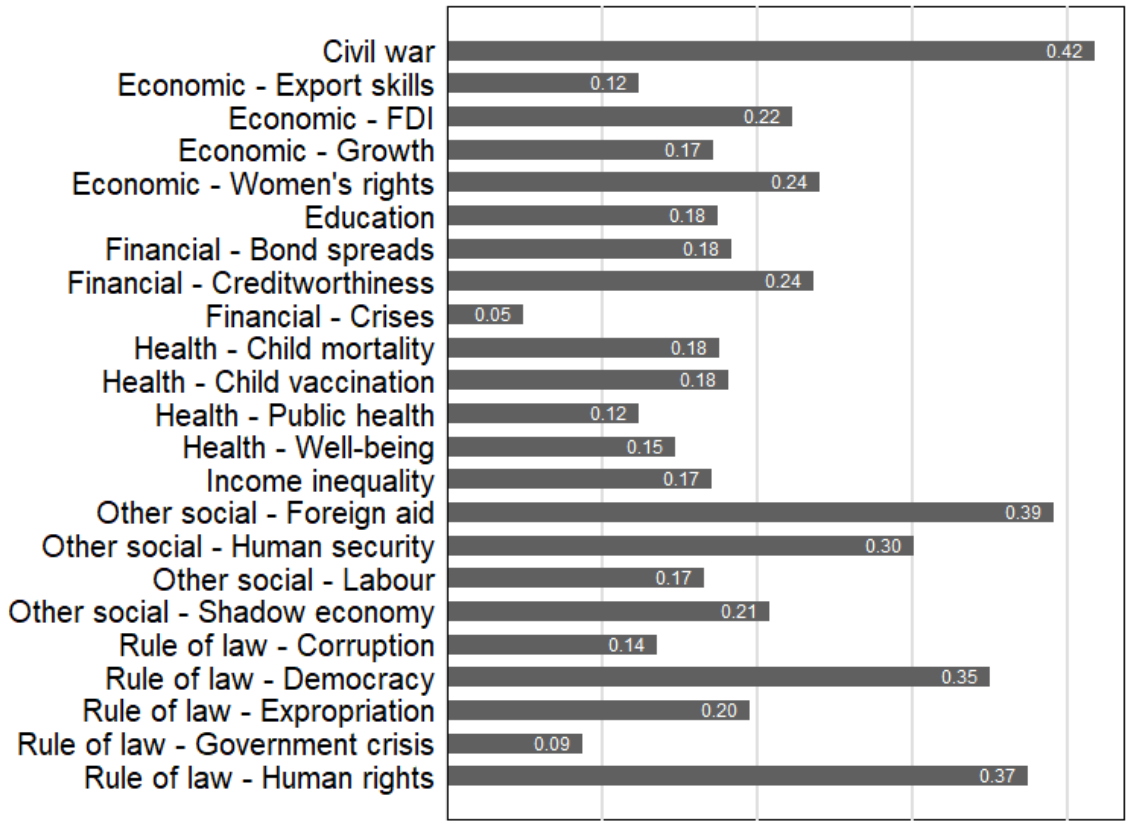


Note: The horizontal line represents the average RES in the entire data. 1=Yes; 0=No.

In Figure 9 we report on average RES levels per field of study. Of course the outcome variable does not directly determine RES (but its particular operationalization may have an indirect impact, by changing the range of observations). So figure 9 does not suggest that some fields should inherently be associated with lower or higher RES. Rather the point of this exercise is to draw the attention of scholars to challenges in inference from existing studies. For example, studies focusing on the IMF's effects on civil wars, foreign aid, human security, democracy and human rights seem to return results that are indeed quite generalizable. In contrast, it appears that studies on financial and government crises could try to make their conclusions less dependent on a small number of country-years that effectively drive their results.[6]

[6] Note that the relatively rare occurrence of crises does not necessarily in itself determine that RES must be low. As demonstrated above, estimation methods, operationalization of treatments, and model specification are important. These shape the added value of non-crisis observations too, which may not be small or even uniform.

Figure 9: Relative Effective Sample (RES) for different outcome types



Using multivariate OLS regression analysis, we now explore some determinants of relative effective sample. Recall that the unit of analysis is a single variable operationalizing IMF involvement (the treatment); treatments nest in the models in which they are specified, which nest in the replicated studies. To avoid confusion in a regression analysis of regressions analyses, our terminology will refer to 'models' that are the subject of our analysis, and various 'regressions' that we run to analyse those models and the treatments within them. Our dependent variable is RES. Our key predictors relate to differences in the ways in which scholars test for the impact of IMF interventions. One important distinction is the type of treatment, which alternatively includes the use of a global dummy for participation in any type of IMF program, dummies for specific programs, and count variables for the number of conditions in the IMF programs. Some scholars have also used a count variable for the years of exposure to IMF programs, while others have constructed measures of influence.[7] Following our earlier insights, we add dummies indicating whether the original model in which each treatment was specified included country-fixed effects and/or year-fixed effects. Note that none of our own estimated regressions reported below include fixed effects themselves. We further control for the (logged) total number of confounders in the analysed model in which the treatment was specified, and the nominal size of that model's sample. In addition to this baseline set of controls, we add variables for specific methodological choices in the analysed models, notably whether their methodology includes any of the inferential methods boosting internal validity. We also construct a dummy for the use of an interaction treatment term, which allows for estimating conditional

[7] We combine some of these variables, as a too fine-grained breakdown would rely on very few observations each.

relationships of IMF treatments. Because such variables partition the predictor space, we would naturally expect a drop in effective observations. Furthermore, we consider whether the analysed model includes multiple IMF treatments, such as both an IMF dummy and a count variable for IMF conditionality, or a battery of dummies for participating in particular types of programs. Finally, we elicit whether the analysed model and its treatments were part of a study from a field other than Political Science, International Political Economy, or Economics. This could be meaningful if standards of acceptable practice differ across disciplines. Variable definitions and descriptive statistics of all the variables can be found in Appendix B (Table B2).

Table 1 shows our main results using multivariate linear regression and standard errors clustered on models. Consistently across all models and treatments, we find that studies which rely on a generic IMF variable have a significantly higher relative effective sample, compared to studies using more specific IMF variables. The effect is sizeable: studies with such an IMF variable have an at least 8 percent larger relative effective sample than studies with more specific IMF variables. We also calculated the RES for the variables in our own regressions and found that importantly, this result is also effectively based on a large number of observations, given that the RES for the Generic IMF variable ranges between 0.73 and 0.83 throughout the regressions in Table 1.

