NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Hippolyte Weneyam Balima, Anna Sokolova

IMF PROGRAMS AND ECONOMIC GROWTH: A META-ANALYSIS

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: ECONOMICS
WP BRP 241/EC/2020

IMF Programs and Economic Growth: A Meta-Analysis *

Hippolyte Weneyam Balima † Anna Sokolova‡

September 17, 2020

Abstract

We examine 994 estimates of the effects of IMF programs on economic growth as reported by 36 studies. The mean reported effect is positive, but the estimates vary widely. We use meta-regression analysis to disentangle sources of this variation, addressing model uncertainty with Bayesian Model Averaging and LASSO. We find that estimates vary systematically depending on data and methods employed by the researchers. Reported effects of IMF programs tend to be more positive for samples that include countries with high levels of institutional and economic development, when measured on longer horizons, estimated using more recent data or obtained with the propensity score matching technique. Estimates appear to depend on the types of IMF programs being considered, as general resource programs tend to result in less favorable growth outcomes compared to programs that lend from concessional resources. Authors with IMF affiliation tend to report estimates that are somewhat higher than those of outside researchers.

JEL Codes: *F3, F4, O19*

Keywords: IMF programs; Economic growth; Meta-Analysis; Bayesian Model Averaging

---

*Data and code supporting results in this study are available from the authors upon request. The views expressed in this paper are those of the authors. Anna Sokolova acknowledges support from the Basic Research Program at the National Research University Higher School of Economics (HSE) and from the Russian Academic Excellence Project '5-100'.

†Email: `hyppobalima@yahoo.fr`

‡Department of Economics, University of Nevada Reno, 1664 N Virginia St, Reno, NV 89557 USA. National Research University Higher School of Economics, International Laboratory for Macroeconomic Analysis, Myasnitskaya Ulitsa 20, Moscow, Russia, 101000. Email: `asokolova@unr.edu`

1 Introduction

One of the key purposes of the International Monetary Fund (IMF) stated in its Articles of Agreement is "to give confidence to members by making the general resources of the Fund temporarily available to them ...". Since its inception in 1944, the IMF has provided lending to more than 150 countries through about 1,300 IMF-supported programs. Over time, the IMF's agenda as that of an international lender has expanded dramatically, reaching a cumulative total commitment level of about 1.6 trillion dollars at the end of 2019 (see Figure 1), the equivalent of about 5.7 percent of average world GDP over the period 1960-2019. This increased activity has prompted many economists—both from within and from outside the Fund—to investigate how IMF-supported programs affect program participants.

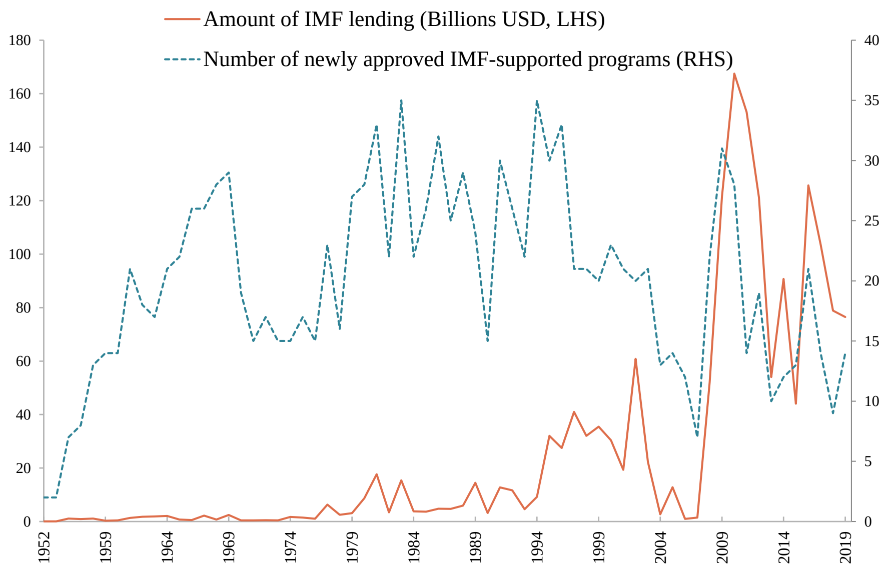
Over the years, one of the most heated debates has centered around the effects that the IMF-supported programs have on economic growth in the participant countries. While some papers document a positive effect of IMF involvement on growth (e.g. Dicks-Mireaux *et al.* 2000, Mercer-Blackman & Unigovskaya 2004), others find the effect to be negative (e.g. Barro & Lee 2005, Dreher 2006), or dependent on other factors (e.g. institutional development, see Binder & Bluhm 2017). Researchers in this field have produced a plethora of evidence that, nevertheless, does not yield a consensus on the direction of the effect and the underlying mechanisms.

We collect 994 estimates of the effects of IMF programs on economic growth as reported by 36 studies. The mean effect is around 0.36, suggesting that, according to the literature, IMF programs tend to raise the growth rates of participant countries by 0.36 percentage points. At the same time, the estimates vary widely—we record a standard deviation of 1.47. In this paper, we examine the sources of this variation using meta-analysis, a tool for performing quantitative synthesis of the literature.

In economics, meta-analysis has been employed to study a variety of topics, such as the effects of changes in minimum wage on employment (Card & Krueger 1995, Doucouliagos & Stanley 2009); the link between currency unions and trade (Rose & Stanley 2005), and trade and distance (Disdier & Head 2008); the effects of active labor market programs (Card *et al.* 2017); habit formation (Havranek *et al.* 2017) and the excess sensitivity in consumption (Havranek & Sokolova 2020), etc. To our knowledge, our study is the first meta-analysis of the literature estimating the effects of IMF programs on economic growth.

For each of the 994 estimates that we found, we record information detailing the specific context in which the estimate was obtained; based on this information, we construct a set of 25 potential explanatory variables. We first conduct our analysis in a frequentest setting, using a fixed subset of the controls that, in our opinion, reflects the most crucial aspects of study design (e.g. estimation method, broadly defined country groups under investigation, see Section 3). We then re-frame the problem in the Bayesian context, taking into account model uncertainty—that is, the uncertainty over which of the 25 controls belong to the 'true' data generating process. We perform Bayesian Model Averaging on the full set of the 25 controls that provides a finer level of detail (e.g. by controlling for specific instruments used when estimating with instrumental variables, specific geographic region studied, see Section 4). We check the robustness of these results by performing model selection with LASSO, an alternative tool for navigating model uncertainty.

Figure 1: IMF-supported programs since 1952.



Notes: The amount of IMF lending is depicted on the left vertical axis; the number of newly approved IMF-supported programs is depicted on the right vertical axis. In the source database each program amount was denominated in USD between 1952 and 1971, and in Special Drawing Right (SDR) after 1971. We use the end of period SDR per USD exchange rate to convert post-1971 program amounts into USD. *Source: IMF's Strategy, Policy and Review Department, Fund Arrangements since 1952 database.*

We find that the data and method choices that researchers make have systematic effects on the estimates that their analyses yield. The estimates of the effect of IMF programs on growth appear to vary with estimation horizons: long-run effects (estimated on horizons of at least 10 years) are larger compared to effects estimated for the short-run (on horizons under two years). We do not document a statistically significant difference between the short- and the medium- run effect—however, we find some (weak) evidence suggesting that the medium-run effect may be lower, pointing to a potential non-linearity in the effects of IMF programs on growth at different horizons. We show that samples that include evidence from a mix of developed and developing countries tend to yield higher estimates than those that do not; this may mean that the IMF involvement is more successful in countries with strong institutional records. We also find that the estimated effects of IMF programs appear to have improved over time, with fresher data yielding higher growth effects. We additionally investigate whether this literature is prone to selective reporting of the results; even though we do not find consistent evidence of publication bias, we document that estimates reported by IMF staff members tend to be somewhat higher than those presented by outside researchers.

We find some evidence suggesting that different types of IMF programs may have systematically different effects on growth. Programs that offer funding from general resources, that target middle-income countries and typically have shorter durations, seem to have a lower effect on countries' growth outcomes compared to programs that offer support to poor countries, relying on concessional lending provided on more lenient terms. This result may be interpreted in several ways, with different implications for the role of the IMF. First, it may suggest that concessional resource programs are more successful in fostering growth by design, as typically their goal is not only to address short-term imbalances in participant countries, but to also correct systemic problems plaguing low income countries. Second, as the concessional lending is offered with more lenient repayment conditions, it likely puts the debtor country under less financial stress which may be more advantageous for growth. Finally, IMF programs come with a set of policy measures ('conditionality') that a participant country is encouraged to implement. These policies could help unlock reform implementation for governments that face stringent political constraints; at the same time, conditionality is sometimes criticized as being inappropriate for developing countries.[1] Concessional programs that are offered on humanitarian grounds tend to require less conditionality overall[2]—if conditionality is indeed harmful for growth, this would explain why concessional programs may result in better growth outcomes.

An important caveat that studies in this field must address is the fact that participation in IMF programs is not random: countries turn to the IMF for help when facing economic downturns. Estimation methods rely on different assumptions in addressing this endogenous selection problem. We find that one method in particular—propensity score matching—produces estimates that are systematically higher compared to those obtained with other techniques. On the one hand, this may be because propensity score matching does not rely on exclusion restrictions that other methods invoke, a feature

[1]Stiglitz (2002) discusses the shortcomings of neoliberal economic policies that are often part of conditionality requirements that IMF negotiates with developing countries participating in its programs.

[2] See discussions in Dreher (2004) and Stone (2008).

that may make its results more reliable. On the other hand, it is not always possible to link a participation observation to a non-participant 'match' with a similar propensity score—especially when it comes to participants with high participation likelihoods; these observations would then be discarded.[3] Therefore, an alternative explanation for propensity score matching estimates being systematically higher is that this method relies on samples that differ systematically from samples considered by the rest of the literature.

An important implication of these findings is that estimates of the effect of IMF programs on growth vary dramatically with study context. Even though we find the overall effect to be positive, the magnitude varies with specific method and data choices. For example, for a mixed sample of developed and developing countries IMF programs would raise growth by an average of 0.913 percentage points, whereas for a sample that exclusively features developing countries, this estimate would drop to about 0.353. These results highlight the need for a comprehensive approach to evaluating the role of the IMF, that examines why the effects of IMF involvement are more favorable under some sets of circumstances and less favorable in other instances, and addresses the context-specific shortcomings of IMF programs.

2 The Data

The effect of IMF programs on growth is often studied with the regression model of the form:

$$y_{it} = \beta_0 + \beta^{\text{IMF}} \cdot \text{IMF}_{it} + \gamma X_{it} + \epsilon_{it} \tag{1}$$

where y_{it} is a measure of economic performance in country i at time t (e.g. annual growth rate of real GDP), IMF_{it} is a proxy for IMF involvement (e.g. a dummy indicator for whether there is an active IMF loan agreement), X_{it} is a vector of controls (e.g. other monetary or fiscal factors that could be affecting growth), and ϵ_{it} is the error term—see, e.g. Bordo & Schwartz (2000), Baqir *et al.* (2005), Barro & Lee (2005), Dreher (2006). The coefficient β^{IMF} captures the effect of the presence of the IMF in a country on the country's economic performance. A positive β^{IMF} would mean that IMF programs have stimulating effect on growth, while a negative value would imply that they constrict economic activity.

In addition to studying the effect of the presence of IMF programs, some studies attempt to disentangle specific channels through which IMF could influence growth outcomes of the participant countries. A number of authors investigate whether growth is affected by the IMF loan amount (e.g Barro & Lee 2005, Dreher 2006, Binder & Bluhm

[3] This problem has been previously discussed in Hardoy (2003) and Atoyan & Conway (2006).

2017)—in these studies, IMF_{it} would capture the amount of the IMF loan granted to a participant country. A positive β^{IMF} would then imply that countries receiving more IMF funding tend to experience faster economic growth; this could mean that more generous IMF programs are more effective in restoring economic activity. A negative β^{IMF}, on the other hand, would suggest that IMF financial involvement is associated with worse growth outcomes for program participants. This result could arise if, for example, large IMF loans allowed opportunistic policymakers to postpone much needed structural reforms that are deemed politically costly.

IMF loans typically come with a set of mutually negotiated conditions that a participating country agrees to meet. For each program, these conditions would be specified in the Fund program documents, as well as the Letters of Intent and Memorandum of Economic and Financial Policies prepared by the country's authorities. These conditions may require countries to implement a number of structural reforms, to address fiscal slippages, to promote market-based policies, etc (see overview in Mussa & Savastano 1999, IMF 2005, see IEO 2018 for recent examples). We will refer to such policy requirements as 'IMF conditionality' throughout the text. A number of studies assess how growth is affected by the extent to which participant countries comply with these requirements specified within the IMF-supported programs (e.g. Killick 1995, Dreher 2006); in these studies, IMF_{it} is a measure of countries' compliance with IMF conditionality; a positive β^{IMF} would mean that countries that adopt a higher portion of the agreed upon policies are experiencing higher growth rates compared to countries that do not comply; on the other hand, a negative coefficient could mean that IMF conditionality tends to impair growth of the program participants.

In all three cases described above—i.e. with IMF_{it} measuring IMF presence, loan amount or compliance—a positive β^{IMF} would imply that IMF programs improve countries' growth outcomes. For this reason, in this paper we consider the estimates of β^{IMF} obtained with all three approaches.[4] In the analysis that follows, we will take measures to account for these differences in the precise interpretation of β^{IMF} by introducing corresponding controls as well as comparing the overall results with those obtained on more homogenous subsamples.

We collect estimates of β^{IMF} from empirical studies that examine how economic growth is affected by either the IMF presence, the IMF loan amounts or by compliance with the IMF conditionality. We employ Google Scholar to identify the relevant studies on

[4] In addition to the research we discussed, a number of papers estimate effect of recidivism on growth outcomes, studying how the prolonged and repeated use of IMF programs affects average growth over long periods of time (e.g. Easterly 2005). We do not include these results here, as we believe that the coefficient interpretation would be substantially different: unlike this latter strand of literature, all three research questions discussed above study growth *following* the implementation of an IMF program.

the topic using a search query that contains: "IMF programs" or "IMF bailouts", and "economic growth". In addition, we review the Google Scholar citations of a number of key papers in this literature: Khan & Knight (1981), Pastor (1987), Haque & Khan (1998), Fischer (1997) , Mussa & Savastano (1999), Dicks-Mireaux *et al.* (2000), Barro & Lee (2005), Easterly (2005), and Dreher (2006). We then go through the list of papers we found and download those that, judging by the abstracts, may contain empirical estimates of β^{IMF}. We end our search on November 11, 2018. We apply two inclusion criteria to the downloaded sample of primary studies. First, the study must contain empirical estimates of the effect of IMF programs on economic growth. Second, the estimates must be accompanied by the associated standard errors, or some statistics from which the standard error can be computed. Applying these two criteria, we end up with 36 primary studies reporting a total of 994 estimates.[5] We list these studies in Appendix C.

Table 1: Estimates of the effect of IMF programs on growth:
Sample statistics across features of study design

	Unweighted				Weighted				
	Mean	Median	5%	95%	Mean	Median	5%	95%	N
All	0.36	0.04	-1.50	2.70	0.21	0.02	-2.02	2.81	994
Dep.var=growth rate	0.28	0.03	-1.50	2.49	0.04	0.00	-2.02	2.05	901
Exp. var.=Dummy	0.35	0.11	-1.78	2.86	0.04	0.02	-2.21	2.45	684
Exp. var.=Amount	0.09	-0.00	-0.40	0.80	0.06	-0.01	-0.46	0.80	209
Exp. var.=Compliance	-0.73	-0.46	-5.51	2.83	-0.14	0.04	-1.37	0.11	8
Mixed sample	0.34	0.01	-1.82	3.30	0.49	0.02	-1.69	4.01	294
Developing only	0.26	0.03	-1.37	2.11	-0.12	-0.00	-2.21	1.76	607
Short Run effect	0.34	0.04	-1.45	2.43	0.03	0.01	-2.02	1.84	523
Medium Run effect	-0.07	-0.01	-1.82	1.74	-0.21	-0.04	-3.73	1.78	318
Long Run effect	1.65	1.46	-0.07	3.75	0.96	0.57	0.04	3.39	60
Mixed res.	0.45	0.07	-1.67	3.02	0.21	0.02	-2.21	3.35	564
General res. only	-0.29	-0.01	-1.41	0.77	-0.40	-0.04	-2.02	1.29	192
Concessional res. only	0.40	0.04	-1.38	2.10	0.40	0.33	-1.17	1.84	145
IMF staff	1.03	0.88	-0.68	3.39	0.29	0.04	-1.17	1.92	141
Dep.var=Growth change	1.04	1.19	-1.53	3.45	1.43	1.72	-1.80	4.05	69
Dep.var=Growth difference	1.24	1.03	-0.45	3.13	1.24	1.03	-0.45	3.13	24

Notes: 5% and 95% refer to the corresponding percentiles. The left panel presents summary statistics from the raw data; the right panel reports summary statistics for data weighted by the inverse of the number of estimates reported in each study; this gives studies with high and low numbers of reported estimates roughly equal weights in the overall sample.

