Supplemental Materials for: Have Your Cake and Eat it Too? Cointegration and Dynamic Inference from Autoregressive Distributed Lag Models

Andrew Q. Philips

March 27, 2017

Contents

1 Programs to Assist in Implementing the Pesaran, Shin and Smith (2001) ARDL Procedure **4**

1.1 pssbounds 4

1.2 dynpss 6

1.3 pss 12

2 Summary of Monte Carlo Results **13**

3 Additional Monte Carlo Results **16**

3.1 Lags and Overfitting in the Monte Carlo Analysis 17

3.2 Type II Error of the Cointegration Tests: Varying Adjustment Parameters and Long-Run Multipliers 20

3.3 Discordant Cointegration Tests 23

3.3.1 Type I Error, Correct Discordant 25

3.3.2 Type I Error, Incorrect Discordant 25

3.3.3 Type II Error, Correct Discordant 28

3.3.4 Type II Error, Incorrect Discordant 29

3.3.5 Suggestions for Practitioners for Cointegration Testing 32

3.4 Fractional Integration and the ARDL Procedure 33

3.4.1 No Fractional Integration, Finite vs. Infinite Variance 36

3.4.2 No Fractional Integration, Finite vs. Infinite Variance 37

3.4.3 Suggestions for Practitioners 40

3.5 How Well Can the ARDL Procedure Recover Cointegrating Effects? . 41

3.6 Can the ARDL-Bounds Procedure Avoid Spurious Cointegrating Effects? 58

3.7 How Well Can the ARDL Procedure Recover Stationary Relationships? 70

3.8 Can the ARDL Procedure Avoid Spurious Stationary Relationships? . 81

3.9 A More Conservative Assessment of Type I Error for the Cointegration Tests 90

4 Proof of the Equivalence of the Triangular Error-Correction Representation to the Standard Representation 95

5 Three Replications 97

5.1 Replication I: Kelly and Enns (2010) 97

5.1.1 Different Conclusions About the Time Series Properties of Welfare Policy Mood 100

5.2 Replication II: Volscho and Kelly (2012) 104

5.3 Replication III: Ura (2014) 107

5.4 A Comparison of the Replications to the Replications of Grant and Lebo (2016) 112

1 Programs to Assist in Implementing the Pesaran, Shin and Smith (2001) ARDL Procedure

I have designed a number of programs to aid in implementing the ARDL procedure proposed in the main paper. Current working versions are available for download on GitHub.[1] The Stata command `pssbounds` aids users by providing the critical values necessary for conducting the bounds test for cointegration (Philips 2016*b*). The Stata command `dynpss` is designed to dynamically simulate ARDL models over time, allowing the user to plot predicted values and their response to changes in the regressors (Philips 2016*a*). Examples are shown on the program websites. In addition, this approach has been adapted for `R` under the package `pss` (Jordan and Philips 2016), which is also available on GitHub.[2] As of December 2016, only the `pssbounds` function is fully supported by `pss`; the `dynpss` function is still in the development and testing stage. I discuss these programs below.

1.1 pssbounds

The Stata command `pssbounds` is designed to provide users with the critical values necessary to conduct the autoregressive distributed lag (ARDL) bounds testing procedure recommended by Pesaran, Shin and Smith (2001). The most current ver-

[1] Users can download `dynpss` at the following link: http://andyphilips.github.io/dynpss, and `pssbounds` at: http://andyphilips.github.io/pssbounds. Download the zip file, unzip and place the program (.ado) and the help file (.hlp) in your "ado/plus/" folder. To find out where this is, type `sysdir` in Stata.

[2] Later versions will be released on CRAN. The current version can be found here, along with instructions about how to download from GitHub to `R`: https://github.com/andyphilips/pss.

sion can be found on GitHub: http://andyphilips.github.io/pssbounds. For help on installing user-written .ado files in Stata, see Footnote 1.