The conditional correlations with the control variables are equally insightful. In particular, we confirm that studies using country-fixed effects have lower effective samples, while year-fixed effects are not systematically related to the effective sample. These findings are themselves based on relatively moderately-sized effective samples, with RES=0.13 for country-FEs and RES=0.31 for year-FE. Neither the number of confounders in a model, nor the size of its nominal sample are consistently related to the effective sample. However, studies using causal-inference techniques have a significantly lower effective sample—at least 8.3 percent compared to observational studies ($p<0.01$). This result itself is based on effective samples of about 40% of all observations, which we regard as high, considering that it compares with more than one standard deviation above the average RES of 0.18 in our data (but of course the RES calculations for variables in our own regressions are not part of our dataset of replicated studies). Equally taxing are interaction terms, with a predicted reduction in the effective sample by at least 8.8 percent ($p<0.01$). However, this result is effectively based on 5% of all observations, so it's hard to make general inferences based on it. Models with multiple treatments reduce the relative effective sample by 5.6 percent ($p<0.01$), while studies in non-IPE journals appear to have a slightly higher effective sample. Both results have good external validity, with a RES above 0.30.

Table 2 allows us to assess the effective samples of estimates for IMF conditionality, given that scholars are often interested in isolating the effects of IMF policy conditions versus other dimensions of IMF program assistance. In line with our descriptive findings earlier, studies probing the impact of IMF conditionality have significantly lower relative effective samples. Coefficient magnitudes are sizeable: IMF conditionality coefficients effectively rely on at least 12.4 percent fewer observations compared to models with an IMF dummy or other treatments that do not operationalize conditionality ($p<0.01$). The result is also based on a large effective sample. The RES is as high as 0.67 in the first model, and 0.64 in the fully specified model in the last column.

Table 3 digs deeper into the impact of using different methods of causal inference for effective samples. We have already found that these methods—motivated by attempts to enhance internal validity—come at the price of external validity. We now find more specifically that the use of instrumental variables significantly reduces effective sample size by about 3 percent ($p<0.05$). Simultaneous-equation estimators do not significantly reduce effective samples, while selection models are related to reduced samples by

up to 12 percent ($p<0.01$). This confirms our earlier finding suggesting that researchers trade off internal validity for external validity when employing causal inference designs. However, some approaches appear to navigate the tradeoff better than others. In terms of the external validity of these findings, we find that coefficients for instrumental variables and simultaneous equations are based on moderately large samples (RES=0.30), while the coefficient for selection models has a lower representativeness (RES=0.11).

We probe the robustness of results and conduct further analyses in Appendix B. First, we are interested in whether contextual variables related to publication practices matter. To that end, we measure whether the published estimate is significant at the 5%-level and respectively whether it features in an appendix at the end of the paper or in online supplemental material. These contextual variables are not consistent across all specifications. Yet, where significant, they suggest that statistically significant estimates appear to be based on fewer effective observations, while treatments presented in appendices have higher effective samples (Table B3). Both findings have good external validity themselves, with relative effective samples of up to 40%.

In addition, we probe the robustness of our findings to an alternative way to cluster standard errors, using either robust standard errors or clustering by study. While our results are unaffected with robust standard errors, study-clustered standard errors inflate the standard errors in some variables. Importantly, we continue to find evidence that generic IMF variables are associated with larger effective samples, whereas estimates obtained from estimators that privilege internal validity significantly depress the effective sample (Table B4).

We also use an alternative specification of the main dependent variable which addresses concerns about non-linearity. In particular, while our effective sample share is bounded to the unit interval, using linear regression on the untransformed RES might generate predictions out of these bounds. There is also significant bunching at the lower end of the distribution of relative effective sample sizes. To address these problems, we use a logistic transformation of RES, computed as $ln(RES/(1-RES))$. As our results are qualitatively unaffected by this transformation, we prefer using linear models on the raw shares for ease of interpretation in our main analysis (Table B5).

Overall, we established huge variation in relative effective samples—as a measure of external validity—in studies on the effectiveness of IMF interventions. We considered a range of model features to explain variation in effective sample sizes. We found that the use of a generic IMF variable—such as a IMF program participation dummy—is associated with larger effective samples than the use of specific IMF variables—such as for specific lending facilities. In contrast, instrumental variables and selection models engender a significant reduction in the effective sample. We also analysed how representative our own findings were by calculating *REPSTAT* for these variables (Figure 11). With few exceptions, our estimates have high external validity, as they are generally based on large effective samples.

Table 1: Determinants of relative effective sample sizes

	(1)		(2)		(3)		(4)		(5)	
Generic IMF variable	0.077***	(0.011)	0.080***	(0.010)	0.085***	(0.010)	0.075***	(0.010)	0.076***	(0.010)
Confounders	-0.019*	(0.010)	0.009	(0.010)	0.010	(0.010)	0.022**	(0.011)	0.020*	(0.011)
Nominal sample	-0.002	(0.005)	0.007	(0.005)	0.006	(0.005)	0.010**	(0.005)	0.015***	(0.005)
Country-FE	-0.081***	(0.023)	-0.086***	(0.021)	-0.091***	(0.021)	-0.086***	(0.020)	-0.090***	(0.020)
Year-FE	-0.021	(0.013)	0.017	(0.015)	0.017	(0.015)	0.028*	(0.016)	0.028*	(0.016)
Causal inference approach			-0.083***	(0.013)	-0.084***	(0.014)	-0.085***	(0.013)	-0.101***	(0.014)
Interaction					-0.088***	(0.020)	-0.100***	(0.023)	-0.096***	(0.023)
Multiple treatments							-0.056***	(0.013)	-0.056***	(0.013)
Non-IPE journal									0.031***	(0.006)
Observations	977		977		977		977		977	
R-squared	0.144		0.206		0.222		0.246		0.254	

Notes: Linear regression estimated via Ordinary Least Squares. Standard errors clustered on models in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.