We present the sample statistics in Table 1. The overall mean of the reported estimates

[5]Most studies estimate model (1) using growth rates expressed in percentage terms as the left-hand side variable; β^{IMF} can then be interpreted as a growth rate increase in percentage points associated with an IMF programs. However, we also found four studies that express the left-hand-side variable in shares (e.g. use differences in log output without multiplying by 100). To make results coming from this latter group comparable to the rest of the sample, we multiply these estimates and standard errors by 100.

is positive: IMF programs appear to improve growth outcomes by about .36 percentage points. At the same time, the estimates of the growth effect seem to vary substantially with the studies' design: studies that consider IMF presence (and use a dummy as the explanatory variable) on average report positive effects on growth; researchers that investigate the effects of loan amount (and use loan to GDP on the right-hand side) find mean effect that is close to zero, while those who consider the effect of compliance with conditionality (and use a measure of compliance on the right-hand side) report negative results. Matters become more complicated when we weight the data by the inverse number of estimates reported per study and observe the means shrinking toward zero—suggesting that the sizable effects seen for the unweighted means could be driven by results coming from a few studies reporting large estimates rather than many studies reporting similar results.

These results may point toward a disconnect between different channels through which the IMF involvement affects growth. At the same time, these observed inconsistencies across estimate signs may also be due to differences in data and estimation strategies employed by researchers. For example, it may be that, coincidentally, studies that examine different channels through which IMF may affect growth also happen to consider different country groups, estimate the effects for different horizons, or study different kinds of IMF programs.

The majority of studies in our sample consider developing countries; however, a subset of 294 estimates is obtained for mixed samples of developing and advanced economies, showing a slightly different mean effect. It also appears that there may be systematic differences in the effects of IMF involvement across different horizons. The programs seem to have a moderate positive effect on growth in the first two years after program adoption (termed 'Short Run Effect' in Table 1); at the same time, the estimates become negative for horizons of 2-9 years (termed 'Medium Run'). Finally, for horizons of 10 years and over (i.e. in the 'Long Run') the mean estimated effect on growth is positive again, and much more prominent.

There also seem to be some differences in the estimated growth outcomes depending on the types of programs being examined. The majority of estimates in our sample come from studies that do not distinguish between program types and consider mixed samples of IMF programs. At the same time, there are 192 estimates that pertain exclusively to programs that offer funding from general resources. These types of programs are typically aimed at helping countries overcome short-lived balance of payment issues; according to Table 1, these programs tend to be associated with lower growth outcomes on average. We also collect 195 estimates for programs funded through concessional resources, that is, programs aimed at poor countries that offer more lenient credit conditions (e.g. zero or low interest rates, and longer grace periods) and are designed to address systemic issues in the participant economies. It appears that, on average, these types of programs may be associated with better growth outcomes.

We also notice that papers in which at least one of the co-authors is an IMF staff member tend to come up with high and positive estimates of the effect of IMF programs on growth. This piece of evidence could imply some selective reporting when it comes to results published by researchers affiliated with the IMF, but there are other possible explanations as well: for example, it could be that papers co-authored by IMF staff members tend to focus on specific kinds of programs (e.g. programs for poor countries offering concessional lending), or horizons (e.g. long run horizons) that tend to yield higher growth estimates. We will attempt to disentangle these potential explanations in the subsequent sections.

Finally, apart from papers estimating the effect of IMF involvement on growth rates, we find a few studies estimating how IMF involvement affects *changes* in growth rates (and use $y_{it} - y_{it-1}$ as the left-hand side variable). We also find two studies looking at the difference between growth projected by the IMF and the actual realized growth rate (i.e. $y^p_{it} - y_{it}$). To make these estimates somewhat consistent with the rest of our sample, we multiply the estimates coming from these two studies by (-1); these modified estimates now capture the effect of IMF programs on $y_{it} - y^p_{it}$, with positive estimates implying faster realized growth in presence of IMF programs—similar to the rest of our collected estimates. Table 1 reports statistics for both groups.

So far it appears that there are several features of studies' design that may have systematic effects on the estimates of β^{IMF} that researchers report. In the subsequent sections we take a more systematic approach to uncovering how different choices made by researchers contribute to the magnitudes of the estimates that they produce. We will also attempt to tackle a number of questions that are fundamental to this field of literature: namely, whether some IMF programs are more favorable for growth than others, and whether growth effects differ with horizon and country groups. We will also take a close look at the potential issues underlying estimation methods most commonly employed by researchers in this field. To estimate well the effect of IMF programs on growth, researchers need to address the fact that selection into IMF programs is not random. Estimation methods differ in their treatment of this selection problem—we will try to determine whether this leads to estimates that vary systematically across different estimation techniques.

3 Heterogeneity in Estimates: a Simple Model

Estimates of the effects of IMF programs on growth may vary for a number of reasons. First, the 'true' underlying effects associated with IMF involvement may be different depending on the context. For example, they may depend on program features and the levels of economic and institutional development of the region under investigation. There may also be differences between short-run and long-run effects. Second, even if the underlying effects are roughly similar regardless of the context, there may still be systematic variation due to studies employing different estimation strategies. Countries *choose* whether to enter into IMF programs; this decision is likely based on a variety of factors, including countries' economic performance. If countries turn to the IMF when facing economic downturns (as is usually the case), then, on average, they are likely to have lower growth during the years of IMF involvement—not due to IMF policies, but because of the selection problem. More generally, to estimate the effect of an IMF program the researcher has to come up with an identification strategy; without it, the estimates are likely to exhibit a bias. Finally, the distribution of estimates that we observe may be affected by the preferences of those conducting the studies. In other words, the literature may be prone to selective reporting of the results.

In this section we attempt to disentangle the potential sources of variation in estimates of the effects of IMF programs on growth in participating countries. In subsection 3.1 we introduce a number of controls that, we believe, capture the key dimensions of studies' design (see summary in Appendix A). In subsection 3.2 we employ meta-regression analysis and investigate whether these study features have systematic effects on reported estimates.

3.1 The Controls

Publication and Data

The effects of IMF involvement may vary across countries. One piece of criticism often raised in the context of IMF lending is that access to IMF funds may enable policymakers to put off much needed reforms.[6] Countries with lower levels of institutional development would suffer the most from this type of moral hazard, and, conceivably, experience the least favorable growth outcomes (see discussions in Bordo & Schwartz 2000, Hutchison & Noy 2003). Binder & Bluhm (2017) find empirical support for this notion, reporting worse growth outcomes for countries with poorer institutional record. Furthermore, some authors credit IMF programs with catalyzing private capital flows—Mody & Saravia (2006) find that this catalytic effect depends on whether the IMF program adoption is

[6] See recent discussion in Bas & Stone (2014).

accompanied by a credible commitment to policy reforms by the government, a condition that may be less likely to be met by lower income countries with weaker institutions. Overall, both these arguments suggest that the effects of IMF programs may vary with countries' level of economic and institutional development. Here, we will introduce a broad control to account for these potential differences. While the majority of estimates in our sample is obtained with data coming from developing countries, some pertain to mixed samples of developing and advanced economies that are likely characterized by higher levels of overall development. We construct a control to distinguish between these two groups of estimates.

One difficulty with conducting a meta-analysis is that the sample of estimates under investigation is comprised of the results that authors *choose* to report. It is possible that the literature favors some results over others, and that estimates that are more in line with authors' preferences are more likely to be reported. In other words, the literature may be prone to selective reporting, or the publication bias.[7] We investigate this concern and do not find compelling evidence of selective reporting in our sample (see Appendix D). Nevertheless, we include a control to reflect whether any of the researchers on the author list were affiliated with the IMF (as indicated on each paper)—to capture possible differences in the preferred estimation outcomes across authors with and without the IMF affiliation.

Method

Estimates of the effects of IMF programs on growth can be obtained with a variety of methods, some of which are better suited for the task at hand compared to others (see discussions in Haque & Khan 1998, Bird 2001). For example, a researcher could estimate β^{IMF} in model (1) with an OLS. This method would be valid if countries selected into IMF programs independently of the state of the economy, an assumption that is difficult to justify because countries generally turn to the IMF for help in the years of economic turmoil. Since OLS does not take this selectivity into account, the estimates of β^{IMF} obtained with this method would likely exhibit a downward bias, as OLS would not correct for the fact that lower growth rates may prompt countries to participate in IMF programs in the first place (see Vreeland 2003 for a detailed discussion of the selection problem).

One way around this problem is to use instruments to predict participation in IMF programs, replacing IMF_{it} with a fitted value from the first-stage regression (Barro & Lee 2005,

[7] Brodeur *et al.* (2016) show that researchers tend to under-report results with p-values between 0.25 and 0.10, possibly engaging in specification searches to obtain results that are statistically significant. Ioannidis *et al.* (2017) argue that about 80% of the findings reported in economics literatures are exaggerated and single out meta-analysis as a tool for correcting such biases.

Dreher 2006). A sound identification strategy would require instruments that are good predictors of participation, but do not affect growth other than through their effect on IMF program adoption. Alternatively, a researcher could attempt to construct a counterfactual, an assessment of what would have happened if the country did not participate in the IMF program. Denote by $\Delta y_{\equiv} y_i^{post} - y_i^{pre}$ the difference between growth rates pre- and post- program participation. The before-after approach (BA) relies on the assumption that, within the participant group, if the countries did not choose to enter into IMF programs, the growth rates pre- and post-treatment would have been the same (i.e. Δy_i would have been zero in the counterfactual scenario). It proceeds to compute the effect of program participation by taking an average across all Δy_i. It is plain to see how the core assumption might fail: if there are forces that affect growth of participants that are not a direct consequence of participation, the estimate would be biased.

The difference-in-differences method (DID) is less restrictive. Instead of assuming that in the counterfactual the growth change would have been zero, it proxies for the counterfactual scenario by using growth rates of non-participants. The DID estimate of β^{IMF} would then be a difference in pre- and post- treatment mean growth rates between participants and non-participants. The identifying assumption here is that the participants would have had the same mean growth rates as non-participants, should they have chosen to not participate. This assumption would fail if events that take place between pre- and post-treatment periods have different effect on participants and non-participants (in other words, if the growth trends are not parallel). Hardoy (2003) provides a formal argument and derives the associated biases.

Matching approach attempts to explicitly construct a control group (see Bal-Gunduz 2016). At the first stage, a researcher estimates a model that explains countries' participation decisions via a set of observables and constructs propensity scores, a measure that reflects the likelihood of program participation for each country in a given period. In the second stage, participant observations are matched to the non-participant counterparts with similar propensity scores. The method would then calculate the difference in mean outcomes between participants and their non-participating matches. The key assumption here is that participation can be adequately explained with the observables used—only then the assembled control group would be providing a good counterfactual. A technical difficulty is that the group of non-participant 'matches' often turns out to be quite narrow, as it is difficult to find non-participant 'neighbors' for all of the participant observations (see discussion in Hardoy 2003).

Finally, a number of authors use the Generalized evaluation estimator (GEE) proposed by Goldstein & Montiel (1986) (e.g. Khan 1990 and Hutchison 2003). The GEE procedure consists of two steps. First, authors construct a policy reaction function that captures how key country variables respond to changes in growth rates, using country-years of non-participants. Second, they use the policy function from non-participants to approximate how the key policy variables in participant countries would have reacted if the countries stayed out of IMF programs. For this techniques to work, policy functions need to be stable across countries and time (see Dicks-Mireaux *et al.* 2000).

We construct a set of controls that documents the various strategies used to obtain estimates of the effect of IMF programs on growth. In the empirical section we will attempt to uncover any systematic difference across estimates produced with different methods.

Horizon

Another important consideration is whether to measure the effect of programs on short-run, medium-run or long-run growth figures. The results could potentially be different depending on the answer (see discussion in Bird & Rowlands 2017). Upon turning to IMF for help, member countries receive 25% of their quota in the Fund automatically, before implementing any policy changes. The rest of the funding becomes available gradually, conditional on the countries making progress in implementing policy measures specified in their agreement with the Fund (see Barro & Lee 2005). Depending on the arrangement, countries may be expected to 1) restrain demand in the short run (e.g. though fiscal or monetary contraction) and 2) implement structural reforms to promote sustainable growth in the long run (see Fischer 1997, Mussa & Savastano 1999).

Based on this dynamic, one can expect to see different effects at the very beginning of the program (when country receives the funding, but has not yet implemented any reforms), in the medium run (when country may attempt to implement the agreed-upon policy prescriptions, potentially with contractionary demand effects) and in the long run, after policy adjustments have been adopted. We introduce a set of controls to capture these differences.

Program

As mentioned above, IMF loan disbursement is conditional on countries meeting a set of requirements specified within the program. These requirements would often prompt participants to implement a series of reforms, addressing perceived economic vulnerabilities that hinder countries' development prospects, or the stability of government finances. This framework could benefit growth for a number of reasons: for example, in countries where governments are unable to implement reforms due to political opposition, IMF conditionality could accelerate the reform process. However, critics point out that some of these agreed-upon policy prescriptions—such as financial market liberalization for developing countries—may not in fact always be adequate.[8] It is therefore unclear whether, overall, the IMF conditionality benefits or harms economic growth (see discussions in Meltzer 2000, Bird 2001, Dreher 2006).

The kinds of conditions offered to program participants vary with program types. Some IMF programs are designed to address short-run balance of payments issues (e.g. the Stand-By Arrangements, SBA), with a goal of maintaining stability. By contrast, long-run programs (e.g. the Poverty Reduction and Growth Facility, PRGF) are aimed at decreasing poverty and fostering sustainable growth. Programs that target low-income countries and are meant to be offered on humanitarian grounds typically fit the latter category. They tend to offer funding from concessional (rather than general) resources with more lenient conditions and lower (or even zero) interest rates. Dreher (2004) investigates whether such programs are associated with looser conditionality. He finds evidence suggesting that the PRGF, a concessional resource program, indeed typically involves fewer conditions compared to other programs. In a similar vein, Stone (2008) finds that concessional programs tend to offer conditionality that is more narrow. Bird & Rowlands (2017) find that concessional programs lead to higher growth rates in low-income countries, while the non-concessional programs do not. We introduce a set of controls that capture whether the estimated effect of IMF programs on growth pertains to programs funded from general or concessional resources, or to a mixed sample of both.

We also consider an alternative way of capturing systematic difference across programs—by using information about intended program durations. The conditionality attached to short-run programs would often involve policies that restrain demand to address short-lived balance of payment imbalances (see, e.g. Mussa & Savastano 1999, Ghosh *et al.* 2005). Programs with longer duration may be expected to produce better growth outcomes, as the associated conditionality and funding are more likely to be tailored to the goal of promoting growth. We construct two alternative sets of controls that capture whether the study exclusively examines short-, or medium- to long-run IMF arrangements. The first set of controls relies on author's description in determining which group of arrangements is being studied—we prefer this approach as in some cases the arrangements with the same title may have different intended durations and goals. The second set of controls we construct uses the official classification based on program titles to record whether the programs fall into short-, or medium- to long-run category.

[8] Stiglitz (2002) lays out a detailed criticism of the neoliberal policies prescribed by the Washington Consensus and their adoption in developing countries. Broner & Ventura (2015) present a formal argument detailing how financial liberalization may destabilize domestic capital markets and lead to increased volatility of capital flows.

3.2 Results

We will now investigate whether the features of studies' design discussed in subsection 3.1 affect researchers' inference about the effects of IMF programs on economic growth. To this end, we will use meta-regression analysis, a tool often employed to disentangle sources of heterogeneity in estimates reported by various literatures.[9] We will estimate the following model:

$$\hat{\beta}_{ij}^{\mathrm{IMF}} = \gamma_0 + \sum_{l=1}^{11} \gamma_l X_{l,ij} + \epsilon_{ij}, \tag{2}$$

where $\hat{\beta}_{ij}^{\mathrm{IMF}}$ is the estimate i of β^{IMF} (i.e. the effect of IMF programs on growth) reported in study j; $X_{l,ij}$ are the controls discussed in previous section that capture the specific context in which study j obtains estimate i (see summary in Appendix A); ϵ_{ij} is the disturbance term. One important caveat related to estimating model (2) is that the estimates of β^{IMF} that researchers obtain are likely correlated within studies. We remedy this by clustering the standard errors at the study level; at the same time, we have a relative small number of clusters which means that the standard errors that result from clustered inference could potentially have a downward bias. We therefore follow Cameron *et al.* (2008) and additionally compute p-values using wild bootstrap cluster.