As outlined in the main paper, the user must first run an ARDL model in error correction form. Then, after introducing lagged first differences of the series as necessary in order to ensure white-noise residuals, the user can conduct the bounds F- and t-tests for cointegration. While Pesaran, Shin and Smith (2001) provide asymptotic critical values for the F- and t-statistics, and Narayan (2005) provides finite-sample F-statistics, `pssbounds` provides the user with these critical values in Stata without having to look them up. The command appears as the following:

```
pssbounds, observations(#) fstat(#) case(#) k(#) [tstat(#)]
```

Required options are:

- `observations(#)` is the number of observations (the length of the series) from the ARDL-bounds model. Small-sample critical values of the bounds test depend on the size of the sample; therefore, this option is required.

- `fstat(#)` is the value of the F-statistic from the test that all variables appearing in levels are jointly equal to zero. This can be obtained using Stata's `test` command.[3] This option is required.

- `case(#)` identifies the type of case and thus which restrictions (if any) to impose on the intercept and/or trend term. Case type can be written in Roman numerals (I, II, III, IV, and V) or numerically (1, 2, 3, 4, and 5). Since the

[3] e.g., `test l.y l.x1 l.x2`.

critical values of the bounds test depend on the restrictions placed on the intercept and trend, this option is required. *By far the most common is an unrestricted intercept with no trend term:* `case(3)`. Other types that are supported are:

  - Case I: No intercept and no trend, `case(1)`.
  - Case II: Restricted intercept and no trend, `case(2)`.
  - Case IV: Unrestricted intercept and restricted trend, `case(4)`.
  - Case V: Unrestricted intercept and unrestricted trend, `case(5)`.

- `k(#)` is the number of regressors appearing in levels in the ARDL-bounds model. Since the critical values of the bounds test depend on the number of regressors, this option is required.

Additional options are:

- `tstat(#)` is the value of the one-sided t-test that the coefficient on the lagged dependent variable is equal to zero. Only asymptotic critical values from the bounds test are available for this test, and only for cases 1, 3, and 5.

1.2 dynpss

`dynpss` is a Stata command to dynamically simulate autoregressive distributed lag (ARDL) models like the ones discussed in the main paper. The most current version can be found on GitHub: http://andyphilips.github.io/dynpss. For help on installing user-written .ado files in Stata, see Footnote 1.

Before using `dynpss`, it is assumed the user has already determined the order of integration of the dependent variable, ensured no regressor is of an order of integration higher than I(1) (or contains other issues such as seasonal unit roots), used information criteria (and theory) to identify the best fitting lagged-difference structure, which is used to purge autocorrelation and to ensure the residuals are white noise, and also to have performed the bounds test and determined if there is cointegration (and if there is not, adjusted the model accordingly).

`dynpss` is designed to dynamically simulate the effects of a counterfactual change in one weakly exogenous regressor at a single point in time, holding all else equal, using stochastic simulation techniques. This approach is gaining in popularity as a simple way to show the substantive results of time series models (Williams and Whitten 2011; Philips, Rutherford and Whitten 2016*a*; Gandrud, Williams and Whitten 2016; Philips, Rutherford and Whitten 2016*b*). Since the ARDL model discussed in this paper can produce models that are somewhat complicated to interpret, `dynpss` is designed to ease this burden through the creation of predicted (or expected) values of the dependent variable (along with associated confidence intervals), which can then be plotted to show how a change in one variable "flows" through the model over time.

`dynpss` first runs a linear regression. Then, using a self-contained procedure similar to the popular Clarify program (Tomz, Wittenberg and King 2003), it takes 1000 (or however many simulations a user desires) draws of the vector of parameters from a multivariate normal distribution. These distributions are assumed to have means equal to the estimated parameters from the regression, and a variance equal