Table 2: Probing effective samples of models with IMF conditionality

	(1)		(2)		(3)		(4)		(5)	
IMF conditionality	-0.133***	(0.008)	-0.128***	(0.008)	-0.133***	(0.008)	-0.124***	(0.009)	-0.124***	(0.009)
Confounders	-0.003	(0.007)	0.019**	(0.008)	0.020***	(0.008)	0.028***	(0.009)	0.027***	(0.009)
Nominal sample	0.003	(0.005)	0.010*	(0.005)	0.010*	(0.005)	0.012**	(0.006)	0.017***	(0.006)
Country-FE	-0.047**	(0.022)	-0.053**	(0.022)	-0.058**	(0.023)	-0.056***	(0.021)	-0.060***	(0.021)
Year-FE	-0.006	(0.013)	0.025*	(0.014)	0.025*	(0.015)	0.032**	(0.016)	0.033**	(0.016)
Causal inference approach			-0.069***	(0.011)	-0.069***	(0.012)	-0.072***	(0.012)	-0.085***	(0.013)
Interaction					-0.089***	(0.024)	-0.098***	(0.026)	-0.095***	(0.026)
Multiple treatments							-0.040***	(0.013)	-0.041***	(0.013)
Non-IPE journal									0.027***	(0.008)
Observations	977		977		977		977		977	
R-squared	0.271		0.314		0.330		0.343		0.349	

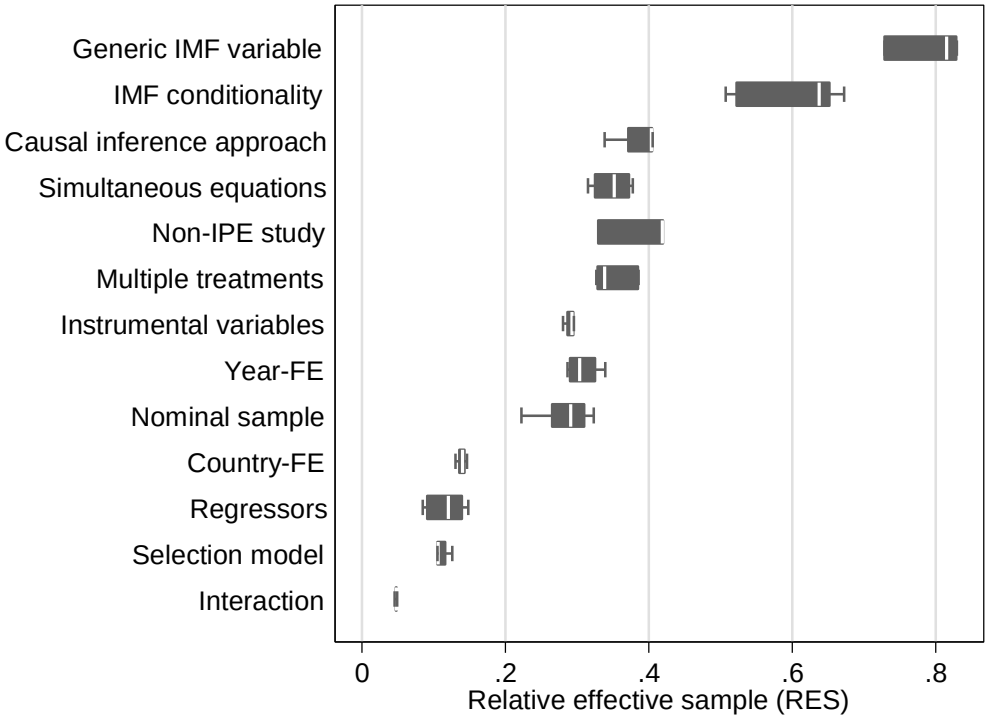
Notes: Linear regression estimated via Ordinary Least Squares. Standard errors clustered on models in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.

Table 3: Probing effective samples for models with different causal inference approaches

	(1)		(2)		(3)		(4)		(5)	
Instrumental variables	-0.034**	(0.014)	-0.030**	(0.014)	-0.027*	(0.014)	-0.030**	(0.014)	-0.035**	(0.015)
Simultaneous equations	-0.016*	(0.008)	-0.009	(0.014)	-0.016	(0.015)	-0.006	(0.015)	-0.022	(0.018)
Selection model	-0.090***	(0.014)	-0.095***	(0.015)	-0.093***	(0.016)	-0.105***	(0.015)	-0.118***	(0.015)
IMF conditionality	-0.149***	(0.008)	-0.144***	(0.008)	-0.146***	(0.008)	-0.140***	(0.009)	-0.139***	(0.009)
Confounders			-0.006	(0.010)	-0.001	(0.011)	0.002	(0.011)	0.004	(0.011)
Nominal sample			-0.004	(0.007)	-0.003	(0.007)	-0.003	(0.007)	0.001	(0.008)
Country-FE			-0.042*	(0.022)	-0.048**	(0.023)	-0.044**	(0.021)	-0.047**	(0.021)
Year-FE			-0.002	(0.016)	0.001	(0.017)	0.003	(0.016)	0.007	(0.017)
Interaction					-0.086***	(0.024)	-0.093***	(0.026)	-0.094***	(0.026)
Observations	977		977		977		977		977	
R-squared	0.289		0.298		0.313		0.328		0.331	

Notes: Linear regression estimated via Ordinary Least Squares. Standard errors clustered on models in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.

Figure 10: Distribution of relative effective samples for the key variables in this study



Note: The figure uses RES statistics of all estimates presented in the main text

Conclusion

We proposed a new statistic to establish the external validity of treatments. Building on research in political methodology, we argued that the effective number of observations relative to the nominal sample is a good indicator of external validity. It is easy to compute after estimation, easy to understand, and sufficiently flexible as it allows to distinguish the representativeness of an estimand at the level of individual treatments. We demonstrated the usefulness of this statistic through a replication exercise of 997 treatments in 508 models from 29 studies on the effectiveness of IMF program interventions.

We established huge variation in effective sample size for specific IMF treatments. These can occur within the same article and even within the same model. In general, external validity in this field is relatively poor. About three quarters of the IMF program treatment estimates effectively rely on fewer than one-quarter of the observations. No study had an effective sample size over 70 percent. In one study recently published in one of the most prominent journals in the field, the headline result is effectively based on only 9 observations, and overwhelmingly driven by the particular cases of one or two countries, or a single time period. These findings cast some doubt about the external validity of the published findings. As a side note, we are also disappointed by the lower-than-expected standards of replication, even in very recent studies and sometimes even in the top journals in the field.

In a second step, we examined the determinants of effective sample size. We found that models using an IMF generic treatment (disregarding program type) generally benefited from enhanced external validity. Treating the outcome with measures of conditionality, adding country fixed effects and having multiple treatments in a single model reduced relative effective sample. Importantly, attempts to about for the endogeneity of the treatment, especially when using selection models, significantly decreased the relative effective sample, suggesting a trade-off between internal validity and external validity.

We make the following practical suggestions for large-N observational studies of the effectiveness of IMF interventions. First, scholars should be aware that the more specific the IMF treatment in the model (e.g. participation in specific program, type and count of conditionality, interacted treatment etc.) the lower external validity of their findings is likely to be relative to a simple IMF participation dummy. In such cases we recommend that RES and EffSamp statistics be reported alongside the conventional *t*-tests for the treatment's coefficient. If external validity indeed is low, the importance of the findings (when hypotheses are either supported or refuted – the significance of the estimates is not associated with our indicators) should be discussed.