We start by examining a relatively homogenous subsample of estimates that measure the effect of IMF presence (as opposed to IMF loan amount or compliance with conditionality) on growth rates (as opposed to on changes in growth rates or differences between actual and projected rates). We will later make this exercise more general and extend the analysis to the full sample. We estimate model (2) and report the results in Table 2. Column (1) lists results for our baseline specification; columns (2)-(4) report results obtained with alternative controls for estimation horizon and IMF program types; column (5) reports results from an alternative specification in which we weight the data by the precision of each estimate thereby giving more precise estimates a higher weight.

We observe a positive association between authors' affiliation with the IMF and the magnitude of growth estimates being reported, but the effect is not very strong—at least not according to the p-values obtained with wild bootstrap clustering. This indicates that the large mean estimates produced by authors with IMF affiliation that we document in Table 1 may partially be due to these studies using specific methods and data. At the same time, we observe a stronger systematic effect related to the country sample under investigation: for the mix of developed and developing countries the estimates of program growth effects are higher compared to estimates obtained for developing countries

[9] For recent examples see Havránek (2015), Balima *et al.* (2017), Card *et al.* (2017), Doucouliagos *et al.* (2018), Sokolova & Sorensen (2018), Havranek & Sokolova (2020).

only. This may be because in countries with higher levels of economic and institutional development there is more public oversight and less scope for moral hazard related to IMF funding, which would make programs more efficient at promoting growth.

We find evidence pointing towards the importance of empirical techniques. As discussed in subsection 3.1, we control for six estimation methods that researchers employ: propensity score matching, BA, DID, IV, GEE and OLS as the reference group. In Table 2 we report the results for the propensity score matching; as we do not find any systematic effects pertaining to other method choices, here we suppress this output for brevity. The full table is available in Table D3 of Appendix D. We find a strong effect associated with the use of propensity score matching, i.e. with constructing control groups of non-participant 'neighbors' that are close to participants in their estimated likelihood of joining IMF programs. Studies that employ matching tend to come up with estimates that are higher by between 0.54-0.74 compared to estimates obtained with OLS. It is intuitive to see why an OLS estimation would measure effects of the IMF involvement with a downward bias: studies that do not control for the fact that countries tend to select into IMF programs in 'bad times' would then attribute the observed bad economic environment to the presence of the IMF. Matching alleviates this bias through an explicit modeling of the selection process. An advantage of matching compared to other methods is that it does not rely on finding variables that explain participation, but are unrelated to growth—unlike, for example, the IV. It is therefore possible that a large positive effect we observe is the result of matching providing a better correction for the endogenous participation decisions. That being said, the matching approach also has several important caveats.

Table 2: Heterogeneity: horizon and program definitions

Variable	(1)	(2)	(3)	(4)	(5)
Publication and Data					
IMF staff	0.509	0.629	0.519	0.474	-0.016
	(0.074)	(0.021)	(0.085)	(0.110)	(0.891)
	[0.348]	[0.130]	[0.380]	[0.429]	[0.855]
Mixed sample	0.626	0.554	0.541	0.624	0.230
	(0.094)	(0.164)	(0.112)	(0.090)	(0.023)
	[0.171]	[0.280]	[0.164]	[0.152]	[0.031]
Method					
Matching	0.544	0.569	0.594	0.621	0.737
	(0.002)	(0.001)	(0.000)	(0.002)	(0.000)
	[0.014]	[0.011]	[0.005]	[0.020]	[0.011]
Not displayed: Before-after, DID, Gen. equil., IV. See full results in Table D3 in Appendix D.					
Horizon					
Medium run effect	-0.348	.	-0.419	-0.360	-0.095
(between 2 and 9 years)	(0.139)	.	(0.146)	(0.126)	(0.000)
	[0.218]	.	[0.198]	[0.203]	[0.343]
Long run effect	0.907	.	0.972	1.026	0.956
(10 years and over)	(0.018)	.	(0.013)	(0.026)	(0.001)
	[0.297]	.	[0.244]	[0.215]	[0.202]
Medium run effect	.	-0.349	.	.	.
(between 2 and 5 years)	.	(0.259)	.	.	.
	.	[0.377]	.	.	.
Long run effect	.	0.517	.	.	.
(6 years and over)	.	(0.137)	.	.	.
	.	[0.404]	.	.	.
Program					
General res. only	-0.837	-0.951	.	.	-0.104
	(0.082)	(0.050)	.	.	(0.000)
	[0.125]	[0.086]	.	.	[0.523]
Concessional res. only	0.252	0.162	.	.	-0.076
	(0.194)	(0.428)	.	.	(0.000)
	[0.338]	[0.555]	.	.	[0.613]
Short run prog. only	.	.	-0.132	.	.
(official definition)	.	.	(0.512)	.	.
	.	.	[0.511]	.	.
Long run prog. only	.	.	0.357	.	.
(official definition)	.	.	(0.128)	.	.
	.	.	[0.294]	.	.
Short run prog. only	.	.	.	-0.640	.
(authors' definition)	.	.	.	(0.156)	.
	.	.	.	[0.261]	.
Long run prog. only	.	.	.	0.250	.
(authors' definition)	.	.	.	(0.241)	.
	.	.	.	[0.367]	.
N of clusters	27	27	27	27	27
Observations	684	684	684	684	684

Notes: The table presents the results of estimating model (2) with OLS with standard errors clustered at the study level. We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and compute p-values using `STATA` command `boottest`, with Rademacher weights and 9999 replications (see Roodman 2018); we report these p-values in square brackets. To produce results for this table, we only use a (relatively homogenous) sample of estimates that pertain to the effect of IMF *presence* on the economic growth (i.e. estimates obtained in a specification with dummy for IMF programs as explanatory variable). (1)=baseline specification; (2)=a specification with an alternative definition of the long run (6 years and over); (3)=a specification with alternative controls for program type, distinguishing between long and short run programs based on the official program definition; (4)=a specification with alternative controls for program type, distinguishing between long and short run programs based on the authors' program definition; (5)=a specification in which all data is weighted by precision, $1/SE(\beta^{\text{IMF}})$. Not displayed are results for alternative method choices (Before-after, DID, GEE, IV) as well as the constant. See full table in Table D3 of Appendix D.

Propensity score matching produces results free of bias if the participation equation captures well the selection process, and remains relatively stable over time and across countries. In doing so, it relies on the assumption that selection into IMF programs can be explained by the set of observables, which would fail if participation was also affected by unobservables (see e.g. Bal-Gunduz 2016, Bas & Stone 2014). Furthermore, as we discussed in subsection 3.1, for smaller datasets there may not be much overlap in the propensity scores of participant and non-participant observations. Under these conditions the effect of IMF programs would be calculated by using only a small sample of 'neighbors', or by expanding the definition of what researchers consider to be a 'match'. A related point raised by Atoyan & Conway (2006) is that, by design, the matching procedure excludes variation coming from country episodes with propensity scores of extreme values, such as participants with very high propensity scores for which it would typically be difficult to find a non-participant match. In other words, matching may be relying on samples of country-years that differ systematically from those used by other methods, samples that exclude participation episodes that are highly predictable: for example, country years characterized by severe crises, or countries that systematically rely on IMF support and therefore are very likely to enter IMF programs in the future. It is possible that the effects of IMF programs on growth are lower for these underrepresented episodes, and that their exclusion brings up the estimates produced with matching.

We also observe evidence of some systematic effects related to estimation horizon. In the baseline specification reported in column (1), we define 'medium run' as the estimation horizon from 2 to 9 years; we consider horizons of 10 years and over to measure 'long run' effects; the reference group contains estimates pertaining to the short run effect (under 2 years). In the long run, IMF programs appear to be associated with more positive effects on growth compared to the short-run—although this effect is statistically weaker under wild bootstrap clustering. For the medium run, we find a negative effect that is not statistically significant. We proceed to investigate how these conjectures change under an alternative definition of the effect horizons. In column (2) of Table 2 we report the results for a specification in which we define the 'long run' to include horizons over 5 years. We observe that the positive effect associated with the long run becomes much less prominent and looses statistical significance, while the (still statistically insignificant) negative effect in the medium run remains roughly the same. One possible interpretation of these results is that the estimation horizon has a nonlinear effect on the estimates of β^{IMF}: in the first two years, before full implementation of conditionality, IMF programs may affect growth through funding and signaling to private creditors, possibly generating a positive or a neutral effect; in the medium run as conditionality gets implemented this effect may slightly diminish because of policies restraining aggregate demand; however, in the long run, after the structural reforms have been implemented, this effect becomes more positive. These insights echo the results of Atoyan & Conway (2006), who find that, while there does not seem to be a positive contemporaneous effects of IMF programs on growth, IMF involvement tends to have positive effects on longer horizons.

We find some evidence suggesting that the effect of IMF programs on growth may be affected by the core features of program design, particularly by the kind of funding the program is offering. We use the reported program titles to determine whether the authors are studying programs that provide lending based on general resources, or the concessional window. The results that feature these controls are reported in columns (1), (2) and (5). We see a prominent negative effect for estimates obtained with general-resource programs (compared to estimates obtained using mixed program samples) and a slight positive (though not statistically significant) effect associated with concessional-resource programs. Programs with concessional lending typically target poor countries, they are more likely to be designed with the goal of reducing poverty and supporting sustainable growth; they also typically provide much more lenient funding opportunities with lower interest rates and longer grace periods. These features may make concessional programs more efficient at fostering growth compared to programs that use general resources, resulting in the discrepancy between the effects of different program types that we observe.

We repeat this exercise using alternative proxies for program design. In specification (3) we use the official program titles to construct controls for whether the program is intended for a short- or a medium- to long-run implementation. The signs of the estimated effects point toward longer-run programs having more positive effects on growth, although overall the results lack statistical significance. Nevertheless, this (albeit weak) evidence is consistent with our previous results: programs designed to address longer run structural problems may be more favorable for growth. Despite similarities in official program titles, program lengths may *de facto* differ. In specification (4) we introduce a robustness check, in which we construct controls for program design based on authors' descriptions of the programs in their sample (as opposed to the official program definitions). The main takeaways are quite similar to specifications presented earlier: estimates for programs with longer-term goals are higher compared to estimates obtained on a mixed sample of programs, while estimates corresponding to short-run programs are lower compared to the mixed sample—although statistically the effect is weak.

We conclude here that program design appears to matter, as longer-term programs that rely on concessional funding produce somewhat better growth outcomes compared to their shorter-term general resource counterparts. There are several possible interpretations for this finding. On the one hand, it may indicate that the conditionality attached to longer run programs is better aligned with the target of promoting growth. On the other hand, concessional programs tend to require less conditionality (see subsection 3.1 discussing Dreher 2004 and Stone 2008), and it is possible that the lack of stringent policy requirements is what leads to favorable growth outcomes under those programs. Finally, as concessional lending comes with more lenient credit conditions, the positive effects on growth we observe could simply result from countries receiving cheaper funding.

To check robustness of our conjectures, we compare the baseline model with the specification in which we weight each data point by the precision of the corresponding estimate, $1/SE(\hat{\beta}^{\mathrm{IMF}})$; we report the results in column (5) of Table 2. The intuitive appeal of this technique is that it gives more weight to estimates that are more precise while correcting for heteroskedasticity (see detailed discussion in Stanley & Doucouliagos 2015). There is also a drawback that is important for our context: suppose some methods generally yield higher standard errors compared to others; the results obtained via those methods would then receive lower weight compared to all other results. In our case, a fraction of estimates is obtained using instrumental variables, a technique that would typically yield higher standard errors compared to other methods. For this reason we use unweighted specifications as our baseline; however, we believe that a comparison can provide useful insights.

As before, in column (5) we see strong positive effects associated with the use of matching, with using a mixed sample of developing and developed countries (as opposed to focusing on developing countries only); we also observe a positive association with measuring the effects on long-run (as opposed to short-run) growth rates. The one result that appears to be challenged by this re-framing of the problem is the importance of program design, which looses in magnitude and statistical power. A likely explanation is that the effects we reported earlier were driven by variation coming from studies with higher standard errors. One important point to consider here is that studies that examine programs of specific type are likely to have less observations compared to those that do not discriminate between programs—and therefore, by design, would have lower precision relative to the latter group. The variation coming from these program-specific estimates would then receive a lower weight in the OLS weighted by precision, which could explain the discrepancy across specifications that we observe.

So far we focused on a narrow sample of estimates that only measure the effects of IMF presence on growth rates that is relatively homogenous. We now turn to investigate heterogeneity in estimates in a broader context. First, we expand our sample to include estimates that capture growth rate responses to IMF loan amounts; we report these results in Table 3, column (2); for comparison, we report the baseline results of Table 2 in column (1). Compared to estimates of IMF presence on growth, these new estimates do not appear systematically different, as the coefficient on the corresponding control, 'Exp. var.=Amount' is not statistically significant. At the same time, our previous conjectures about the effects associated with the use of propensity score matching, the estimation horizon and program types remain intact for this more heterogenous sample.

Once again, we find that IMF programs seem to be associated with more positive effects when averaging across mixed samples of developed and developing countries—as opposed to the samples of developing countries only. As we discussed earlier, this finding could suggest that institutional differences play an important role in determining whether IMF involvement leads to a success or a failure in terms of growth outcomes. We investigate this further by focusing on a subsample of 600 estimates that correspond exclusively to effects of IMF programs in developing countries. We report the results in Table 3, column (3); they are similar to those obtained for other specifications. One interesting observation is that for this subsample, the negative effect associated with the medium-run horizons (compared to short-run horizons) becomes somewhat more prominent, while the positive long-run effect slightly diminishes. This may mean that for countries with lower levels of institutional development the medium-run adjustment during which conditionality is being implemented tends to be more painful.

Finally, column (4) of Table 3 reports the results coming from the most heterogenous version of our dataset, that incorporating 8 estimates of the effect of compliance with IMF conditionality on growth that we collected, as well as estimates that use alternative definitions of the dependent variable discussed in subsection 3.1. Our previous findings appear resilient to this expansion, suggesting that, overall, the key underlying relationships between estimates and the context in which they are obtained remain largely the same, even when the precise definition of the dependent variable and the explanatory variable is altered. In the subsequent section we will use this expanded dataset to evaluate the impact of study design in the context of model uncertainty.

Table 3: Heterogeneity: sample splits

Variable	(1)	(2)	(3)	(4)
Publication and Data				
IMF staff	0.509	0.281	0.476	0.444
	(0.074)	(0.358)	(0.068)	(0.133)
	[0.348]	[0.600]	[0.282]	[0.370]
Mixed sample	0.626	0.504	.	0.522
	(0.094)	(0.059)	.	(0.058)
	[0.171]	[0.111]	.	[0.112]
Method				
Matching	0.544	0.639	0.493	0.636
	(0.002)	(0.001)	(0.006)	(0.001)
	[0.014]	[0.013]	[0.050]	[0.014]
Not displayed: Before-after, DID, Gen. equil., IV. See full results in Table D4 of Appendix D.				
Horizon				
Medium run effect	-0.348	-0.288	-0.491	-0.305
(between 2 and 9 years)	(0.139)	(0.178)	(0.084)	(0.161)
	[0.218]	[0.242]	[0.144]	[0.231]
Long run effect	0.907	0.985	0.767	0.700
(10 years and over)	(0.018)	(0.019)	(0.054)	(0.012)
	[0.297]	[0.217]	[0.440]	[0.256]
Program				
General res. only	-0.837	-0.554	-0.836	-0.535
	(0.082)	(0.044)	(0.031)	(0.060)
	[0.125]	[0.087]	[0.182]	[0.116]
Concessional res. only	0.252	0.124	0.129	0.136
	(0.194)	(0.388)	(0.410)	(0.351)
	[0.338]	[0.478]	[0.541]	[0.440]
Extra explanatory and dependent variables				
Exp. var.=Amount	.	0.025	0.139	0.064
	.	(0.943)	(0.718)	(0.853)
	.	[0.948]	[0.660]	[0.872]
Exp. var.=Compliance	.	.	.	0.062
	.	.	.	(0.917)
	.	.	.	[0.987]
Dep. var.=Growth change	.	.	.	0.058
	.	.	.	(0.844)
	.	.	.	[0.866]
Dep. var.=Growth difference	.	.	.	0.742
	.	.	.	(0.050)
	.	.	.	[0.166]
N of clusters	27	31	23	36
Observations	684	893	600	994

Notes: The table presents the results of estimating model (2) with OLS with standard errors clustered at the study level. We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and report the associated p-values in square brackets. (1)=baseline results—same as column (1) of Table 2—obtained using the sample of estimates that pertain to the effect of IMF *presence* on the economic growth (i.e. with dummy for IMF programs as explanatory variable); (2)=the sample of estimates from column (1) together with estimates measuring the effect of loan *amount* on growth (i.e. with loan amount as explanatory variable); (3)=the subsample of estimates from column (2) measuring effects of IMF *presence* and loan *amount*—for developing countries only; (4)=the full sample of estimates collected, that extends sample of column (2) by adding estimates obtained using compliance with conditionality as explanatory variable, as well as estimates corresponding to alternative dependent variables: growth change over time and difference between actual and projected growth. Not displayed are results for alternative method choices (Before-after, DID, GEE, IV) as well as the constant. See full table in Table D4 of Appendix D.