to the estimated variance-covariance matrix from the regression. In order to re-introduce stochastic uncertainty back into the model when creating predicted values, `dynpss` simulates σ^2 by taking draws from a scaled inverse χ^2 distribution. The distribution is scaled by the residual degrees of freedom (n-k), as well as the estimated $\hat{\sigma}^2$ from the regression (Gelman et al. 2014, pp. 43,581). This ensures that draws of σ^2 are bounded by zero and one. Simulated parameters and sigma-squared values are then used to create predicted $\hat{Y}_t$ values over time by setting all covariates to certain values (typically means), and introducing stochastic uncertainty back into the prediction by taking a draw from a multivariate normal distribution with mean zero and variance $\hat{\sigma}^2$. The program then obtains the average $\hat{Y}_t$ and percentile confidence intervals of the distribution of predicted values at a particular point in time. These are then saved, allowing a user to make a table or (more commonly) a graph of the results.

The command appears as the following:

```
dynpss depvar indepvars [, options]
```

the options below are required:

- `lags(numlist)` is a numeric list of the number of lags to include for each variable. The number of desired lags is listed in the order in which the variables `depvar` and `indepvars` appear. For instance, in a model with two weakly exogenous variables, we lag all variables by specifying: `lags(1 1 1)`. Note that the lag on `depvar` (the first "`1`") must always be specified. To estimate a model without a lag for a particular variable, simply replace the number with a

"."; for instance, if we did not want a lag on the first regressor, we type: `lags(1 . 1)`. If a higher number of lags are specified, `dynpss` will add consecutive lags. For instance, `lags(3 . .)` will introduce lags of y_t at $t-1$, $t-2$, and $t-3$ into the model.

- `shockvar(varname)` is a single independent variable from the list of `indepvars` that is to be shocked. It will experience a counterfactual shock of size `shockval(#)` at time `time(#)`.

- `shockval(#)` is the amount to shock `shockvar(varname)` by. Most commonly, a +/- one standard deviation shock is specified.

The following options are not required:

- `diffs(numlist)` is a numeric list of the number of contemporaneous first differences (i.e., $t-(t-1)$) to include for each variable. Note that the first entry (the placeholder for the `depvar`) will always be empty (denoted by "."), since the first difference of the dependent variable cannot appear on the right-hand side of the model.[4]

- `lagdiffs(numlist)` is a numeric list of the number of lagged first differences to include for each variable. For instance, to include a lag at $t-2$ for `depvar`, a lag at $t-1$ for the first weakly exogenous regressor, and none for the second, specify `lagdiff(2 1 .)`. NOTE: the current version does not allow for consecutive

[4]It can however, appear in *lagged* first differences, as shown below.

lagged differences.[5]

- `level(numlist)` is a numeric list of variables to appear in levels (i.e., not lagged or differenced but appearing contemporaneously). If both `level( )` and `ec` are specified, `dynpss` will issue a warning message.[6]

- `ec` if specified, `depvar` will be estimated in first differences. If estimating an error correction model, users will need to use this option.

- `range(#)` is the length of the scenario to simulate. By default, this is $t = 20$. Note that the range must be larger than `time( )`.

- `sig(#)` specifies the significance level for the percentile confidence intervals. The default is for 95% confidence intervals.

- `time(#)` is the scenario time in which the shock occurs to `shockvar( )`. The default time is $t = 10$.

- `saving(string)` specifies the name of the output file. If no filename is specified, the program will save the results as "dynpss_results.dta".

- `forceset(numlist)` by default, the program will estimate the ARDL model in equilibrium; all lagged variables and variables appearing in levels are set to their sample means. All first differences and any lagged first differences are set to zero. This option allows the user to change the setting of the lagged (or

[5]Using the `lagdiff(2 1 .)` example, this means that Δy_{t-2} and $\Delta x1_{t-1}$ will be included, but not: Δy_{t-1}.

[6]Of course, users may have a valid reason to include a variable in levels; for instance, a dummy variable.

unlagged if using `levels( )`) levels of the variables. This could be useful when estimating a dummy variable. For instance, when we wish to see the effect of a movement from zero to one.

- `sims(#)` is the number of simulations (default is 1000). If confidence intervals are particularly noisy, it may help to increase this number. Note that you may also need to increase the matsize in Stata.