Second, as large-*n* studies aspire to greater external validity, when choosing model specification and operationalization of IMF treatment scholars should consider the size of the effective sample alongside proper causal identification. If there is a trade-off between internal and external validity of findings, scholars would do well to report models that alternatively prioritize each and discuss the difference in findings. Better still, optimal models should aspire to minimize the cost in external validity associated with good causal identification.

Our results extend the findings of the IMF effectiveness literature in a critical dimension. Over the past decades, researchers have been vexed by the challenge of identifying the treatment effect of IMF programs. The focus on internal validity has relegated external validity to a neglected second-order issue. Consequently, our knowledge about to what extent the findings from this research program generalize across different countries and different time periods has remained extremely limited. We begin to fill this gap and suggest that researchers face a critical trade-off: the promise of enhanced internal validity comes at the cost of external validity.

Literature

Abouharb, M. Rodwan, and David L Cingranelli. 2009. "IMF Programs and Human Rights, 1981--2003." *Review of International Organizations* 4: 47--72.

Andrijić, Marijana, and Tajana Barbić. 2021. "When the Going Gets Tough… the Effect of Economic Reform Programmes on National Well-Being." *Sustainablility* 13(20): 11557.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1): 250–67.

Barro, Robert J, and Jong-Wha Lee. 2005. "IMF Programs: Who Is Chosen and What Are the Effects?" *Journal of Monetary Economics* 52(7): 1245–69. http://linkinghub.elsevier.com/retrieve/pii/S0304393205000851.

Bas, Muhammet A, and Randall W Stone. 2014. "Adverse Selection and Growth under IMF Programs." *Review of International Organizations* 9(1): 1–28. http://link.springer.com/10.1007/s11558-013-9173-1.

Biglaiser, Glen, and Karl DeRouen. 2010. "The Effects of IMF Programs on US Foreign Direct Investment in the Developing World." *Review of International Organizations* 5: 73–95.

Biglaiser, Glen, Hoon Lee, and Joseph L. Staats. 2016. "The Effects of the IMF on Expropriation of Foreign Firms." *Review of International Organizations* 11(1): 1–23.

Biglaiser, Glen, and Ronald J McGauvran. 2022. "The Effects of IMF Loan Conditions on Poverty in the Developing World." *Journal of International Relations and Development* (0123456789).

Birchler, Kassandra, Sophia Limpach, and Katharina Michaelowa. 2016. "Aid Modalities Matter: The Impact of Different World Bank and IMF Programs on Democratization in Developing Countries." *International Studies Quarterly* 60(3): 427–39.

Bird, Graham, and Dane Rowlands. 2002. "Do IMF Programmes Have a Catalytic Effect on Other International Capital Flows?" *Oxford Development Studies* 30(3): 229–49.

Blanton, Robert G., Bryan Early, and Dursun Peksen. 2018. "Out of the Shadows or into the Dark? Economic Openness, IMF Programs, and the Growth of Shadow Economies." *Review of International Organizations* 13(2): 309–33.

Blanton, Robert G, Shannon Lindsey Blanton, and Dursun Peksen. 2015. "The Impact of IMF and World Bank Programs on Labor Rights." *Political Research Quarterly* 68(2): 324–36.

Boockmann, Bernhard, and Axel Dreher. 2003. 19 European journal of political economy *The Contribution of the IMF and the World Bank to Economic Freedom*. http://linkinghub.elsevier.com/retrieve/pii/S0176268003000168.

Breen, Michael, and Patrick J.W. Egan. 2019. "The Catalytic Effect of IMF Lending: Evidence from Sectoral FDI Data." *International Interactions* 45(3): 447–73.

Brown, Kathleen. 2023. *IMF Survival Instincts: Risk Exposure and the Design of Loan Programs*. Toulouse.

Caraway, Teri L, Stephanie J Rickard, and Mark S Anner. 2012. "International Negotiations and Domestic Politics: The Case of IMF Labor Market Conditionality." *International Organization* 66(1): 27–61. http://journals.cambridge.org/abstract_S0020818311000348.

Chapman, Terrence, Songying Fang, Xin Li, and Randall W. Stone. 2015. "Mixed Signals: IMF Lending and Capital Markets." *British Journal of Political Science* 47(2): 329–49.

Chletsos, Michael, and Andreas Sintos. 2021. "Hide and Seek: IMF Intervention and the Shadow Economy." *Structural Change and Economic Dynamics* 59: 292–319.

———. 2022. "The Effects of IMF Conditional Programs on the Unemployment Rate." *European Journal of Political Economy* 102272.

Cho, Hye Jee. 2014. "Impact of IMF Programs on Perceived Creditworthiness of Emerging Market Countries: Is There a 'Nixon-Goes-to-China' Effect?" *International Studies Quarterly* 58(2): 308–21.

Chwieroth, Jeffrey M. 2014. "Professional Ties That Bind: How Normative Orientations Shape IMF Conditionality." *Review of International Political Economy* 22(4): 757–87. http://www.tandfonline.com/doi/full/10.1080/09692290.2014.898214.

Daoud, Adel, and Bernhard Reinsberg. 2019. "Structural Adjustment, State Capacity and Child Health: Evidence from IMF Programmes." *International Journal of Epidemiology* 48(2): 445–54.

David, Antonio C., Jaime Guajardo, and Juan F. Yépez. 2022. "The Rewards of Fiscal Consolidations: Sovereign Spreads and Confidence Effects." *Journal of International Money and Finance* 123(102602).

Demir, Firat. 2022. "IMF Conditionality, Export Structure and Economic Complexity: The Ineffectiveness of Structural Adjustment Programs." *Journal of Comparative Economics* (August 2021). https://doi.org/10.1016/j.jce.2022.04.003.

Detraz, Nicola, and Dursun Peksen. 2016. "The Effect of IMF Programs on Women's Economic and Political Rights." *International Interactions* 42(1): 81–105.

Dreher, Axel. 2006. "IMF and Economic Growth: The Effects of Programs, Loans, and Compliance

with Conditionality." *World Development* 34(5): 769–88.

Dreher, Axel, and Martin Gassebner. 2012. "Do IMF and World Bank Programs Induce Government Crises? An Empirical Analysis." *International Organization* 66(02): 329–58. http://www.journals.cambridge.org/abstract_S0020818312000094.

Dreher, Axel, and Nathan M Jensen. 2007. "Independent Actor or Agent? An Empirical Analysis of the Impact of U.S. Interests on International Monetary Fund Conditions." *Journal of Law and Economics* 50(1): 105–24. http://www.jstor.org/stable/10.1086/508311.

Dreher, Axel, Jan-Egbert Sturm, and James R Vreeland. 2009. "Global Horse Trading: IMF Loans for Votes in the United Nations Security Council." *European Economic Review* 53(7): 742–57. http://www.sciencedirect.com/science/article/pii/S0014292109000312.