4 Heterogeneity in Estimates: Model Uncertainty

In Section 3 we studied how the estimates of the effect of the IMF involvement on growth vary with key choices that researchers make. We constructed a set of core controls that depicts the context of the studies in relatively broad strokes, highlighting what we believe are the most prominent aspects of study design. In doing so, we made a value judgment that is based on our prior beliefs about what these crucial aspects of study design are. In this section we will follow an alternative approach. Instead of pre-selecting a small set of most crucial features of study design, we will start with a bigger set of 25 controls that describe the study context in a greater level of detail, and allow for model uncertainty—in other words, we will take into account the fact that we do not know for certain which subset of these controls belongs to the 'true' data generating process.

In subsection 4.1 we introduce the extra explanatory variables that add more detail to our description of the studies' context. In subsection 4.2 we employ Bayesian Model Averaging to assess the relative performance of all 2^{25} possible combinations of the controls and construct inference by averaging results across all models, assigning higher weight to models that fit the data best. As a robustness check we report the results from estimating with LASSO, a method that offers an alternative solution to the model uncertainty problem. Finally, we construct our prediction of what estimates of the effect of the IMF involvement on growth would be given specific method and data choices.

4.1 The Additional Controls

Publication and Data

Over time, the emphasis that the IMF places on growth outcomes has increased. This shift was reflected through changes to the Fund's Guidelines on Conditionality, as well as through the subsequent introduction of Extended Fund Facility (EFT) in 1974 with longer program durations, the Enhanced Structural Adjustment Facility (ESAF) in 1987 and the Poverty Reduction and Growth Facility (PRGF) in 2000 aimed at promoting growth in low-income countries (see IEO 2019). The effects that IMF programs have on growth in participant countries may have changed following these institutional developments. To account for possible evolution of β^{IMF} over time, we record the average year of the data that was used to obtain each estimate.

In subsection 3.1 we introduced a broad control for whether the sample includes data from developing countries only, or a mix of developing and developed economies—to account for whether the effects of IMF programs differ with the overall levels of economic and institutional development. There may, however, also be systematic differences across specific geographical regions. Hutchison & Noy (2003) argue that the findings of negative effects of the IMF involvement on output growth are largely driven by experiences of Latin America, partially due to a prolonged history of economic volatility in the region and to low program completion rates. Gebregziabher (2015) finds mixed results while studying stabilization programs in the African region. We include controls to capture estimates of growth effects pertaining to Africa and Latin America.

We also include two controls to capture study quality. Other things equal, studies that rely on larger dataset are likely to have more statistical power—we control for the number of observations used by the studies. Furthermore, published studies are more likely to have gone through a peer review process—compared to the unpublished work; we add a corresponding control to reflect this distinction.

Finally, even though the majority of studies examine per capita growth rates that are most reflective of changes to the standards of living, a fraction of estimates in our sample is obtained using growth rates that do not account for population growth. We add a corresponding control.

Method

In subsection 3.1 we discussed the selection bias associated with countries entering IMF programs: namely, the fact that countries tend to turn to the IMF for help under dire circumstances. It is therefore important to model the countries' choice to participate in IMF programs. So far we touched on factors that affect the countries' demand for IMF loans; however, selection into IMF programs and the approved loan amounts likely also depend on factors related to loan supply. Barro & Lee (2005) point out that the political standing of countries with the IMF is likely an important determinant of IMF loan approval, loan size and the attached conditionality. They therefore suggest to use instruments that reflect the countries' political influence, such as the IMF quota size, the number of IMF staff members coming from the country, trade with the major economies as well as the extent to which country votes in the UN align with the votes of the major advanced economies—variables that also tend to be good predictors of program participation. We introduce a control variable that we term *Instruments: proximity* that equals one for studies that use any of these variables (that may reflect political proximity between the country and the Fund) as instruments for program participation decision or approved loan amount.

Aside from the countries' political influence, participation is likely to depend on the history of the IMF presence within the country. Assuming that this history does not have an effect on contemporary growth rates, the number of prior years the country spends in IMF programs can be used as an instrument for program participation.[10] We introduce

[10] Atoyan & Conway (2006) note that this assumption may be too strong, as IMF involvement may in fact have a lasting effect on growth rates.

a control that reflects whether this instrument was used to obtain a given estimate.

Finally, we add controls to capture whether the estimate was obtained within a specification featuring time and country fixed effects.

4.2 Results

With the addition of the controls discussed above the number of potential explanatory variables in our model goes up to 25. On the one hand, this adds a finer level of detail to our investigation of heterogeneity in estimates. On the other hand, we do not know *ex ante* which of the 25 potential controls belong to the 'true' data generating process for estimates of β^{IMF}. It is likely that not all of the proposed 25 variables contribute to the variation in estimates in a meaningful way; adding controls that do not have systematic effect on the estimates to our empirical model may render it misspecified. In other words, we are facing a problem of model uncertainty.

In Section 3 we studied a subset of control variables which we believe to be particularly relevant—thus addressing the model uncertainty problem by making a value judgment. In doing so we may have excluded some of the 25 explanatory variables that could also be important and included some variables that, in fact, do not have a systematic effect on estimates of β^{IMF}. In this section we will follow an alternative approach to resolving the model uncertainty problem that does not rely on such *ad hoc* assumptions: we will employ Bayesian Model Averaging (BMA).

The intuition behind the BMA approach can be summarized as follows. There are 2^{25} possible combinations of the 25 explanatory variables that we singled out (or 'models'). While some of these models may do better than others when it comes to explaining our data, we cannot know with certainty which, if any, specific combination of the control variables represents the 'true' data generating process—due to limited data. However, we can attempt to assign probabilities to each of the 2^{25} models (called Posterior Model Probabilities, or PMPs) that reflect how likely it is that each combination of control variables represents the underlying data generating process. We can then form an expectation about the 'true' coefficient values by estimating coefficients for all 2^{25} models and calculating a weighted average, using the models' posterior model probabilities as weights.

Fernández *et al.* (2001) employ BMA to tackle model uncertainty in cross-country growth regressions, while Havranek *et al.* (2015), Havranek *et al.* (2017) and Sokolova & Sorensen (2018) use it in the context of a meta-analysis similar to ours. Steel (2017) provides a detailed discussion of model uncertainty and model averaging in economics; Koop (2003) presents a formal introduction to BMA. To implement BMA, we take advantage of the `BMS` package for `R` developed by Zeugner & Feldkircher (2015).

Figure 2 presents the results of implementing BMA for our data. The figure gives an overview of the estimated model space. Each column appearing on the graph represents one model; a blank cell means that the corresponding explanatory variable (listed on the left) is not included in the model, while a colored cell means that the variable is included and the sign of the estimated coefficient is positive (for cells colored in blue, darker in grayscale) or negative (for cells colored in red, lighter in grayscale). The models are sorted based on their posterior model probability in the descending order; the horizontal axis reports the corresponding cumulative model probabilities.

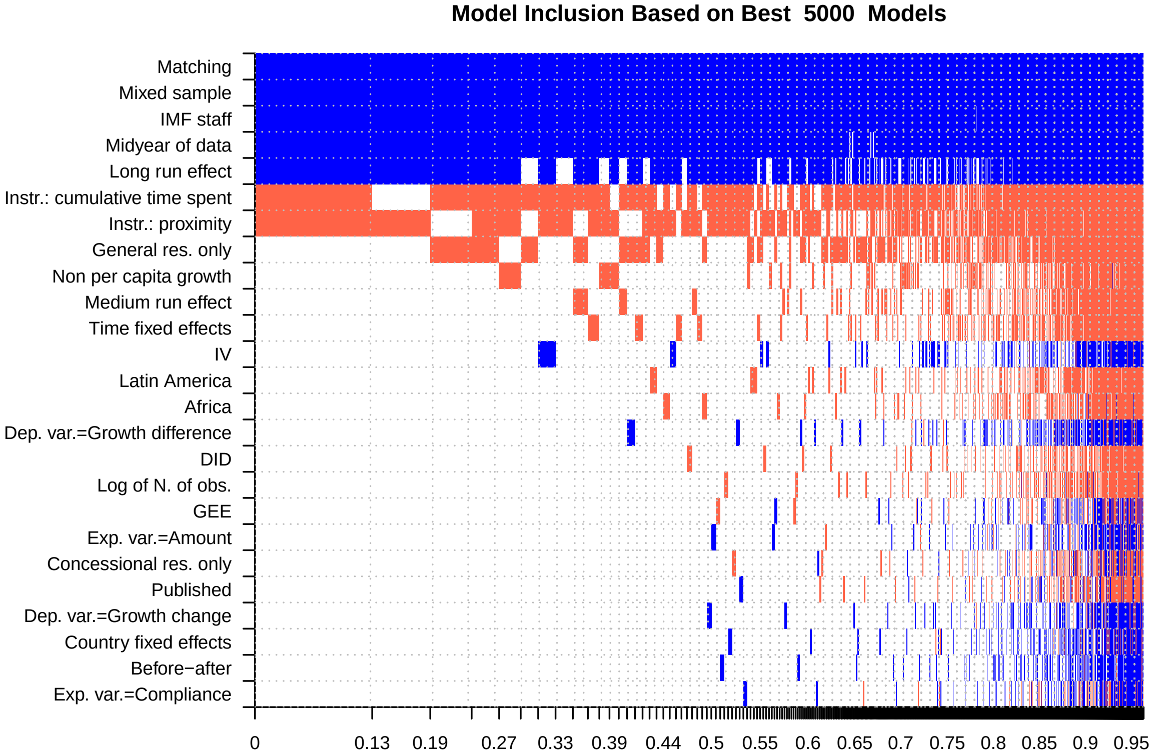
The model with the highest posterior model probability (of 13%) includes only a small subset of the 25 explanatory variables listed at the top of Figure 2. These variables tend to also appear in other models with high PMP, with signs that are consistent throughout the whole model space. By contrast, the variables listed towards the bottom of Figure 2 only appear in models that are the 'least likely', with signs that may change depending on the combination of explanatory variable that is being considered.

To make the comparison of the relative performance of our explanatory variables more formal, we evaluate the likelihood of each of the 25 controls belonging to the 'true' model—called Posterior Inclusion Probability, PIP—by adding up the PMPs of all models in which a given control is included. The variables displayed on the left on Figure 2 are listed based on their PIP, in the descending order. The subset of variables listed at the top (i.e. variables with the highest PIP) includes controls that also showed statistical significance in the OLS exercise reported in Table 2 and Table 3 of Section 3.

We report the BMA estimation results in the left panel of Table 4. For each explanatory variable, we report the posterior mean and standard deviation as well as the associated PIP that reflects how likely it is that the variable belongs to the 'true' model. To see how this compares with the frequentist approach, we also include a robustness check in which we perform an OLS with a set of variables for which the posterior inclusion probability exceeds 50%; we report these results in the right panel of Table 4. The BMA results seem consistent with the insights of Section 3: as before, the effect of IMF programs on growth appears to be more positive when estimated on the mixed sample of developing and developed countries and with the use of the propensity score matching technique. It is also more positive in the long run (compared to the short run).

Compared to the analysis of Section 3, here we see less evidence of heterogeneous effects across program types: even though, as before, the point estimate associated with programs that rely on general resources is negative, the likelihood of this control belonging to the 'true' data generating process is estimated to be relatively low (around 39%). As before, the evidence presented here suggests that researchers who are affiliated with the IMF tend to report estimates that are higher than the results reported by those

Figure 2: Bayesian Model Averaging and variable inclusion



Notes: The figure depicts the results of estimating different variable combinations with the BMA, obtained using the `BMS` package for `R` developed by Zeugner & Feldkircher (2015). Each column represents one model. Cells that are white indicate that a variable listed on the vertical axis is not included in the given model; a red cell (lighter in grayscale) means that the variable is included with a negative corresponding effect, while a blue cell (darker in grayscale) indicates that the variable is included, and the estimated effect is positive. Along the vertical axis the explanatory variables are sorted by their posterior inclusion probability, with variables that are most likely to belong in the 'true' data generating process listed at the top. Along the horizontal axis the models are sorted by their posterior model probability, with models that are most likely to capture the underlying data generating process depicted on the left. We report the corresponding numerical results in Table 4. A detailed description of all variables is available in Table A1.

without the IMF affiliation. Compared to previous analysis, this result appears much more statistically significant, with high posterior inclusion probability in the BMA and low p-values in the frequentist check.

Table 4: Heterogeneity and model uncertainty: BMA

	BMA			OLS with selected variables		
Response variable:	Post. Mean	Post. SD	PIP	Coef.	P-value	P-value (wild)
Publication and Data						
IMF staff	0.571	0.149	0.995	0.543	0.002	0.046
Mixed sample	0.525	0.133	0.996	0.559	0.011	0.027
Latin America	-0.029	0.134	0.070			
Africa	-0.013	0.072	0.059			
Midyear of data	0.028	0.008	0.981	0.028	0.000	0.001
Log of N. of obs.	-0.002	0.014	0.041			
Non per capita growth	-0.025	0.075	0.134			
Published	-0.001	0.021	0.032			
Method						
Matching	0.640	0.136	0.999	0.658	0.000	0.006
Before-after	0.003	0.043	0.029			
DID	-0.005	0.039	0.041			
GEE	0.001	0.054	0.036			
IV	0.015	0.065	0.078			
Instruments: proximity	-0.433	0.325	0.706	-0.600	0.176	0.189
Instruments: cumulative time spent	-1.539	0.976	0.790	-1.755	0.025	0.225
Time fixed effects	-0.028	0.093	0.114			
Country fixed effects	0.001	0.018	0.030			
Horizon						
Medium run effect	-0.023	0.076	0.116			
Long run effect	0.476	0.293	0.801	0.610	0.010	0.135
Program						
General res. only	-0.140	0.198	0.387			
Concessional res. only	-0.001	0.033	0.033			
Extra explanatory and dependent variables						
Exp. var.=Amount	0.002	0.028	0.034			
Exp. var.=Compliance	0.000	0.062	0.027			
Dep. var.=Growth change	0.002	0.036	0.031			
Dep. var.=Growth difference	0.024	0.145	0.059			
Const.	-0.429	.	1.000	-0.492	0.001	0.005
N of clusters	36			36		
Observations	994			994		

Notes: Here we report the numerical results for Bayesian Model Averaging estimation corresponding to Figure 2. The left panel presents the BMA results; SD denotes standard deviation, PIP stands for the posterior inclusion probability. The right panel presents a frequentist check, in which we select a set of variables that, according to BMA, belong to the 'true' data generating process with the likelihood above 50% (i.e. $PIP > 0.5$). We run an OLS estimation with clustered standard errors using these controls. We also compute wild bootstrap clustered p-values and report them under 'P-value (wild)'. A detailed description of all variables is available in Table A1.