- `burnin(#)` allows `dynpss` to iterate out so starting values are stable. This option is rarely used. However, if using the option `forceset( )`, the predicted values will not be in equilibrium at the start of the simulation, and will take some time to converge on stable values. To get around this, one can use the `burnin` option to specify a number of simulations to "throw out" at the start. By default, this is 20. Burnins do not change the simulation range or time; to simulate a range of 25 with a shock time at 10 and a burnin of 30, specify: `burnin(30) range(25) time(10)`.

- `graph` although `dynpss` saves the means of the predicted values and user-specified confidence intervals in `saving`, users can use this option to automatically plot the dynamic results using a spikeplot. As an alternative, by adding the option `rarea`, the program will automatically create an area plot. Predicted means along with 75, 90, and 95 percent confidence intervals are shown using the area plot.

- `expectedval` by default, `dynpss` will calculate predicted values of the dependent variable for a given number of simulations. For every simulation, the

predicted value comes from a systematic component as well as a single draw from the stochastic component. With the `expectedval` option, the program instead calculates expected values of the dependent variable such that the average of 1000 stochastic draws now becomes the estimate of the stochastic component for each of the simulations. This effectively removes the stochastic uncertainty introduced in calculating $\hat{Y}_t$. Predicted values are more conservative than expected values. Note that `dynpss` takes longer to run if calculating expected values.

Users can find examples using `dynpss` on the program's website: http://andyphilips.github.io/dynps

1.3 pss

The `R` package `pss` has been developed to implement both `pssbounds` and `dynpss` in `R` (Jordan and Philips 2016). Since this package is still under development, the current version only allows for the `pssbounds` function to be run. Users can find this package on GitHub by using this link: https://github.com/andyphilips/pss; it also contains details about how to use the `devtools` package (Wickham and Chang 2015) to load `pss` into `R`. In future iterations of this package, `pss` will be uploaded to CRAN.

The `pssbounds` function appears as the following:

```
pssbounds(obs, fstat, tstat = NULL, case, k)
```

the following options are required:

- `obs` is the number of observations (i.e., length of the series) from the ARDL-bounds model. Since the critical values of the bounds test depend on the size of the sample, this option is required.

- `fstat` is the value of the F-statistic from the test that all variables appearing in levels are jointly equal to zero.

- `case` identifies the type of case of the restrictions on the intercept and/or trend term. Case type can be given in Roman numerals ("I","II","III","IV","V") or numerically (1,2,3,4,5). Since the critical values of the bounds test depend on the assumptions placed on the intercept and trend, this option is required.

- `k` is the number of regressors appearing in levels in the ARDL-bounds model. Since the critical values of the bounds test depend on the number of regressors, this option is required.

An additional option, `tstat`, is the value of the one-sided t-test that the coefficient on the lagged dependent variable is equal to zero. Only asymptotic critical values are currently available, and only for cases I, III, and V.

2 Summary of Monte Carlo Results

In the next section, I implement eight Monte Carlo experiments. For brevity, the results from these experiments are summarized in Table 1. I first examine the level of Type II error when varying the adjustment parameter and long-run multiplier. I

find that "fast-moving" adjustment parameters are easier to detect with cointegration tests. Thus, the slow-moving adjustment parameter results shown in the main article should be seen as an especially difficult scenario for cointegration testing. I also evaluate how often a given cointegration test avoids Type I or II error when the other three tests commit these errors. I find that if the bounds test concludes cointegration, it is very likely it exists. If the bounds test does not find cointegration in short series ($\boldsymbol{T} \leq 50$) with more than one regressor, it is advisable to side with the other three cointegration tests (assuming they all conclude cointegration). In long series, users should only rely on the bounds test since the rate of Type II error is low. In addition, I examine how well the ARDL-bounds and Engle-Granger cointegration tests perform for fractionally cointegrated data-generating processes. Results suggest that both tests have high rates of Type I error when the series have finite variance, but less so (especially for the ARDL-bounds) when it has infinite variance. Both procedures have extremely low rates of Type II error.