Dreher, Axel, and Stefanie Walter. 2010. "Does the IMF Help or Hurt? The Effect of IMF Programs on the Likelihood and Outcome of Currency Crises." *World Development* 38(1): 1–18.

Easterly, William. 2005. "What Did Structural Adjustment Adjust?: The Association of Policies and Growth with Repeated IMF and World Bank Adjustment Loans." *Journal of Development Economics* 76(1): 1–22.

Edwards, Martin S. 2006. "Signalling Credibility? The IMF and Catalytic Finance." *Journal of International Relations and Development* 9: 27–52.

Forster, Timon et al. 2019. "How Structural Adjustment Programs Affect Inequality: A Disaggregated Analysis of IMF Conditionality, 1980–2014." *Social Science Research* 80: 83–113.

Forster, Timon, Alexander Kentikelenis, Thomas Stubbs, and Lawrence King. 2020. "Globalization and Health Equity: The Impact of Structural Adjustment Programs on Developing Countries. Social Science & Medicine." *Social Science & Medicine* 267(112496).

Gehring, Kai, and Valentin F Lang. 2020. "Stigma or Cushion? IMF Programs and Sovereign Creditworthiness." *Journal of Development Economics* 146: https://doi.org/10.1016/j.jdeveco.2020.102507.

Goes, Iasmin. 2023. "Examining the Effect of IMF Conditionality on Natural Resource Policy." *Economics and Politics* 35(1): 227–85.

Hartzell, Caroline A., Matthew Hoddie, and Molly Bauer. 2010. "Economic Liberalization via IMF Structural Adjustment: Sowing the Seeds of Civil War?" *International Organization* 64(2): 339–56.

Kaplan, Stephen B, and Sujeong Shim. 2021. *Global Contagion and IMF Credit Cycles: A Lender of Partial Resort?* Washington D.C.

Kentikelenis, Alexander E, Thomas H Stubbs, and Lawrence P King. 2015. "Structural Adjustment and Public Spending on Health: Evidence from IMF Programs in Low-Income Countries." *Social Science and Medicine* 126: 169–76. http://www.sciencedirect.com/science/article/pii/S0277953614008351.

———. 2016. "IMF Conditionality and Development Policy Space, 1985-2014." *Review of International Political Economy* 23(4): 543–82.

Kern, Andreas, Elias Nosrati, Bernhard Reinsberg, and Dilek Sevinc. 2023. "Crash for Cash: Offshore Financial Destinations and IMF Programs." *European Journal of Political Economy* 102359.

Kern, Andreas, Bernhard Reinsberg, and Claire Lee. 2024. "The Unintended Consequences of IMF Programs: Women Left behind in the Labor Market." *Review of International Organizations*.

Kern, Andreas, Bernhard Reinsberg, and Patrick E. Shea. 2024. "Why Cronies Don't Cry? IMF Programs, Chinese Lending, and Leader Survival." *Public Choice*: https://doi.org/10.1007/s11127-023-01114-4.

Lang, Valentin F. 2020. "The Economics of the Democratic Deficit: The Effect of IMF Programs on Inequality." *Review of International Organizations*: forthcoming.

Lang, Valentin F, and Andrea F Presbitero. 2018. "Room for Discretion? Biased Decision-Making in International Financial Institutions." *Journal of Development Economics* 130: 1–16.

Lee, Su-Hyun, and Byungwon Woo. 2021. "IMF= I'M Fired! IMF Program Participation, Political Systems, and Workers' Rights." *Political Studies* 69(3): 514–37.

Lipscy, Phillip Y., and Haillie Na-Kyung Lee. 2019. "The IMF as a Biased Global Insurance Mechanism: Asymmetrical Moral Hazard, Reserve Accumulation, and Financial Crises." *International Organization* 73(1): 35–64.

Lombardi, Domenico, and Ngaire Woods. 2008. "The Politics of Influence: An Analysis of IMF Surveillance." *Review of International Political Economy* 15(5): 711–39. http://www.tandfonline.com/doi/abs/10.1080/09692290802418724.

Marchesi, Silvia, and Emanuela Sirtori. 2011. "Is Two Better than One? The Effects of IMF and World Bank Interaction on Growth." *Review of International Organizations* 6: 287–306.

Metinsoy, Saliha. 2022. *The IMF, Labor Market Reform, and Labor Market Outcomes for Women: Overlaps between Gender and Economic Inequality*. Groningen.

Midtgaard, Trude M., Krishna Chaitanya Vadlamannati, and Indra de Soysa. 2014. "Does the IMF Cause Civil War? A Comment." *Review of International Organizations* 9(1): 107–24.

Nelson, Stephen C., and Geoffrey P.R. Wallace. 2017. "Are IMF Lending Programs Good or Bad for Democracy?" *Review of International Organizations* 12(4): 523–58. http://dx.doi.org/10.1007/s11558-016-9250-3.

Oberdabernig, Doris A. 2013. "Revisiting the Effects of IMF Programs on Poverty and Inequality." *World Development* 46: 113–42.

Pinheiro, Diogo, Jeffrey M Chwieroth, and Alexander Hicks. 2015. "Do International Non-Governmental Organizations Inhibit Globalization? The Case of Capital Account Liberalization in Developing Countries." *European Journal of International Relations* 21(1): 146–70. http://journals.sagepub.com/doi/10.1177/1354066114523656.

Przeworski, Adam, and James R Vreeland. 2000. "The Effect of IMF Programs on Economic Growth." *Journal of Development Economics* 62: 385–421.

Reinsberg, Bernhard, and M. Rodwan Abouharb. 2023. "Partisanship, Protection, and Punishment: How Governments Affect the Distributional Consequences of International Monetary Fund Programs." *Review of International Political Economy* 30(5): 1851–79.

Reinsberg, Bernhard, Alexander Kentikelenis, and Thomas Stubbs. 2021. "Creating Crony Capitalism: Neoliberal Globalization and the Fueling of Corruption." *Socio-Economic Review* 19(2): 607–34.

Reinsberg, Bernhard, Alexander Kentikelenis, Thomas Stubbs, and Lawrence King. 2019. "The World System and the Hollowing Out of State Capacity: How Structural Adjustment Programs Affect Bureaucratic Quality in Developing Countries." *American Journal of Sociology* 124(4): 1222–57. https://www.journals.uchicago.edu/doi/10.1086/701703.

Reinsberg, Bernhard, Andreas Kern, Mirko Heinzel, and Saliha Metinsoy. 2023. "Women's Leadership and the Gendered Consequences of Austerity in the Public Sector: Evidence from IMF Programs." *Governance*: doi: 10.1111/gove.12764.