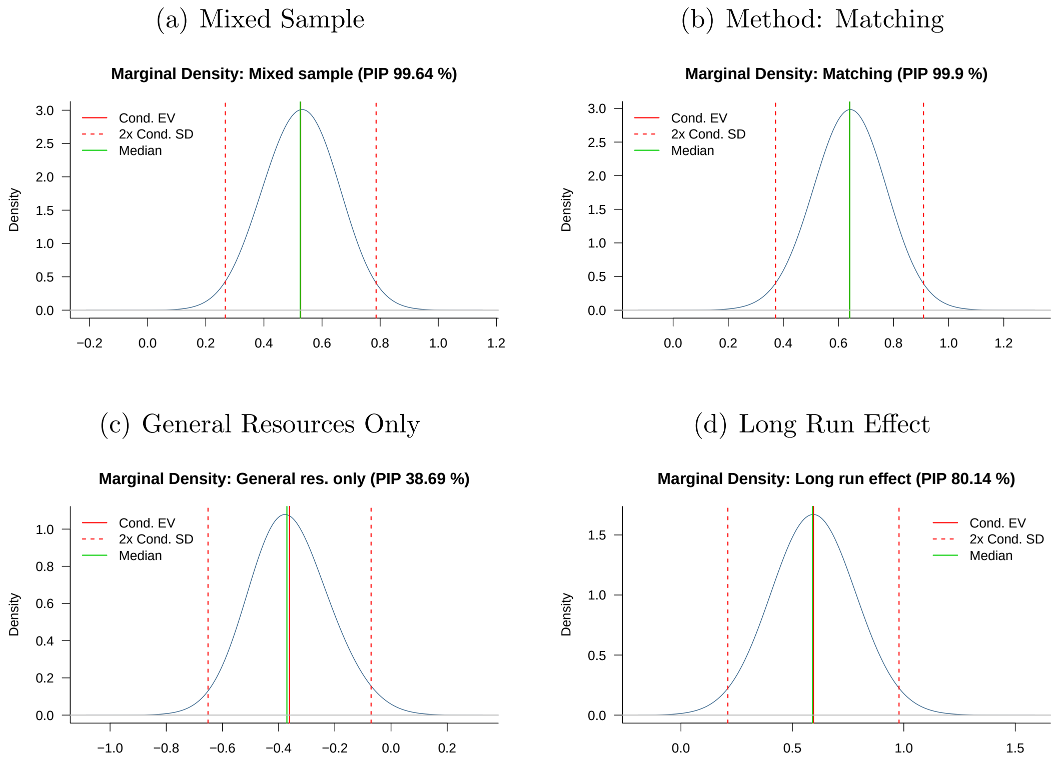
Furthermore, we find that several of the new explanatory variables that we add seem to have a systematic effect on estimates. First, for studies that employ more recent data the estimated growth effect appears to be higher, as evidenced by a positive coefficient with high PIP on the variable *Midyear of data*. This may imply that over time, as the IMF agenda became more focused on fostering growth, the programs started to yield better growth outcomes. Alternatively, the positive evolution of β^{IMF} over time may be due to the overall increase in the amount of funding distributed through IMF-supported programs that occurred in the last decade (see Figure 1).

Second, we find that studies that use instrumental variables that reflect countries' political influence (*Instruments: proximity*) and the countries' history with the IMF (*Instruments: cumulative time spent*) tend to yield lower estimates of the effect of IMF programs on growth. Previously we discussed how the fact that countries choose to participate when their economies are in peril means that researchers need to model participation decisions that countries make—otherwise the effect of participation on growth would be understated. A similar supply-side argument could also be made: it is possible that IMF's decisions to approve loans depend on country characteristics, and countries that are, for example, more likely to implement beneficial reforms and quickly restore growth get approved more often, receiving more generous funding and negotiating conditionality that is less stringent. If this was the case, then comparing participants and non-participants without modeling the IMF's decision-making process could overstate the effect of IMF programs on growth. The instruments listed above could partially capture the likelihood of future participation and negate some of this upward bias—their use would then result in systematically lower growth estimates. That being said, the effects associated with the use of these instruments are not very statistically significant. Furthermore, the effect pertaining to *Instruments: cumulative time spent* is obtained using variation from a small number of estimates and studies, as the number of observations for which this control equals one in our sample is very low.

The estimates reported in Table 4 reflect the unconditional means of the coefficients, i.e. means averaged across *all* models (and weighted by the posterior model probabilities), including the models in which the corresponding variable is not included. For variables with high PIPs (i.e. those that are present in models with high PMPs), the unconditional means should be relatively close to means that are conditioned on the variable being included. By contrast, for variables that are absent from some of the 'good' models, there will likely be a difference between the two means, with the unconditional mean being closer to zero.

Figure 3 displays posterior coefficient densities for four of our variables, conditional on each variable being included; it also plots the coefficient means conditional on variable inclusion. Comparing these results with the unconditional means reported in Table 4 helps explain the gap between the outcomes of the BMA and the results obtained in Section 3 in which we make explicit variable inclusion assumptions. For variables *Mixed Sample* and *Method: Matching* displayed in the top row the posterior inclusion probabilities reported in Table 4 are close to 1; for these variables, the expected coefficient values conditional on inclusion shown on Figure 3 are very close to those reported in Table 4. Conversely, for the variables *General Resources Only* and *Long Run Effect* shown at the bottom, the conditional means are further away from zero than the unconditional means of Table 4. This distinction arises because the posterior inclusion probabilities for these

Figure 3: Posterior coefficient distributions for selected variables



Notes: The figure depicts posterior densities of the effects of *Mixed Sample*, *Method: Matching*, *General Res. Only*, *Long Run Effect*; these results are conditional on variable inclusion, i.e. for each variable, they are computed using all models in which the variable is present. We perform this computation using the `BMS` package for `R`.

controls are markedly lower than 1, as the variables are absent from some of the 'good' models. However, once we make an assumption that these variables belong to the data generating process (and, therefore, condition on inclusion), their expected contribution to the estimated effects of IMF programs on growth becomes much more pronounced. The conditional means for *General Resources Only* and *Long Run Effect* are much closer to the OLS estimates reported in Table 2 and Table 3 of Section 3, because we obtain the latter results under a similar inclusion assumption. For all four variables the conditional means lie more than two standard deviations away from zero, indicating that, conditional on inclusion, each variable would have a prominent impact in a Bayesian setting—much like in the frequentist setting of Section 3.

An important policy question that we still need to address is what these results imply about the overall effect of IMF involvement on growth. For this purpose, we use the results from the frequentist check reported in Table 4 to construct our prediction for $\hat{\beta}^{\mathrm{IMF}}$, for different country groups, techniques and horizons. We construct our baseline estimates using mean values of all explanatory variables, except for *Midyear of data*. As we are primarily interested in the effect of IMF programs in the modern days, we put the value of the 90s percentile for *Midyear of data* instead of the mean.[11]

Table 5: Fitted Estimates by Group

Group	Point Estimate	95% interval	95% interval (wild)
All	0.532	[0.395; 0.670]	[0.346; 0.711]
Country groups			
Developing only	0.353	[0.134; 0.572]	[-0.139; 0.775]
Mixed sample	0.913	[0.623; 1.203]	[0.608; 1.796]
Technique			
All: Matching	1.122	[0.903; 1.341]	[0.762; 1.509]
All: Non-matching	0.389	[0.224; 0.555]	[0.151; 0.641]
All: IV with proximity instruments	-0.145	[-1.018; 0.728]	[-2.093; 0.724]
Horizon			
Long Run effect only	1.096	[0.637; 1.556]	[-0.460; 2.256]
Short and Medium Run effect	0.486	[0.347; 0.625]	[0.266; 0.645]

Notes: The table presents fitted values of estimates of the IMF effects on growth, for different sources of data, estimation techniques and measurement horizons. The estimates are obtained using the frequentist check model reported in Table 4. To obtain the fitted values, we substitute sample means for all variables, except for those related to the category under consideration and the *Midyear of data*: we use the value of the 90s percentile for the latter. Results using mean value for the *Midyear of data* are reported in Table B1 of Appendix B. We report fitted values, 95% confidence intervals and confidence intervals constructed with wild bootstrap cluster.

We report the estimates and the corresponding 95% confidence intervals in Table 5. Overall, the effect appears to be positive: on average, IMF presence tends to boost growth rates by about 0.53 percentage points. That being said, the context matters. For a sample of developing countries, the estimates of the effect tend to be lower compared to estimates obtained with data coming from a mix of developing and developed economies. As we noted earlier, this discrepancy could be due to underlying differences in institutional development across the country groups. Furthermore, the results seem to be greatly affected by the choice of the estimation technique, particularly by whether the study performs propensity score matching, or uses instruments that reflect political proximity between the IMF and the program participants—although the latter effect is not very precise, as the confidence intervals are quite broad. We also observe a difference between point estimates obtained for long- and short- to medium-run horizons, albeit the long-run effect is associated with wide confidence intervals based on wild bootstrap cluster.

One potential concern related to using Bayesian Model Averaging is that the posterior

[11] See B1 in Section Appendix B for a version in which we assign mean value for the *Midyear of data*. The resulting point estimates are somewhat smaller compared to those displayed in Table 5.

inference may be sensitive to the choice of a specific prior for parameters and model space. In particular, Ciccone & Jarociński (2010) argue that BMA results for growth regressions obtained with agnostic priors are very sensitive to data revisions in Penn World Table. Feldkircher & Zeugner (2012) point out that this is because certain prior choices may attribute very high weight to a few 'best' models, skewing the distribution of the posterior model probabilities and amplifying noise in the data. They suggest to remedy this problem by employing flexible data-dependent priors that are more resilient to noise in the data. Figure B1 in Appendix B shows how our results change under different prior assumptions. The posterior means of the coefficients appear to be very similar under different priors. One notable difference is that the posterior inclusion probabilities are generally higher under the flexible prior suggested in Feldkircher & Zeugner (2012) (see 'HyperBRIC and Random' on Figure B1). This means that our baseline prior choice may, in fact, result in us undervaluing the impact of some of the explanatory variables—most notably, the variable *General res. only*, for which the PIP turns out to be much higher under the 'HyperBRIC' prior for model parameters.

Overall, comparing the BMA results with those obtained in Section 3, we note several discrepancies; this suggests that model uncertainty likely matters for the problem at hand. We now examine how these results change when we address model uncertainty with an alternative approach, and employ LASSO. LASSO is designed to be used as a model selection tool. The coefficients in LASSO are obtained by solving a constrained minimization problem that yields corner solutions, setting some of the coefficients to exact zeros (see further details in Appendix B.2). We apply LASSO to our data and report the results in Table B2 of Appendix B.2.

All variables that were estimated to have a posterior inclusion probability over 50% in the BMA are also selected by LASSO. In addition to variables that proved important in the BMA, LASSO also selects the control for whether the study only considers general resource programs; similar to the OLS results of Section 3, the point estimate is negative. Furthermore, LASSO chooses to include the control for Medium Run effects with a negative sign, although in the post-LASSO estimation this coefficients does not appear statistically significant; this provides some evidence suggesting that the horizon effects may be nonlinear. LASSO also chooses to include geographic controls for estimates pertaining exclusively to data from Africa and Latin America, both with negative coefficients. As a result, the difference between estimates for a mixed sample of countries and estimates for developing countries appears less pronounced, likely because part of the negative effect associated with developing countries' data is now attributed to these geographical regions.

We compute the fitted estimates of the effects of IMF programs on growth implied by LASSO for different method and data choices and report them in Table B3. Because LASSO selects additional controls, we are able to provide fitted values for additional groups of estimates—e.g. estimates for general resource programs, medium run, etc. Some of the resulting fitted values are associated with wide confidence intervals indicating that the inference is not based on many observations and should be interpreted with caution. Overall, the point estimates reveal patterns similar to those displayed in Table 5.

5 Conclusion

In this paper we present the first meta-analysis of the literature estimating the effect of IMF-supported programs on economic growth, taking stock of four decades of heated debates among academic researchers and IMF staff members. We collect 994 estimates from 36 studies and use meta-regression analysis to detect sources of systematic variation in the results that are being reported. Estimates of the effect of IMF programs on growth may vary because the underlying 'true' effect depends on the specific context in which it is being measured; estimates may also change depending on the estimation techniques that are being used; finally, the distribution of estimates we observe may be affected by preferences of those conducting the research. We attempt to disentangle these three sources of variation and pin down specific aspects of study context, empirical techniques and researchers' preferences that have systematic effects on the estimates reported by this literature.

First, we show that the study context is important. Samples that include a mix of developed and developing countries tend to yield higher estimates of the effects of IMF programs on growth compared to samples of developing countries only, which may suggest that IMF involvement is more successful in economies with strong institutional record. We also find that the effect that IMF programs have on growth appears to have improved over time, possibly due to the IMF's increased focus on growth outcomes in the recent years. We also show that the growth effects following IMF involvement depend on the horizon on which they are being measured (with more favorable effects in the long run) and on the types of programs being investigated (with general resource programs leading to less favorable outcomes).

Second, we find that estimation results depend on the estimation techniques that are being used. In particular, propensity score matching seems to produce estimates that are systematically higher compared to those obtained using other methods. On the one hand, this may mean that this method is particularly successful at addressing the endogenous participation in IMF program with its selection on observables assumption. On the other hand, this discrepancy could also arise because matching may systematically discard the participation observations for which there are no non-participant neighbors, thus considering a sample that is systematically different from those studied with other methods.

Third, while we do not find compelling evidence of systematic publication bias in the literature, we document that estimates reported by IMF staff members tend to be somewhat higher than those reported in papers authored by researchers without the IMF affiliation. This effect remains intact after we control for the different choices researchers make with respect to their studies' context and estimation techniques; although this effect is not very prominent, we believe it warrants further investigation by future researchers.

Having taken these various features of study design into account, we present a holistic view on what the wealth of the accumulated evidence implies about the magnitude and the direction of the effect of IMF involvement on growth—conditional on specific data and method choices. We find that, overall, the effect appears to be positive, especially when the inference is based on more recent data from mixed samples of developing and developed countries. In this context, IMF programs would tend to increase growth rates by about 0.913 percentage points. Importantly, this magnitude would be lower given some other data choices, such as samples of developing countries only or short- and medium-run estimation horizons.

As IMF programs are now prevalent in many countries around the world, it is crucial to understand how they influence countries' economic performance. Our results highlight the importance of the context in which the effects of IMF programs are being evaluated. Consequently, our recommendation for future studies is to report robustness checks with respect to the elements of study design that we show to have systematic bearing on the estimates of the effects of IMF programs on growth.

References

Atoyan, R. & P. Conway (2006): "Evaluating the impact of IMF programs: A comparison of matching and instrumental-variable estimators." *The Review of International Organizations* **1(2)**: pp. 99–124.

Bal-Gunduz, Y. (2016): "The economic impact of short-term imf engagement in low-income countries." *World Development* **87**: pp. 30–49.

Balima, H. W., E. G. Kilama, & R. Tapsoba (2017): "Settling the Inflation Targeting Debate: Lights from a Meta-Regression Analysis." *IMF Working Papers 17/213*, International Monetary Fund.

Baqir, R., R. Ramcharan, & R. Sahay (2005): "Imf programs and growth: is optimism defensible?" *IMF Staff Papers* **52(2)**: pp. 260–286.

Barro, R. J. & J.-W. Lee (2005): "IMF programs: Who is chosen and what are the effects?" *Journal of Monetary Economics* **52(7)**: pp. 1245–1269.

Bas, M. A. & R. W. Stone (2014): "Adverse selection and growth under imf programs." *The Review of International Organizations* **9(1)**: pp. 1–28.

Belloni, A., D. Chen, V. Chernozhukov, & C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica* **80(6)**: pp. 2369–2429.

Binder, M. & M. Bluhm (2017): "On the conditional effects of imf program participation on output growth." *Journal of Macroeconomics* **51**: pp. 192–214.

Bird, G. (2001): "Imf programs: Do they work? can they be made to work better?" *World Development* **29(11)**: pp. 1849–1865.

Bird, G. & D. Rowlands (2017): "The effect of imf programmes on economic growth in low income countries: An empirical analysis." *The Journal of Development Studies* **53(12)**: pp. 2179–2196.

Bordo, M. D. & A. J. Schwartz (2000): "Measuring real economic effects of bailouts: historical perspectives on how countries in financial distress have fared with and without bailouts." In "Carnegie-Rochester Conference Series on Public Policy," volume 53, pp. 81–167. Elsevier.

Brodeur, A., M. Lé, M. Sangnier, & Y. Zylberberg (2016): "Star wars: The empirics strike back." *American Economic Journal: Applied Economics* **8(1)**: pp. 1–32.

Broner, F. & J. Ventura (2015): "Rethinking the Effects of Financial Liberalization." *Working Papers 509*, Barcelona Graduate School of Economics.

Cameron, A. C., J. B. Gelbach, & D. L. Miller (2008): "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics* **90(3)**: pp. 414–427.

Card, D., J. Kluve, & A. Weber (2017): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the European Economic Association* **16(3)**: pp. 894–931.

Card, D. & A. B. Krueger (1995): "Time-series minimum-wage studies: A meta-analysis." *American Economic Review* **85(2)**: pp. 238–43.

Ciccone, A. & M. Jarociński (2010): "Determinants of economic growth: Will data tell?" *American Economic Journal: Macroeconomics* **2(4)**: pp. 222–246.