While cointegration testing is important, so too is the ability to get our substantive hypotheses correct.[7] The last four Monte Carlo experiments in the Supplemental Materials examine the ability to recover the short- and long-run effect, coefficient on the lagged independent variable, and adjustment parameter for a cointegrating relationship. I find that the short-run effect is often recovered, but that the long-run effect is often over- or underestimated. Adjustment parameters tend to converge on their true value as the length of the series increase. I next examine how well the ARDL-bounds and GECM recover null effect sizes (i.e., when the I(1) series are *not*

[7]This is similar to the approach taken by Enns et al. (2016).

Table 1: Summary of Additional Monte Carlo Experiments

Location	Test	Parameters Manipulated	Summary
SM Figure 1	Type II error of cointegration test	LRM, adj. param.	Low Type I error with fast adjustment parameter (nearing -1.0). E-G and bounds test have lower Type II error when adjustment parameter slows (nearing -0.0), with E-G performing best of all. LRM has no effect on cointegration tests.
SM Figure 2 and 3	Type I error, cointegration correct and incorrect discordant	T, k, ϕ_1	When $k = 2,3,4$, bounds test correctly finds evidence of *no* cointegration when all other tests incorrectly conclude cointegration 5 to 13 percent of the time; Johansen rank performs slightly better than bounds when $k = 1$. Bounds test consistently lowest at committing Type I error when all other tests do not. When $k = 1$, Johansen BIC commits Type I error when all others do not at rates over 20 percent. For $k = 3,4$ E-G is often over 20 percent; increasing T does not alter results substantially.
SM Figure 3 and 4	Type II error, cointegration, correct and incorrect discordant	T, k	E-G correctly concludes cointegration when all other tests do not at between 10 and 45 percent. Rate is increasing in T, and highest when $k = 2,3$. Johansen BIC fails to find cointegration when all other tests do when $T = 80$, Johansen test performs poorly when $k = 1$. Bounds test fails to find cointegration when all other tests do when $k > 1$ and $T = 35, 50$.
SM section 3.4	Fractional (Co)Integration (Type I and Type II)	$T, I(d-b)$	Both E-G and bounds detect fractional cointegration (both finite and infinite variance) at nearly 100 percent. Extremely high rates of Type I error when DGP has finite variance. Bounds has lower Type I error and slightly higher Type II error (when $T = 35$) than E-G.
SM section 3.5	Recovery of Cointegrating Effects	T, ρ	GECM and ARDL-bounds nearly always recover correct short-run effect. Long-run effect often over/under estimated, while the coefficient on the lagged independent variable is recovered accurately. Adjustment parameter biased towards -1.0 when $T = 35$, with large improvements as T increases. Spread of estimates slightly larger for ARDL-bounds.
SM section 3.6	Avoidance of Spurious I(1) Effects	T, ρ	GECM and ARDL-bounds nearly always recover correct short-run effect of zero. Autocorrelation increases spread of long-run effect estimates, which are often statistically significant from zero. Adjustment parameter biased towards -1.0 when T is small.
SM section 3.7	Recovery of Stationary Effects	T, ρ	GECM and ARDL-bounds nearly always recover correct short-run effect. Long-run effect not biased from true value but high sampling variability in small T and high ρ. Adjustment parameter recovered unless ρ is high.
SM section 3.8	Avoidance of Spurious Stationary Effects	T, ρ	GECM and ARDL-bounds nearly always recover correct short-run effect. Long-run effect often correctly recovered when ρ is high. Adjustment parameter based towards -1.0 when $T = 35$ and $\rho = 0$; estimates move towards 0 when $T = 80$ and $\rho = 0.5$.