Reinsberg, Bernhard, Daniel O Shaw, and Louis Bujnoch. 2022. "Revisiting the Security--Development Nexus: Human Security and the Effects of IMF Adjustment Programmes." *Conflict Management and Peace Science*: 07388942221111064.

Reinsberg, Bernhard, Thomas Stubbs, and Alexander Kentikelenis. 2022a. "Compliance, Defiance, and the Dependency Trap: International Monetary Fund Program Interruptions and Their Impact on Capital Markets." *Regulation & Governance* 16(4): 1022–41.

———. 2022b. "Unimplementable by Design? Understanding (Non-)Compliance with International Monetary Fund Policy Conditionality." *Governance* 35(3): 689–715.

Reinsberg, Bernhard, Thomas Stubbs, Alexander Kentikelenis, and Lawrence King. 2019. "The Political Economy of Labor Market Deregulation during IMF Interventions." *International Interactions* 45(3): 532–59.

———. 2020. "Bad Governance: How Privatization Increases Corruption in the Developing World." *Regulation and Governance* 14(4): 698–717.

Roodman, David. 2009. "How to Do Xtabond2: An Introduction to Difference and System GMM in Stata." *Stata Journal* 9(1): 86–136. http://www.nuffield.ox.ac.uk/users/bond/file_HowtoDoxtabond8_with_foreword.pdf.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *Journal of Politics* 78(3): 941–55.

Shim, Sujeong. 2022. "Who Is Credible? Government Popularity and the Catalytic Effect of IMF Lending." *Comparative Political Studies* 55(13): 2147–77.

Smith, Alastair, and James R Vreeland. 2006. "The Survival of Political Leaders and IMF Programs: Testing the Scapegoat Hypothesis." In *Globalization and the Nation State: The Impact of the IMF and the World Bank*, , 263–89.

Stone, Randall W. 2004. "The Political Economy of IMF Lending in Africa." *American Political Science Review* 98(4): 577–92. http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=265353.

Stone, Randall W, and Martin C. Steinwand. 2008. "The International Monetary Fund: A Review of the Recent Evidence." *Review of International Organizations* 3(2): 123–49.

Stubbs, Thomas H., Alexnder E. Kentikelenis, Rebecca Ray, and Kevin P. Gallagher. 2022. "Poverty, Inequality, and the International Monetary Fund: How Austerity Hurts the Poor and Widens Inequality." *Journal of Globalization and Development* 13(1): 61–89.

Stubbs, Thomas H et al. 2017. "The Impact of IMF Conditionality on Government Health Expenditure: A Cross-National Analysis of 16 West African Nations." *Social Science and Medicine* 174(3): 220–27.

Stubbs, Thomas H, Alexander E Kentikelenis, and Lawrence P King. 2016. "Catalyzing Aid? The IMF and Donor Behavior in Aid Allocation." *World Development* 78: 511–28. http://www.sciencedirect.com/science/article/pii/S0305750X15002326.

Stubbs, Thomas H, Bernhard Reinsberg, Alexander E Kentikelenis, and Lawrence P King. 2020. "How to Evaluate the Effects of IMF Conditionality: An Extension of Quantitative Approaches and an Empirical Application to Government Education Expenditures." *Review of International*

Organizations 15(1): 29–73.

Vadlamannati, Krishna Chaitanya, Gina Maria G. Østmoe, and Indra de Soysa. 2014. "Do IMF Programs Disrupt Ethnic Peace? An Empirical Analysis, 1985–2006." *Journal of Peace Research* 51(6): 711–25.

Vreeland, James R. 2002. "The Effect of IMF Programs on Labor." *World Development* 30(1): 121–39.

———. 2003. *The IMF and Economic Development*. Cambridge University Press.

Williams, Laron K. 2012. "Pick Your Poison: Economic Crises, International Monetary Fund Loans and Leader Survival." *International Political Science Review* 33(2): 131–49.

Woo, Byungwon. 2013. "Conditional on Conditionality: IMF Program Design and Foreign Direct Investment." *International Interactions* 39(3): 292–315.

Appendix A – Calculating *REPSTAT* indicators under simultaneous equations estimators

In estimation of simultaneous equations, a log-likelihood estimator chooses coefficients for the variables in all of the equations in a given system, which maximize the relevant distribution function. While one of the equations is typically regarded as the outcome equation, it is estimated in one stage with the other equations. This poses a challenge for *REPSTAT* indicators, which are calculated based on a given autonomously-estimated outcome equation.

In estimation of simultaneous equations, the added value of each observation in generating a particular coefficient in a particular equation should be calculated considering all of the variables in all of the equations. Since all of the variables in the system co-determine each other's coefficients, the simplest way to calculate *REPSTAT* indicators is to regress the treatment variable against all of the other variables in the system, except the outcome variable (the dependent variable in the outcome equation); In other words, to specify all of them in Regression (2). However, this calculation would be similar if all variables were specified in a single all-encompassing outcome equation. A more appropriate calculation of *REPSTAT* indicators should consider that (1) the log-likelihood estimator considers error terms of equations, so it matters how variables are grouped into equations; and that (2) some variables appear in more than one equation, thus constraining the search for the optimal set of coefficients more than other variables.

Considering all this, we calculate *REPSTAT* indicators based on the outcome equation, but for this purpose we replace the values of any of its regressors that appear also in the non-outcome equations with its residuals as calculated against variables in those equations. Specifically, residuals are calculated by regressing each such variable against all of the other variables in a non-outcome equation, regardless of which side of that equation they are on.[8] The rationale is that we must isolate outcome-regressors' added explanatory value to the outcome, which excludes explanatory value of non-outcome equation variables, and non-outcome dependent variables also constrain the outcome. Note that outcome-regressors may be specified in non-outcome equations as dependent variables or independent variables, and may or may not include the treatment variable(s) of interest.