Dicks-Mireaux, L., M. Mecagni, & S. Schadler (2000): "Evaluating the effect of imf lending to low-income countries." *Journal of Development Economics* **61(2)**: pp. 495–526.

Disdier, A.-C. & K. Head (2008): "The Puzzling Persistence of the Distance Effect on Bilateral Trade." *The Review of Economics and Statistics* **90(1)**: pp. 37–48.

Doucouliagos, C., P. Laroche, D. L. Kruse, & T. D. Stanley (2018): "Where Does Profit Sharing Work Best? A Meta-Analysis on the Role of Unions, Culture, and Values." *IZA Discussion Papers 11617*, Institute of Labor Economics (IZA).

Doucouliagos, H. & T. D. Stanley (2009): "Publication selection bias in minimum-wage research? A meta-regression analysis." *British Journal of Industrial Relations* **47(2)**: pp. 406–428.

Dreher, A. (2004): "A public choice perspective of imf and world bank lending and conditionality." *Public Choice* **119(3/4)**: pp. 445–464.

Dreher, A. (2006): "IMF and economic growth: The effects of programs, loans, and compliance with conditionality." *World Development* **34(5)**: pp. 769–788.

Easterly, W. (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)**: pp. 1–22.

Egger, M., G. D. Smith, M. Scheider, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *British Medical Journal* **316**: pp. 629–634.

Eicher, T. S., C. Papageorgiou, & A. E. Raftery (2011): "Default priors and predictive performance in bayesian model averaging, with application to growth determinants." *Journal of Applied Econometrics* **26(1)**: pp. 30–55.

Feldkircher, M. & S. Zeugner (2012): "The impact of data revisions on the robustness of growth determinants—a note on 'determinants of economic growth: Will data tell?'." *Journal of Applied Econometrics* **27(4)**: pp. 686–694.

Fernandez, C., E. Ley, & M. F. J. Steel (2001): "Benchmark priors for bayesian model averaging." *Journal of Econometrics* **100(2)**: pp. 381–427.

Fernández, C., E. Ley, & M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions." *Journal of Applied Econometrics* **16(5)**: pp. 563–576.

Fischer, S. (1997): "Applied economics in action: Imf programs." *American Economic Review, P&P* **87(2)**: pp. 23–27.

Gebregziabher, F. (2015): "Adjustment and Long?Run Economic Performance in 18 African Countries." *Journal of International Development* **27(2)**: pp. 170–196.

Ghosh, A. R., C. Christofides, J. Kim, L. Papi, U. Ramakrishnan, A. Thomas, & J. Zalduendo (2005): *The design of IMF-supported programs*, volume 241. International Monetary Fund.

Goldstein, M. & P. Montiel (1986): "Evaluating fund stabilization programs with multicountry data: Some methodological pitfalls." *IMF Staff Papers* **33(2)**: pp. 304–344.

Haque, N. U. & M. S. Khan (1998): "Do IMF-Supported Programs Work? A Survey of the Cross-Country Empirical Evidence." *IMF Working Papers 98/169*, International Monetary Fund.

Hardoy, I. (2003): "Effect of IMF programs on growth: A reappraisal using the method of matching." *Institute for Social Research Paper.*

Havranek, T., R. Horvath, Z. Irsova, & M. Rusnak (2015): "Cross-country heterogeneity in intertemporal substitution." *Journal of International Economics* **96(1)**: pp. 100–118.

Havranek, T., M. Rusnak, & A. Sokolova (2017): "Habit formation in consumption: A meta-analysis." *European Economic Review* **95**: pp. 142 – 167.

Havranek, T. & A. Sokolova (2020): "Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say 'Probably Not'." *Review of Economic Dynamics* **35**: pp. 97–122.

Havránek, T. (2015): "Measuring Intertemporal Substitution: The Importance Of Method Choices And Selective Reporting." *Journal of the European Economic Association* **13(6)**: pp. 1180–1204.

Hutchison, M. (2003): "A Cure Worse Than the Disease? Currency Crises and the Output Costs of IMF-Supported Stabilization Programs." In "Managing Currency Crises in Emerging Markets," NBER Chapters, pp. 321–360. National Bureau of Economic Research, Inc.

Hutchison, M. M. & I. Noy (2003): "Macroeconomic effects of IMF-sponsored programs in Latin America: output costs, program recidivism and the vicious cycle of failed stabilizations." *Journal of International Money and Finance* **22(7)**: pp. 991–1014.

IEO (2018): "Structural conditionality in imf-supported programs." , Independent Evaluation Office of the International Monetary Fund, DC. Evaluation update 2018.

IEO (2019): "Adjustment and growth in imf-supported programs." , Independent Evaluation Office of the International Monetary Fund, DC. July 31, 2019.

IMF (2005): "Review of the 2002 conditionality guidelines." , International Monetary Fund, Washington, DC.

Ioannidis, J. P. A., T. D. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *Economic Journal* **127(605)**: pp. 236–265.

Khan, M. S. (1990): "The macroeconomic effects of fund-supported adjustment programs." *IMF Staff Papers* **37(2)**: pp. 195–231.

Khan, M. S. & M. D. Knight (1981): "Stabilization programs in developing countries: A formal framework (programmes de stabilisation dans les pays en développement: cadre formel) (programas de estabilización en los países en desarrollo: Un marco formal)." *Staff Papers (International Monetary Fund)* **28(1)**: pp. 1–53.

Killick, T. (1995): *IMF Programmes in Developing Countries - Design and Impact.* London: Routledge.

Koop, G. (2003): *Bayesian Econometrics.* John Wiley & Sons.

Ley, E. & M. F. Steel (2009): "On the effect of prior assumptions in Bayesian model averaging with applications to growth regressions." *Journal of Applied Econometrics* **24(4)**: pp. 651–674.

MELTZER, A. (2000): "The report of the international financial institution advisory commission: comments on the critics; reform of the international architecture." In "CESifo Forum," volume 1, pp. 9–17. München: ifo Institut für Wirtschaftsforschung an der Universität München.

MERCER-BLACKMAN, V. & A. UNIGOVSKAYA (2004): "Compliance with imf program indicators and growth in transition economies." *Emerging Markets Finance and Trade* **40(3)**: pp. 55–83.

MODY, A. & D. SARAVIA (2006): "Catalysing Private Capital Flows: Do IMF Programmes Work as Commitment Devices?" *Economic Journal* **116(513)**: pp. 843–867.

MUSSA, M. & M. SAVASTANO (1999): "The IMF Approach to Economic Stabilization." *NBER Macroeconomics Annual* **14**: pp. 79–122.

PASTOR, M. (1987): "The effects of imf programs in the third world: Debate and evidence from latin america." *World Development* **15(2)**: pp. 249 – 262.

ROODMAN, D. (2018): "Boottest: Stata module to provide fast execution of the wild bootstrap with null imposed."

ROSE, A. K. & T. D. STANLEY (2005): "A Meta-Analysis of the Effect of Common Currencies on International Trade ." *Journal of Economic Surveys* **19(3)**: pp. 347–365.

SOKOLOVA, A. & T. A. SORENSEN (2018): "Monopsony in Labor Markets: A Meta-Analysis." *IZA Discussion Papers 11966*, Institute of Labor Economics (IZA).

STANLEY, T. D. (2005): "Beyond publication bias." *Journal of Economic Surveys* **19(3)**: pp. 309–345.

STANLEY, T. D. & H. DOUCOULIAGOS (2015): "Neither fixed nor random: Weighted least squares meta-analysis." *Statistics in Medicine* **34(13)**: pp. 2116–27.

STEEL, M. F. J. (2017): "Model Averaging and its Use in Economics." *MPRA Paper 81568*, University Library of Munich, Germany.

STIGLITZ, J. E. (2002): *Globalization and its discontents / Joseph E. Stiglitz.* W. W. Norton New York, 1st ed. edition.

STONE, R. W. (2008): "The scope of imf conditionality." *International Organization* **62(4)**: pp. 589–620.

TIBSHIRANI, R. (1996): "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* **58(1)**: pp. 267–288.

VREELAND, J. R. (2003): *The IMF and economic development.* Cambridge University Press.

ZEUGNER, S. & M. FELDKIRCHER (2015): "Bayesian model averaging employing fixed and flexible priors – the BMS package for R." *Journal of Statistical Software* **68(4)**: pp. 1–37.

Appendix A Description of Variables

Table A1: Definitions and summary statistics of explanatory variables

Variable	Description	Mean	Std. dev.
Publication and Data			
IMF staff	=1 if IMF staffer is on a co-author list (as indicated on the paper).	0.216	0.412
Mixed sample	=1 if sample includes developing and developed countries (baseline: sample includes only developing countries).	0.320	0.467
	Additional variables in Section 4:		
Latin America	=1 if sample only includes countries from Latin America.	0.021	0.144
Africa	=1 if sample only includes countries from Africa.	0.093	0.290
Midyear of data	The average year of the data used minus 1976 (i.e. the earliest average midyear in our sample).	15.766	6.515
Log of N. of obs.	The logarithm of the number of observations.	6.615	1.219
Non per capita growth	=1 if the dependent variable is constructed using non-per capita growth rates (baseline: dependent variables constructed using per capita growth rates).	0.409	0.492
Published	=1 if study is published in a peer-reviewed journal.	0.406	0.491
Method			
Matching	=1 if the estimate is obtained with the propensity score matching technique (baseline: OLS).	0.217	0.413
Before-after	=1 if the estimate is obtained with the before-after approach (baseline: OLS).	0.036	0.187
DID	=1 if the estimate is obtained with the difference-in-differences method (baseline: OLS).	0.146	0.353
GEE	=1 if the estimate is obtained with the generalized evaluation estimator (baseline: OLS).	0.051	0.221
IV	=1 if the estimate is obtained with the instrument variables method (baseline: OLS).	0.252	0.434
	Additional variables in Section 4:		
Instruments: proximity	=1 if the study uses instruments, and the instrument set includes a measure of political proximity of the country to the IMF (e.g. the country's quota in the IMF, UN voting patterns and membership features, country's staff at the IMF, etc.).	0.110	0.313
Instruments: cumulative time spent	=1 if the study uses instruments, and the instrument set includes a measure of cumulative time spent in IMF programs prior.	0.005	0.071
Time fixed effects	=1 if the estimate is obtained in a specification with time fixed effects.	0.152	0.359
Country fixed effects	=1 if the estimate is obtained in a specification with country fixed effects.	0.386	0.487
Horizon			
Medium run effect	=1 if the horizon on which the effect is measured is between 2 and 9 years from program start (baseline: short run effect, i.e. horizon shorter than 2 years after program start).	0.320	0.467

Continued on next page

Table A1: Definitions and summary statistics of explanatory variables (continued)

Variable	Description	Mean	Std. dev.
Long run effect	=1 if the horizon on which the effect is measured is 10 years and over from program start (baseline: short run effect, i.e. horizon shorter than 2 years after program start).	0.075	0.264
Program			
General res. only	=1 if the sample only includes IMF programs funded through general resources (baseline: the sample includes both programs with general and concessional funding).	0.217	0.413
Concessional res. only	=1 if the sample only includes IMF programs funded through concessional resources (baseline: the sample includes both programs with general and concessional funding).	0.146	0.353
Extra explanatory and dependent variables			
Exp. var.=Amount	=1 if the explanatory variable is the IMF loan amount (baseline: explanatory variable is dummy for IMF program presence).	0.210	0.408
Exp. var.=Compliance	=1 if the explanatory variable is the measure of countries' compliance with IMF conditionality (baseline: explanatory variable is dummy for IMF program presence).	0.013	0.114
Dep. var.=Growth change	=1 if the dependent variable is the change in GDP growth rate (baseline: dependent variable is the growth rate).	0.069	0.254
Dep. var.=Growth difference	=1 if the dependent variable is the difference between actual and projected GDP growth rate (baseline: dependent variable is the growth rate).	0.024	0.154
N of studies	36		
Observations	994		

Notes: Collected from 36 studies estimating the effect of IMF programs on economic growth.

Appendix B Model Uncertainty: Additional Results and Robustness

Appendix B.1 Additional Results and Robustness with BMA

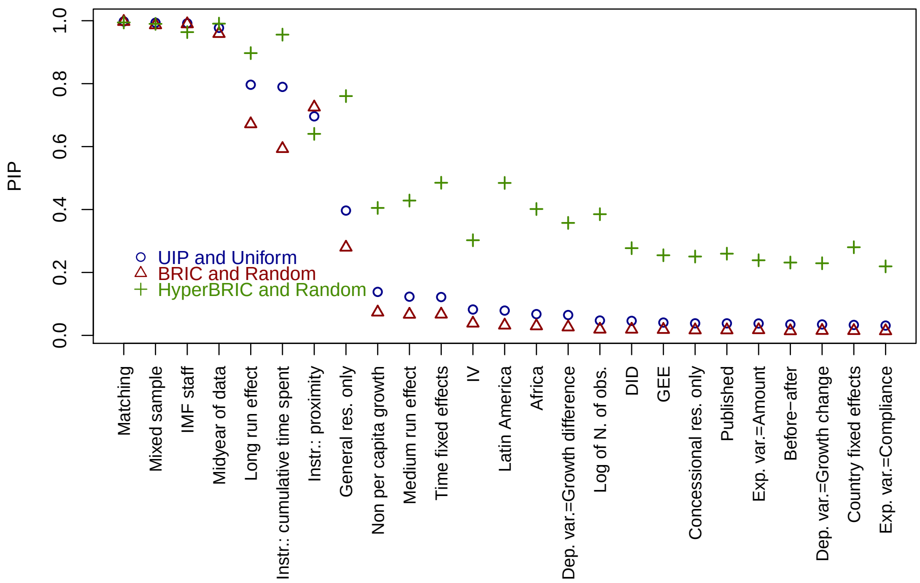
Table B1: Best Practice Estimates: sample mean for *Midyear of data*

Group	Point Estimate	95% interval	95% interval (wild)
All	0.358	[0.235; 0.482]	[0.196; 0.484]
Country groups			
Developing only	0.179	[-0.040; 0.398]	[-0.187; 0.445]
Mixed sample	0.739	[0.469; 1.009]	[0.472; 1.589]
Technique			
All: Matching	0.948	[0.766; 1.130]	[0.699; 1.200]
All: Non-matching	0.215	[0.053; 0.378]	[0.000; 0.398]
All: IV with proximity instruments	-0.318	[-1.194; 0.557]	[-2.813; 0.697]
Horizon			
Long Run effect only	0.923	[0.476; 1.369]	[-0.112; 2.177]
Short and Medium Run effect	0.312	[0.185; 0.440]	[0.134; 0.439]

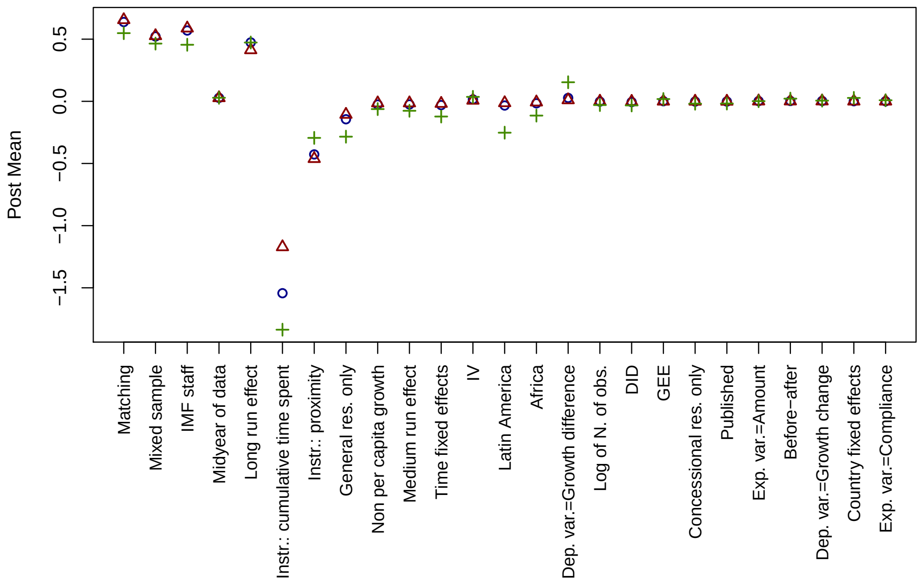
Notes: The table presents fitted values of estimates of the IMF effects on growth, for different sources of data, estimation techniques and measurement horizons. The estimates are obtained using the frequentist check model reported in Table 4. To obtain the fitted values, we substitute sample means for all variables, except for those related to the category under consideration. Here we use the sample mean for the variable *Midyear of data*, unlike in Table 5 where we employed the value of the 90s percentile. We report fitted values, 95% confidence intervals and confidence intervals constructed with wild bootstrap cluster.