Consider for example the following set of two simultaneous equations, Y_i being the outcome:

$$(8)\ Y_i = \alpha_Y + \beta_T T_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$(9)\ Z_i = \alpha_Z + \gamma_T T_i + \gamma_1 X_{1i} + \gamma_3 X_{3i} + \mu_i$$

In this example, the treatment and one of the outcome's regressors – X_1 – are both specified as predictors of a variable Z that is not part of the outcome equation. We first predict the residuals of T and X_1 based on Equations (10) and (11) respectively:

$$(10)\ T_i = \alpha_{ZT} + \gamma_Z Z_i + \gamma_{X1} X_{1i} + \gamma_{X3} X_{3i} + \mu_i$$

$$(11)\ X_{1i} = \alpha_{Z1} + \gamma_Z Z_i + \gamma_T T_i + \gamma_{X3} X_{3i} + \mu_i$$

We then calculate *REPSTAT* indicators based on Equation (8a):

[8] This assumes that there is at least one overlapping variable among the system's equations. While this is not necessarily a technical requirement in all forms of simultaneous equations estimation, it is highly likely from a theoretical perspective, and always the case in the literature reviewed for this study.

$$(8a)\ Y_i = \alpha_Y + \beta_T Resid(T)_i + \beta_1 Resid(X1)_i + \beta_2 X_{2i} + \varepsilon_i$$

When there are more than one non-outcome equations, cascading residuals of outcome-equation regressors are calculated based on all of the equations in which they appear. Specifically:

a) Residuals are first calculated in non-outcome equation #1 for any variable on any side of that equation that appears on any side in any other equation;

b) Any such residuals are then specified in non-outcome equation #2, for the variables that are included in it on any side, along with original values of the other variables of equation #2 if any; a new series of residuals is then calculated in non-outcome equation #2 (against all variables on any side there) for any variable that appears on any side in any other equation other than #1 and #2;

c) The resulting residuals are then specified in non-outcome equation #3, for the variables that are included in it on any side, together with residuals calculated in equation #1 for variables that are included in equation #3 but not in equation #2 if any, and original values of variables in equation #3 that do not appear in either equation #1 or #2 if any; and a new series of residuals is calculated in non-outcome equation #3 (against all variables on any side there) for any variable that appears on any side in any other equation other than #1 #2 and #3;

d) This calculation of cascading residuals continues until all non-outcome equations are exhausted, and the last calculation of residuals is specified in the outcome equation for any variable that appeared in any other equation.[9]

This perhaps requires establishing an appropriate sequence of equations, which may follow the recursion order if the system is recursive. Otherwise, for efficiency of calculation the non-outcome equations can be sequenced by the number of their regressors, starting with the equation with fewest regressors. However, our experience shows that the sequence of non-outcome equations does not matter. For convenience, we simply followed the order of non-outcome equation as they were originally specified in the replicated studies.

Things get even more complicated when not all equations are fully observed over the entire dataset. Some simultaneous estimators, such as the Seemingly Unrelated Bivariate Probit estimator, use only data that are fully observed in all of the equations. In contrast, Conditional Mixed Process (CMP) uses the union of the different non-outcome equations' sets of observations: Each observation enters the CMP calculation based on all of the fully observed equations for that observation.[10] Following this logic, for CMP estimators the dataset is divided into groups of observations, depending on the combination of fully observed equations, and the calculation of cascading residuals outlined above is run separately for each group. In each observation only variables from fully observed equations can contribute to the explanation, and other variables contribute zero added value. This means that outcome-regressors retain their original values for *REPSTAT* calculations in observations for which all the non-outcome equations are missing. In observations for which the outcome equation itself is not fully observed, zero values are entered for

[9] Note that for variables that appear in the outcome equation and in more than one non-outcome equation, residuals are calculated more than once; it is the last calculation that is used for *REPSTAT* calculation in the outcome equation, because it extracts the variable's added value over all other variables in all of the non-outcome equations.

[10] See for example Roodman's explanation at: https://www.statalist.org/forums/forum/general-stata-discussion/general/1291779-biprobit-and-bitobit-using-cmp-command.

outcome-regressors for *REPSTAT* calculations. Using the above example, in groups of observations where Equation (8) is observed but Equation (9) is not, original values of T and X_1 are entered in Equation (8a) for $Resid(T)$ and $Resid(X_1)$ respectively; in groups of observations where Equation (9) is observed but Equation (8) is not, zero values are entered for $Resid(T)$ and $Resid(X_1)$. The calculations outlined in Equations (2)-(7) are then performed over the entire dataset, using Equation (8a) in place of Equation (1).

Appendix B – Descriptive statistics and robustness checks

Table B1: Replicated studies by outcomes (treatment frequencies)

Area	Outcome	Replicated studies	Frequency
Civil war	Civil war	Midtgaard, Vadlamannati, and de Soysa (2014) Vadlamannati, Østmoe and de Soysa (2014)	24
	Total		24
Economic development	Economic growth	Bas and Stone (2014) Marchesi and Sirtori (2011)	16
	FDI	Biglaiser and DeRouen (2010) Breen and Egan (2019) Goes (2023)	80
	Export skills	Demir (2022)	124
	Women's rights	Detraz and Peksen (2016)	14
	Total		234
Education	Education	Stubbs *et al*. (2020)	46
	Total		46
Financial markets	Bond spreads	Shim (2022)	9
	Creditworthiness	Cho (2014) Gehring and Lang (2020)	16
	Financial crises	Lipscy and Lee (2019)	16
	Total		41
Health	Child mortality	Daoud and Reinsberg (2019) Forster *et al*. (2020)	37
	Child vaccination	Daoud and Reinsberg (2019)	48
	Public health system	Daoud and Reinsberg (2019) Forster *et al*. (2020) Kentikelenis, Stubbs, and King (2015)	49
	Well-being (surveys)	Andrijic and Barbic (2021)	20
	Total		154
Income inequality	Income inequality	Lang (2020) Forster *et al*. (2019)	168
	Total		168
Other social development	Foreign aid	Stubbs, Kentikelenis, and King (2016)	14
	Human security	Reinsberg, Shaw, and Bujnoch (2022)	16
	Labour	Reinsberg *et al*. (2019~~b~~)	10
	Shadow economy	Blanton, Early, and Peksen (2018)	8
	Total		48
Rule of law	Corruption	Reinsberg, Kentikelenis, and Stubbs (2021)	105
	Democracy	Birchler, Limpach and Michaelowa (2016)	44
	Expropriation	Biglaiser, Lee and Staats (2016)	28
	Government crisis	Dreher and Gassebner (2012)	76
	Human rights	Abouharb and Cingranelli (2009)	9
	Total		262

Table B2: Descriptive statistics

Variable	Definition	count	mean	sd	min	max
Relative Effective Sample (RES)	The effective number of observations divided by the nominal number of observations	977	0.18	0.133	0.001	0.697
Generic IMF variable	Dummy for a treatment that is a generic measure of IMF participation, which disregards program type, such as a dummy for any IMF program being active in a given year, its first difference, or a count of the total number of IMF conditions (omitted category)	977	0.575	0.495	0	1
IMF dummy	Dummy for a treatment that is a dummy in its original model to characterize IMF participation, including any programs, specific facilities, or differenced dummies (omitted category)	977	0.373	0.484	0	1
IMF conditionality	Dummy for a treatment that measures IMF conditionality	977	0.535	0.499	0	1
Causal inference approach	Dummy for a treatment included in a model that uses any of the three causal inference approaches described below (Instrumental variables, Simultaneous equations or Selection model), Some models use instrumental variables inside simultaneous equations, so the frequency of this dummy is lower than the sum of its components.	977	0.66	0.474	0	1
Instrumental variables	Dummy for a treatment that is instrumented (i.e. is based on predicted values from some first-step model)	977	0.185	0.389	0	1
Simultaneous equations	Dummy for a treatment that is included in model that uses a simultaneous equations approach (with auxiliary equation(s) for potentially endogenous variables)	977	0.47	0.499	0	1
Selection model	Dummy for a treatment that is included in a model that uses a selection model for this treatment variable	977	0.062	0.242	0	1

Table B2: Descriptive statistics (cont.)

Variable	Definition	count	mean	sd	min	max
Interaction	Dummy for a treatment that is interacted with another regressor	977	0.04	0.196	0	1
Confounders	Number of regressors in the analysed model in which the treatment was specified, exclusive of all of the treatment variables in the model (log-transformed +1 for inclusion in the regression)	977	12.718	6.951	0	55
Country-FE	Dummy for a treatment that is included in a model that uses country-fixed effects	977	0.908	0.289	0	1
Year-FE	Dummy for a treatment that is included in a model that uses year-fixed effects	977	0.786	0.41	0	1
Nominal sample	Number of observations in the nominal sample of the analysed model (log-transformed for inclusion in the regression)	977	2259.805	1387.319	93	5675
Multiple treatments	Dummy for a treatment that is included in a model that includes multiple treatments (different operationalizations of IMF intervention). For interacted treatments, the constitutive variables are not counted here.	977	0.783	0.412	0	1
Non-IPE journal	Dummy for a treatment in a model that is included in a study published in a ranked journal that is not classified under the disciplines of Political Science, International Relations, and Economics	977	0.357	0.479	0	1
Estimate in appendix	Dummy for a treatment in a that model is presented in an appendix at the end of the paper or in online supplemental material	977	0.245	0.43	0	1
Estimate significant	Dummy for a treatment that is statistically significant at $p<0.05$ in its model.	977	0.278	0.448	0	1

Table B3: Results with additional contextual variables on publication practices

	(1)		(2)		(3)		(4)		(5)	
Generic IMF variable	0.085***	(0.012)	0.079***	(0.011)	0.084***	(0.011)	0.072***	(0.011)	0.072***	(0.011)
Estimate in appendix	0.005	(0.010)	0.021**	(0.009)	0.017*	(0.009)	0.019**	(0.009)	0.019**	(0.009)
Estimate significant	0.005	(0.010)	-0.011	(0.012)	-0.009	(0.012)	-0.031***	(0.011)	-0.030***	(0.011)
Confounders			-0.027**	(0.011)	-0.024**	(0.011)	-0.015	(0.012)	-0.013	(0.012)
Nominal sample			-0.003	(0.005)	-0.003	(0.005)	-0.001	(0.005)	-0.001	(0.005)
Country-FE			-0.084***	(0.023)	-0.088***	(0.023)	-0.082***	(0.023)	-0.081***	(0.022)
Year-FE			-0.025*	(0.015)	-0.024	(0.015)	-0.020	(0.015)	-0.018	(0.015)
Interaction					-0.079***	(0.023)	-0.091***	(0.026)	-0.092***	(0.026)
Multiple treatments							-0.066***	(0.013)	-0.066***	(0.013)
Non-IPE journal									-0.008	(0.008)
Observations	977		977		977		977		977	
R-squared	0.100		0.149		0.162		0.190		0.191	

Notes: Linear regression estimated via Ordinary Least Squares. Standard errors clustered on models in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.

Table B4: Results with different clustering

	(1)		(2)		(3)		(4)	
Generic IMF variable	0.077***	(0.008)	0.076***	(0.007)	0.077*	(0.040)	0.076*	(0.037)
Number of regressors	-0.019**	(0.008)	0.020**	(0.009)	-0.019	(0.025)	0.020	(0.017)
Nominal sample	-0.002	(0.005)	0.015***	(0.005)	-0.002	(0.014)	0.015	(0.011)
Country-FE	-0.081***	(0.022)	-0.090***	(0.021)	-0.081*	(0.047)	-0.090**	(0.043)
Year-FE	-0.021*	(0.011)	0.028**	(0.014)	-0.021	(0.023)	0.028	(0.029)
Causal inference approach			-0.101***	(0.012)			-0.101***	(0.034)
Interaction			-0.096***	(0.022)			-0.096**	(0.037)
Multiple treatments			-0.056***	(0.011)			-0.056*	(0.028)
Non-IPE journal			0.031***	(0.007)			0.031**	(0.012)
Observations	977		977		977		977	
R-squared	0.144		0.254		0.144		0.254	

Notes: Linear regression estimated via Ordinary Least Squares. Robust standard errors (first two columns) and standard errors clustered on studies (last two columns) in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.

Table B5: Results with transformed dependent variable

	(1)		(2)		(3)		(4)		(5)	
Generic IMF variable	0.831***	(0.107)	0.847***	(0.102)	0.922***	(0.102)	0.893***	(0.102)	0.896***	(0.102)
Number of regressors	-0.170**	(0.077)	-0.018	(0.078)	0.001	(0.080)	0.037	(0.085)	0.024	(0.084)
Nominal sample	-0.146***	(0.045)	-0.098**	(0.046)	-0.109**	(0.045)	-0.096**	(0.046)	-0.055	(0.051)
Country-FE	-0.459***	(0.170)	-0.487***	(0.160)	-0.560***	(0.167)	-0.544***	(0.167)	-0.577***	(0.170)
Year-FE	-0.031	(0.133)	0.180	(0.137)	0.183	(0.142)	0.215	(0.148)	0.217	(0.150)
Causal inference approach			-0.457***	(0.115)	-0.469***	(0.119)	-0.474***	(0.119)	-0.595***	(0.126)
Interaction					-1.178***	(0.282)	-1.216***	(0.290)	-1.188***	(0.290)
Multiple treatments							-0.168	(0.120)	-0.173	(0.120)
Non-IPE journal									0.244***	(0.073)
Observations	977		977		977		977		977	
R-squared	0.133		0.152		0.181		0.183		0.188	

Notes: Linear regression estimated via Ordinary Least Squares. Standard errors clustered on models in parentheses. Significance levels: * p<.1 ** p<.05 *** p<.01. Constant not reported. Observations relate to variables operationalizing IMF intervention (treatments) in models included in studies of the effectiveness of IMF programs. Country-FE and Year-FE are dummies for treatments specified in models that include such fixed effects. None of the estimated regressions in this study include fixed effects themselves.