Figure B1: Robustness checks: BMA with alternative priors

(a) Posterior inclusion probabilities under alternative priors



(b) Posterior means under alternative priors



Notes: 'UIP and Uniform' stands for unit information prior for parameters and the uniform model prior for model space; these are our chosen baseline priors that were used to obtain the results appearing in Figure 2 and Table 4. UIP is a prior that contains the amount of information similar to that of one observation; uniform prior for model space implicitly gives higher weight to average model size. These priors were reported to perform well in predictive estimations (see Eicher *et al.* 2011). 'BRIC and Random' are the benchmark g-prior for parameters (proposed by Fernandez *et al.* 2001) and the beta-binomial model prior for the model space (suggested by Ley & Steel 2009). HyperBRIC is a data-dependent hyper-g prior for model parameters (discussed in Feldkircher & Zeugner 2012).

Appendix B.2 Results with LASSO

In this section we address model uncertainty with LASSO instead of BMA. BMA directly models and quantifies uncertainty over the space of all possible combinations of explanatory variables, assigning each model the likelihood of representing the 'true' data generating process. LASSO, a model selection tool introduced by Tibshirani (1996), follows a conceptually different approach. Starting from the premiss that the 'true' underlying model is likely to be sparse, LASSO attempts to select one combination of explanatory variables which would fit the data best—therefore, unlike BMA, it does not base the coefficient estimates on averaged results coming from all possible models. LASSO obtains coefficient estimates by solving a modified OLS minimization problem that adds an upper bound on the sum of absolute values of the estimates. OLS, an unconstrained minimization procedure, does not assign exact zeros to coefficient values—and therefore fails at achieving sparsity. By adding a constraint on the sum of absolute coefficient values to the OLS minimization, LASSO transforms the original procedure into a problem that sometimes yields corner solutions and assigns exact zeros to some regression coefficients.

One important caveat of this technique is that a researcher implementing LASSO must choose a specific value for the upper bound on the sum of regression coefficients. This is usually done via cross-validation, an approach that we follow as well. Nevertheless, Belloni *et al.* (2012) note that the presence of a specific upper bound results in LASSO coefficient estimates shrinking toward zero. To correct for this bias, they suggest to additionally report the results from the post-LASSO estimation procedure which discards the variables that LASSO assigned the exact zeros, and performs an OLS on the remaining set of controls. We also follow this practice.

We implement LASSO with 10-fold cross-validation in `STATA` using the `cvlasso` routine; we choose the value for the upper bound on the coefficients to minimize the mean-squared prediction error. We report these results in Table B2; we also show the results from the post-LASSO procedure. Overall, the patterns emerging here are similar to the ones we have seen in Section 3 and Section 4. First, the variables discarded by the LASSO did not display statistical significance in any of the specifications studied prior. Second, comparing coefficient estimates produced by LASSO with those obtained in Section 3 and Section 4, we note that the signs of the coefficient estimates appear very consistent with evidence reported earlier. Third, we note that, overall, LASSO selects a larger number of explanatory variables.

We now turn to study what this evidence implies about the fitted values of the effects of the IMF programs on growth; we compute the point estimates conditional on different method and data choices and report the results in Table B3. Compared to Table 5, here we are able to present results for more subgroups—as LASSO chooses more explanatory variables compared to our baseline BMA.

Table B2: Why do estimates vary? LASSO

	LASSO	OLS using selected variables		
Response variable:	Coef.	Coef.	P-value	P-value (wild)
Publication and Data				
IMF staff	0.484	0.473	0.060	0.283
Mixed sample	0.254	0.357	0.177	0.267
Latin America	-0.245	-0.543	0.043	0.185
Africa	-0.112	-0.300	0.017	0.157
Midyear of data	0.024	0.028	0.000	0.007
Log of N. of obs.	0	.	.	.
Non per capita growth	0	.	.	.
Published	0	.	.	.
Method				
Matching	0.450	0.491	0.002	0.023
Before-after	0	.	.	.
DID	0	.	.	.
GEE	0	.	.	.
IV	0	.	.	.
Instruments: proximity	-0.026	-0.095	0.880	0.840
Instruments: cumulative time spent	-1.573	-2.119	0.041	0.235
Time fixed effects	-0.121	-0.209	0.354	0.392
Country fixed effects	0	.	.	.
Horizon				
Medium run effect	-0.130	-0.106	0.393	0.430
Long run effect	0.472	0.606	0.012	0.209
Program				
General res. only	-0.272	-0.375	0.131	0.155
Concessional res. only	0	.	.	.
Extra explanatory and dependent variables				
Exp. var.=Amount	0	.	.	.
Exp. var.=Compliance	0	.	.	.
Dep. var.=Growth change	0	.	.	.
Dep. var.=Growth difference	0.256	0.549	0.188	0.263
Const.	-0.192	-0.255	0.134	0.166
N of clusters	36	36		
Observations	994	994		

Notes: The left panel reports the results of estimating LASSO using cross-validation with 10 folds where the bound is chosen to minimize the mean-squared prediction error. We implement this in `STATA` using the `cvlasso` command. The right panel presents results from OLS estimation using variables selected by LASSO. We report regular p-values as well as p-values from wild bootstrap cluster.

Table B3: Fitted Estimates by Group

Group	Point Estimate	95% interval	95% interval (wild)
All	0.532	[0.392; 0.672]	[0.275; 0.794]
Country groups			
Developing only	0.418	[0.177; 0.658]	[-0.141; 0.956]
Latin America	-0.086	[-0.526; 0.354]	[-1.533; 23.544]
Africa	0.157	[0.016; 0.298]	[-2.042; 1.450]
Mixed sample	0.814	[0.477; 1.151]	[0.507; 1.830]
Technique			
All: Matching	0.937	[0.662; 1.213]	[0.542; 1.430]
All: Non-matching	0.425	[0.250; 0.601]	[0.108; 0.778]
All: IV with proximity instruments	0.341	[-0.846; 1.528]	[-1.119; 2.115]
Horizon			
Long Run effect only	1.126	[0.650; 1.602]	[-0.149; 2.642]
Medium Run effect only	0.414	[0.228; 0.600]	[0.197; 0.632]
Short Run effect only	0.520	[0.337; 0.704]	[0.192; 0.846]
Program			
General resource only	0.238	[-0.139; 0.615]	[-0.250; 0.673]
Concessional and mixed resource	0.614	[0.421; 0.807]	[0.315; 0.939]

Notes: The table presents fitted values of estimates of the IMF effects on growth, for different sources of data, estimation techniques and measurement horizons. The estimates are obtained using the post-LASSO estimation results reported in Table B2. To obtain the fitted values, we substitute sample means for all variables, except for those related to the category under consideration and the *Midyear of data*: we use the value of the 90s percentile for the latter. We report fitted values, 95% confidence intervals and confidence intervals constructed with wild bootstrap cluster.

Appendix C Studies Used in Meta-analysis

We used the following search query to find the relevant studies:

> Our search querry is: 'IMF programs' or 'IMF bailouts', and 'economic growth'.

Papers in Study

Atoyan, R. & P. Conway (2006): "Evaluating the impact of IMF programs: A comparison of matching and instrumental-variable estimators." *The Review of International Organizations* **1(2)**: pp. 99–124.

Baqir, R., R. Ramcharan, & R. Sahay (2003): "The consistency of imf programs." In "Macroeconomic challenges in low income countries, research workshop, October," pp. 23–24.

Baqir, R., R. Ramcharan, & R. Sahay (2005): "Imf programs and growth: is optimism defensible?" *IMF Staff Papers* **52(2)**: pp. 260–286.

Barro, R. J. & J.-W. Lee (2005): "IMF programs: Who is chosen and what are the effects?" *Journal of Monetary Economics* **52(7)**: pp. 1245–1269.

Bas, M. A. & R. W. Stone (2014): "Adverse selection and growth under imf programs." *The Review of International Organizations* **9(1)**: pp. 1–28.

Binder, M. & M. Bluhm (2017): "On the conditional effects of imf program participation on output growth." *Journal of Macroeconomics* **51**: pp. 192–214.

Bird, G. & D. Rowlands (2017): "The effect of imf programmes on economic growth in low income countries: An empirical analysis." *The Journal of Development Studies* **53(12)**: pp. 2179–2196.

Bordo, M. D. & A. J. Schwartz (2000): "Measuring real economic effects of bailouts: historical perspectives on how countries in financial distress have fared with and without bailouts." In "Carnegie-Rochester Conference Series on Public Policy," volume 53, pp. 81–167. Elsevier.

Butkiewicz, J. L. & H. Yanikkaya (2005): "The effects of imf and world bank lending on long-run economic growth: An empirical analysis." *World Development* **33(3)**: pp. 371–391.

Conway, P. (1994): "Imf lending programs: Participation and impact." *Journal of Development Economics* **45(2)**: pp. 365–391.

Dicks-Mireaux, L., M. Mecagni, & S. Schadler (2000): "Evaluating the effect of imf lending to low-income countries." *Journal of Development Economics* **61(2)**: pp. 495–526.

Doroodian, K. (1993): "Macroeconomic performance and adjustment under policies commonly supported by the international monetary fund." *Economic Development and Cultural Change* **41(4)**: pp. 849–864.

Dreher, A. (2006): "IMF and economic growth: The effects of programs, loans, and compliance with conditionality." *World Development* **34(5)**: pp. 769–788.

Eichengreen, B., P. Gupta, & A. Mody (2008): "Sudden Stops and IMF-Supported Programs." In "Financial Markets Volatility and Performance in Emerging Markets," NBER Chapters, pp. 219–266. National Bureau of Economic Research, Inc.

Fidrmuc, J. & S. Kostagianni (2015): "Impact of imf assistance on economic growth revisited 1." *Economics & Sociology* **8(3)**: p. 32.

Goldstein, M. & P. Montiel (1986): "Evaluating fund stabilization programs with multicountry data: Some methodological pitfalls." *IMF Staff Papers* **33(2)**: pp. 304–344.

Gündüz, Y. B. (2016): "The economic impact of short-term imf engagement in low-income countries." *World Development* **87**: pp. 30–49.

Gylfason, T. (1987): *Credit policy and economic activity in developing countries with IMF stabilization programs.* International Finance Section, Department of Economics, Princeton University.

Hardoy, I. (2003): "Effect of IMF programs on growth: A reappraisal using the method of matching." *Institute for Social Research Paper.*

Hutchison, M. (2003): "A Cure Worse Than the Disease? Currency Crises and the Output Costs of IMF-Supported Stabilization Programs." In "Managing Currency Crises in Emerging Markets," NBER Chapters, pp. 321–360. National Bureau of Economic Research, Inc.

Hutchison, M. M. (2004): "Selection bias and the output costs of imf programs." , EPRU Working Paper Series.

Hutchison, M. M. & I. Noy (2003): "Macroeconomic effects of IMF-sponsored programs in Latin America: output costs, program recidivism and the vicious cycle of failed stabilizations." *Journal of International Money and Finance* **22(7)**: pp. 991–1014.

IEO (2002): "Evaluation of prolonged use of imf resources." , Independent Evaluation Office of the International Monetary Fund, DC.

IMF (2012): "Review of facilities for low-income countries." , International Monetary Fund.

Khan, M. S. (1990): "The macroeconomic effects of fund-supported adjustment programs." *IMF Staff Papers* **37(2)**: pp. 195–231.

Kilman, J. & K. Olsson (2013): "The imf and economic growth: An analysis of lending to developing countries during 1983-2010." *Thesis, Lund University* .

Marchesi, S. & E. Sirtori (2011): "Is two better than one? the effects of imf and world bank interaction on growth." *The Review of International Organizations* **6(3-4)**: pp. 287–306.

Mercer-Blackman, V. & A. Unigovskaya (2004): "Compliance with imf program indicators and growth in transition economies." *Emerging Markets Finance and Trade* **40(3)**: pp. 55–83.

Mumssen, C., Y. Bal-Gunduz, C. H. Ebeke, & L. Kaltani (2013): "IMF-Supported Programs in Low Income Countries; Economic Impact over the Short and Longer Term." *IMF Working Papers 13/273*, International Monetary Fund.

Newiak, M. & T. Willems (2017): "Evaluating the Impact of Non-Financial IMF Programs Using the Synthetic Control Method." *IMF Working Papers 17/109*, International Monetary Fund.

Nsouli, S. M., R. V. Atoyan, & A. Mourmouras (2004): "Institutions, Program Implementation, and Macroeconomic Performance." *IMF Working Papers 04/184*, International Monetary Fund.

Ozturk, I. (2008): "Evaluating the macroeconomic impacts of imf programmes in latin america, 1975-2004: A gee analysis." *South African Journal of Economic and Management Sciences* **11(2)**: pp. 190–202.

Ozturk, I. (2011): "The macroeconomic effects of imf programs in mena countries." *African Journal of Business Management* **5(11)**: pp. 4379–4387.

Quaglia, R. (2016): "Evaluating the impact of imf loan participation on real gdp growth, gross capital formation growth, and unemployment, using a difference-in-differences estimation." *Lake Forest College Publications* .

Ramos, R. (2008): *Do IMF programmes stabilize the economy?* Thesis, CEMFI.

Yang, G. (2013): "Evaluate the effect of imf's longer-term concessional lending programs on growth in the development background of subsaharan region." *Undergraduate Economic Review* **10(1)**: p. 2.

Appendix D Additional Results

Appendices D is only presented here for the convenience of reviewers. If the manuscript is accepted for publication, this material will be relegated to an online appendix.

Appendix D.1 Publication Bias

In this section we will investigate whether the literature estimating the effects of IMF programs on economic growth is prone to selective reporting of the results—in other words, whether there is publication bias in this literature. The problem of selective reporting has been found to be prominent in some fields in economics. For example, Card & Krueger (1995) and Doucouliagos & Stanley (2009) find evidence of publication bias in the literature studying the effect of minimum wage regulations on employment; Balima *et al.* (2017) uncover publication bias in the literature estimating the macroeconomic effect of inflation targeting adoption; Havranek & Sokolova (2020) document strong publication bias for studies estimating shares of rule-of-thumb consumers with micro-level data due to the underreporting of negative estimates.

The intuition behind testing for publication bias can be described as follows. In the absence of selective reporting, estimates appearing in the empirical literature should be distributed symmetrically around the 'true' underlying parameter—due to noise in the data.[12] A skewed distribution of estimates could indicated that certain results tend to be discarded, providing evidence of publication bias. We apply these ideas here. One important consideration is that for this line of argument to work, we would need to focus on estimates that pertain to one underlying 'true' effect, rather than several different parameters. In other words, we would need to identify a subset of estimates that are relatively homogenous.

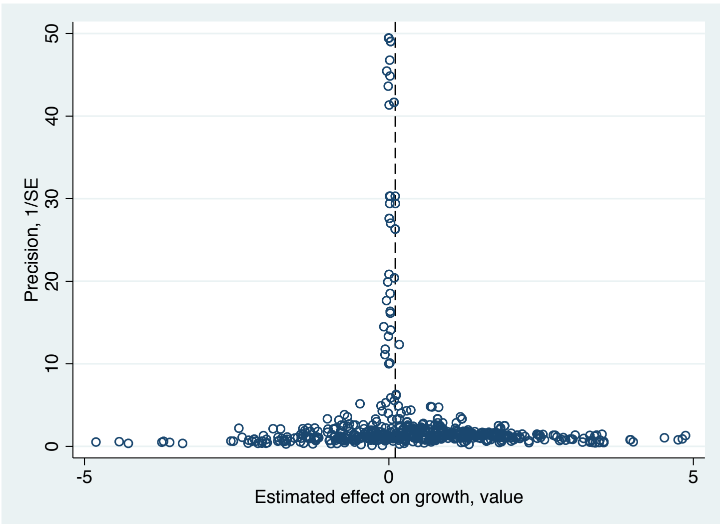
As we discuss in Section 2, researchers consider different channels through which the IMF involvement affects growth: while the majority of the literature focuses on the effect of the IMF presence (and uses dummy variables to capture whether there is an active IMF program), there are also studies that consider how growth is affected through loan amounts and countries' compliance with conditionality. Although these effects may end up sharing similar signs and being affected by features of study design in similar directions, there is no theoretical reason for these effects to be of similar magnitudes. Similarly, while the majority of estimates is obtained with growth rates as dependent variables, a small subset measures the effect of IMF programs on *changes* in growth rates and differences between actual and projected growth. This, again, could mean that these studies are estimating different underlying parameters.

[12] This logic only applies to cases of unconstrained estimation procedures.

To ensure that the sample we consider is relatively homogenous, here we focus on a subset of estimates that come from the specification in which the dependent variable is a growth rate, and the estimate measures how growth is affected by IMF presence (as opposed to loan amount or compliance with conditionality). This strategy leaves us with 684 estimates which comprise about 70% of the full dataset. We also consider this subsample in Section 3 (see Table 2).

Figure D1 presents a funnel plot depicting estimates against their precision. In the absence of selective reporting, the graph should assume the shape of a symmetrical inverted funnel: estimates should be distributed symmetrically, with more precise estimates clustering closer to the underlying 'true' effect (see Egger *et al.* 1997). The funnel plot shown on Figure D1 appears to have these properties: even though the tails are not exactly symmetrical, we cannot state with confidence that one tail appears much more prominent than the other. We therefore conclude that the graphical test does not appear to suggest strong bias.

Figure D1: Funnel plot of estimates of β^{IMF}



Notes: The figure plots estimates of the effect of IMF presence on growth rates in patricipating countries against their precision, $1/SE(\hat{\beta}_{ij}^{\text{IMF}})$.

To investigate this further, we proceed with a formal funnel asymmetry test. In the absence of publication bias, there should be no correlation between estimates and their standard errors, as their ratios would roughly follow a t-distribution. If estimates of the 'wrong' sign get discarded, or if researchers tend to underreport results that are not statistically significant, we would likely observe some correlation between estimates and the standard errors (see detailed discussion in Stanley 2005). We now apply these ideas to our sample of 684 estimates and consider a regression model:

$$\hat{\beta}_{ij}^{\text{IMF}} = \beta_0 + \lambda \cdot SE(\hat{\beta}_{ij}^{\text{IMF}}) + u_{ij}, \tag{3}$$

where $\hat{\beta}_{ij}^{\text{IMF}}$ is the i-th estimate of the effect of IMF presence on growth rates reported by study j, $SE(\hat{\beta}_{ij}^{\text{IMF}})$ is the corresponding standard error and u_{ij} is the disturbance term. In the absence of publication bias, the coefficient λ should be zero, i.e. there should be no correlation between the estimates and their standard errors. The estimates are likely to be correlated within studies; we therefore cluster standard errors at the study level and additionally compute wild bootstrapped clustered p-values, to address the concern that the number of clusters we are using is relatively small.

We report the results of estimating (3) in Table D1 for five alternative specifications. The first column documents the results of estimating (3) with a simple OLS; we then estimate a version of the model that includes study-level fixed effects and report the results in the second column. In the third column we report the results for a specification in which we regress median estimates reported by each study on median standard errors, thereby only employing variation across studies to evaluate λ. In column four we report the results for a specification in which we use estimates' precision as a weight to address potential heteroskedasticity (see discussion in Stanley & Doucouliagos 2015). Finally, as some studies report many more estimates than others and are therefore overrepresented in our sample, we attempt to equalize each studies' weight by weghting each data point by the inverse of the number of estimates reported in the corresponding study. For a more detailed discussion of the above techniques see, e.g. Sokolova & Sorensen (2018).

The results reported in Table D1 are somewhat conflicting. On the one hand, the majority of specifications produce negative estimates of λ, and the standard p-values indicate some statistical significance. On the other hand, the p-values computed using wild bootstrap clustering are, for the most part, far above 10%. Furthermore, the estimate of λ is positive in the specification employing precision weights.

To test the sensitivity of these results to the treatment of outliers, we winsorize the outliers at 1% (0.05% each tail); we report the results in Table D2. Despite the very mild change to the overall dataset, the evidence presented in Table D2 markedly differs from results obtained for the untreated sample, showing much less overall evidence for the statistical significance of λ. This implies that results shown in Table D1 that appeared to be somewhat closer to the 10% statistical significance mark for estimates of λ are likely driven by the outliers rather than the systematic underlying effects.[13]

We conclude that there does not seem to be any consistent evidence in support of

[13] The same is not true of our main results reported in Section 3: the core evidence presented in Table 2 remains intact even after we apply a stronger outlier treatment, winsorizing 5% of the outliers.

Table D1: Testing for publication bias

Panel A: All estimates					
	OLS	FE	BE	Precision	Study
SE dummy	-0.307	-0.354	-0.793	0.339	-0.318
	(0.012)	(0.001)	(0.032)	(0.093)	(0.120)
	[0.312]	[0.329]	[0.185]	[0.120]	[0.076]
Constant	0.605	0.643	0.601	0.073	0.315
	(0.002)	(0.000)	(0.066)	(0.000)	(0.102)
	[0.005]	.	[0.014]	[0.131]	[0.118]
Studies	27	27	27	27	27
Observations	684	684	27	684	684
Panel B: Published estimates only					
	OLS	FE	BE	Precision	Study
SE dummy	-0.448	-0.425	-1.109	0.234	-0.403
	(0.000)	(0.000)	(0.062)	(0.559)	(0.114)
	[0.198]	[0.197]	[0.211]	[0.627]	[0.016]
Constant	0.829	0.807	0.922	0.168	0.357
	(0.004)	(0.000)	(0.125)	(0.209)	(0.309)
	[0.092]	.	[0.032]	[0.118]	[0.345]
Studies	13	13	13	13	13
Observations	201	201	13	201	201

Notes: The table presents estimates of regression (3). We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and compute p-values using `STATA` command `boottest`, with Rademacher weights and 9999 replications (see Roodman 2018); we report these p-values in square brackets. To produce results for this table, we only use a (relatively homogenous) sample of estimates that pertain to the effect of IMF *presence* on the growth rate (i.e. estimates obtained in a specification with dummy for IMF programs as explanatory variable and GDP growth rate as dependent variable). (OLS)=Ordinary least squares; (FE)=Fixed effect estimation; (BE)=OLS using median estimates and standard errors reported in each study; (Precision)=a specification with precision weights (i.e. $1/SE(\hat{\beta}_{ij}^{\text{IMF}})$); (Study)=a specification with the inverse number of estimates reported in each study as weights.

selective reporting, as results appear to be conflicting across specifications with respect to the sign of the effect, and also are likely driven by a small set of outliers. We leave further investigations to future researchers and proceed to look for other sources of variation in estimates.

Table D2: Testing for publication bias (outliers winsorized at 1%)

	OLS	FE	BE	Precision	Study
Panel A: All estimates					
SE dummy	-0.301	-0.354	-0.793	0.369	-0.519
	(0.295)	(0.269)	(0.032)	(0.099)	(0.124)
	[0.387]	[0.335]	[0.185]	[0.124]	[0.198]
Constant	0.583	0.625	0.601	0.072	0.460
	(0.016)	(0.017)	(0.066)	(0.000)	(0.042)
	[0.005]	.	[0.014]	[0.105]	[0.041]
Studies	27	27	27	27	27
Observations	684	684	27	684	684
Panel B: Published estimates only					
	OLS	FE	BE	Precision	Study
SE dummy	-0.707	-0.660	-1.109	0.321	-0.430
	(0.123)	(0.137)	(0.062)	(0.515)	(0.216)
	[0.177]	[0.285]	[0.211]	[0.589]	[0.014]
Constant	1.023	0.981	0.922	0.143	0.369
	(0.013)	(0.021)	(0.125)	(0.248)	(0.333)
	[0.064]	.	[0.032]	[0.145]	[0.368]
Studies	13	13	13	13	13
Observations	201	201	13	201	201

Notes: The table presents estimates of regression (3). We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and compute p-values using `STATA` command `boottest`, with Rademacher weights and 9999 replications (see Roodman 2018); we report these p-values in square brackets. To produce results for this table, we only use a (relatively homogenous) sample of estimates that pertain to the effect of IMF *presence* on the growth rate (i.e. estimates obtained in a specification with dummy for IMF programs as explanatory variable and GDP growth rate as dependent variable). Furthermore, here we winsorize the outliers in each tail at 0.5%. (OLS)=Ordinary least squares; (FE)=Fixed effect estimation; (BE)=OLS using median estimates and standard errors reported in each study; (Precision)=a specification with precision weights (i.e. $1/SE(\hat{\beta}_{ij}^{\text{IMF}})$); (Study)=a specification with the inverse number of estimates reported in each study as weights.

Appendix D.2 Full Tables 2 and 3

Table D3: Heterogeneity: horizon and program definitions

Variable	(1)	(2)	(3)	(4)	(5)
Publication and Data					
IMF staff	0.509	0.629	0.519	0.474	-0.016
	(0.074)	(0.021)	(0.085)	(0.110)	(0.891)
	[0.348]	[0.130]	[0.380]	[0.429]	[0.855]
Mixed sample	0.626	0.554	0.541	0.624	0.230
	(0.094)	(0.164)	(0.112)	(0.090)	(0.023)
	[0.171]	[0.280]	[0.164]	[0.152]	[0.031]
Method					
Matching	0.544	0.569	0.594	0.621	0.737
	(0.002)	(0.001)	(0.000)	(0.002)	(0.000)
	[0.014]	[0.011]	[0.005]	[0.020]	[0.011]
Before-after	0.000	0.013	0.036	0.030	-0.066
	(0.999)	(0.980)	(0.953)	(0.958)	(0.852)
	[0.998]	[0.969]	[0.912]	[0.923]	[0.792]
DID	-0.146	-0.124	0.028	-0.085	0.087
	(0.656)	(0.701)	(0.917)	(0.790)	(0.598)
	[0.712]	[0.725]	[0.915]	[0.818]	[0.660]
GEE	0.364	0.429	-0.294	0.233	-0.291
	(0.492)	(0.430)	(0.353)	(0.635)	(0.152)
	[0.593]	[0.557]	[0.474]	[0.720]	[0.228]
IV	0.129	0.082	0.020	0.099	-0.060
	(0.736)	(0.835)	(0.960)	(0.796)	(0.000)
	[0.748]	[0.832]	[0.963]	[0.805]	[0.615]
Horizon					
Medium run effect	-0.348	.	-0.419	-0.360	-0.095
(between 2 and 9 years)	(0.139)	.	(0.146)	(0.126)	(0.000)
	[0.218]	.	[0.198]	[0.203]	[0.343]
Long run effect	0.907	.	0.972	1.026	0.956
(10 years and over)	(0.018)	.	(0.013)	(0.026)	(0.001)
	[0.297]	.	[0.244]	[0.215]	[0.202]
Medium run effect	.	-0.349	.	.	.
(between 2 and 5 years)	.	(0.259)	.	.	.
	.	[0.377]	.	.	.
Long run effect	.	0.517	.	.	.
(6 years and over)	.	(0.137)	.	.	.
	.	[0.404]	.	.	.
Program					
General res. only	-0.837	-0.951	.	.	-0.104
	(0.082)	(0.050)	.	.	(0.000
	[0.125]	[0.086]	.	.	[0.523]
Concessional res. only	0.252	0.162	.	.	-0.076
	(0.194)	(0.428)	.	.	(0.000)
	[0.338]	[0.555]	.	.	[0.613]
Short run prog. only	.	.	-0.132	.	.
(official definition)	.	.	(0.512)	.	.
	.	.	[0.511]	.	.
Long run prog. only	.	.	0.357	.	.
(official definition)	.	.	(0.128)	.	.
	.	.	[0.294]	.	.
Short run prog. only	.	.	.	-0.640	.
(authors' definition)	.	.	.	(0.156)	.
	.	.	.	[0.261]	.
Long run prog. only	.	.	.	0.250	.
(authors' definition)	.	.	.	(0.241)	.
	.	.	.	[0.367]	.
Const.	0.058	0.085	-0.042	0.002	0.130
	(0.788)	(0.703)	(0.845)	(0.993)	(0.000)

Continued on next page

Table D3: Heterogeneity: horizon and program definitions

Variable	(1)	(2)	(3)	(4)	(5)
	[0.833]	[0.781]	[0.875]	[0.994]	[0.039]
N of clusters	27	27	27	27	27
Observations	684	684	684	684	684

Notes: Here we present the full version of Table 2. We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and compute p-values using `STATA` command `boottest`, with Rademacher weights and 9999 replications (see Roodman (2018)); we report these p-values in square brackets. To produce results for this table, we only use a (relatively homogenous) sample of estimates that pertain to the effect of IMF *presence* on the economic growth (i.e. estimates obtained in a specification with dummy for IMF programs as explanatory variable). (1)=baseline specification; (2)=a specification with an alternative definition of the long run (6 years and over); (3)=a specification with alternative controls for program type, distinguishing between long and short run programs based on the official program definition; (4)=a specification with alternative controls for program type, distinguishing between long and short run programs based on the authors' program definition; (5)=a specification in which all data is weighted by precision, $1/SE(\beta^{\text{IMF}})$.

Table D4: Heterogeneity: sample splits

Variable	(1)	(2)	(3)	(4)
Publication and Data				
IMF staff	0.509	0.281	0.476	0.444
	(0.074)	(0.358)	(0.068)	(0.133)
	[0.348]	[0.600]	[0.282]	[0.370]
Mixed sample	0.626	0.504	.	0.522
	(0.094)	(0.059)	.	(0.058)
	[0.171]	[0.111]	.	[0.112]
Method				
Matching	0.544	0.639	0.493	0.636
	(0.002)	(0.001)	(0.006)	(0.001)
	[0.014]	[0.013]	[0.050]	[0.014]
Before-after	-0.000	0.011	-0.033	0.004
	(0.999)	(0.983)	(0.953)	(0.995)
	[0.998]	[0.973]	[0.952]	[0.992]
DID	-0.146	-0.070	-0.079	-0.087
	(0.656)	(0.790)	(0.713)	(0.746)
	[0.712]	[0.812]	[0.798]	[0.774]
GEE	0.364	0.125	0.200	0.089
	(0.492)	(0.751)	(0.642)	(0.822)
	[0.593]	[0.809]	[0.732]	[0.860]
IV	0.129	0.099	-0.184	0.004
	(0.736)	(0.696)	(0.589)	(0.987)
	[0.748]	[0.715]	[0.673]	[0.989]
Horizon				
Medium run effect	-0.348	-0.288	-0.491	-0.305
(between 2 and 9 years)	(0.139)	(0.178)	(0.084)	(0.161)
	[0.218]	[0.242]	[0.144]	[0.231]
Long run effect	0.907	0.985	0.767	0.700
(10 years and over)	(0.018)	(0.019)	(0.054)	(0.012)
	[0.297]	[0.217]	[0.440]	[0.256]
Program				
General res. only	-0.837	-0.554	-0.836	-0.535
	(0.082)	(0.044)	(0.031)	(0.060)
	[0.125]	[0.087]	[0.182]	[0.116]
Concessional res. only	0.252	0.124	0.129	0.136
	(0.194)	(0.388)	(0.410)	(0.351)
	[0.338]	[0.478]	[0.541]	[0.440]
Extra explanatory and dependent variables				
Exp. var.=Amount	.	0.025	0.139	0.064
	.	(0.943)	(0.718)	(0.853)
	.	[0.948]	[0.660]	[0.872]
Exp. var.=Compliance	.	.	.	0.062
	.	.	.	(0.917)
	.	.	.	[0.987]
Dep. var.=Growth change	.	.	.	0.058
	.	.	.	(0.844)
	.	.	.	[0.866]
Dep. var.=Growth difference	.	.	.	0.742
	.	.	.	(0.050)
	.	.	.	[0.166]

Continued on next page

Table D4: Heterogeneity: sample splits

Variable	(1)	(2)	(3)	(4)
Const.	0.058	0.067	0.227	0.069
	(0.788)	(0.674)	(0.144)	(0.679)
	[0.833]	[0.744]	[0.435]	[0.750]
N of clusters	27	31	23	36
Observations	684	893	600	994

Notes: Here we present the full version of Table 3. We report regular p-values in parenthesis. In addition, we perform wild bootstrap clustering and report the associated p-values in square brackets. (1)=baseline results—same as column (1) of Table 2—obtained using the sample of estimates that pertain to the effect of IMF *presence* on the economic growth (i.e. with dummy for IMF programs as explanatory variable); (2)=the sample of estimates from column (1) together with estimates measuring the effect of loan *amount* on growth (i.e. with loan amount as explanatory variable); (3)=the subsample of estimates from column (2) measuring effects of IMF *presence* and loan *amount*—for developing countries only; (4)=the full sample of estimates collected, that extends sample of column (2) by adding estimates obtained using compliance with conditionality as explanatory variable, as well as estimates corresponding to alternative dependent variables: growth change over time and difference between actual and projected growth.