Note: SM= Supplemental Materials, E-G= Engle-Granger, T = length of series, k = number of regressors, ϕ_1 = level of autocorrelation in a single, stationary regressor, LRM= long-run multiplier, adj. param.= adjustment parameter of the lagged dependent variable, $I(d-b)$ = lag operator of a (not) fractionally integrated series with (in)finite variance, ρ = level of residual autocorrelation.

cointegrating). I find that while the short-run effect of zero is often recovered, both the GECM and ARDL-bounds often show statistically significant long-run effects, underscoring the importance of testing for cointegration before running these models. In the last two experiments, I examine the ability of the GECM and ARDL-bounds models to recover actual and spurious stationary relationships. Similar to the I(1) example, I find that the true short-run effect is almost always recovered, while the long-run effect often is not. This is because the long-run effect is a combination of the coefficient on the lagged independent variable—which is recovered at rates similar to the short-run effect—and the adjustment parameter, which is sensitive to the length of the series and the level of autocorrelation.

The full setup, results, and discussion from these eight Monte Carlo experiments can be found below.

3 Additional Monte Carlo Results

In the main paper, I examined the ability of the bounds test to fail to reject the null hypothesis of no cointegration when the null hypothesis is true (avoid Type I error) and correctly reject the null when it is false (avoid Type II error). This was compared to the most common test for cointegration (the Engle-Granger procedure), as well as the Johansen test for cointegration, using both Rank and BIC statistics. One of the crucial steps before running the bounds test is to ensure that no autocorrelation remains in the residuals. In the main paper, I made a restriction of four lags per variable for nearly all simulations; the one exception being for the simulations with 35

observations, which had a maximum of three lags per variable. For example, if there were three independent variables and 50 observations, up to 16 lags (four for each of the three x_{kt} variables and four for y_t) of the first differences were possible. The simulations were designed to iterate through possible lag combinations, choosing the one with the smallest SBIC. Note that augmenting lags with the same restrictions were allowed for both the Engle-Granger and Johansen procedures.

In order to be as comprehensive as possible, I conducted a number of additional Monte Carlo simulations which are presented in this section. These include an examination of the "discordant rates" of the four cointegration tests, fractional integration and the performance of the ARDL-bounds procedure, and the ability of the ARDL-bounds procedure to recover effect sizes (or lack thereof) of cointegrating or stationary data-generating processes.

3.1 Lags and Overfitting in the Monte Carlo Analysis

Since overfitting in small series poses a danger for inference, I calculated the average number of lagged first-differences needed to minimize SBIC from the two main Monte Carlo experiments in the main paper, along with the standard deviations. These are shown in Table 2.

Recall that for the Type I Monte Carlo simulation in the main paper, I generated up to four x_{kt} series that were completely unrelated to the dependent variable, y_t. Since these series cannot possibly be cointegrating (in addition, the variable x_{1t} was often autoregressive instead of a unit root), it is not surprising a number of lagged

Table 2: Average Number of Lagged First-Differences: Monte Carlo Simulations in the Main Paper

		Type I		**Type II**	
T	Number of X	Mean	Std. Dev.	Mean	Std. Dev
35	1	2.10	1.98	2.04	1.96
	2	3.04	2.73	2.86	2.67
	3	4.35	3.79	3.98	3.71
	4	7.49	4.68	7.18	4.84
50	1	0.18	0.76	0.23	0.86
	2	0.33	1.28	0.43	1.51
	3	0.75	2.31	0.82	2.37
	4	1.93	4.33	2.31	4.66
80	1	0.09	0.52	0.07	0.48
	2	0.08	0.60	0.08	0.55
	3	0.07	0.64	0.12	0.89
	4	0.08	0.77	0.13	1.09

Note: Table shows the average number of lagged first-differences across both Monte Carlo experiments (e.g., "Type I" and "Type II" error). The combination of lags was chosen to minimize SBIC. A restriction of a maximum of 4 lags for the dependent and each independent variable was imposed for $T = 50, 80$ and a restriction of 3 was imposed for $T = 35$.

first-differences were needed in order to optimize SBIC, especially when a series had only 35 observations. Yet, as shown in Table 2, as the number of observations increase, the average number of lagged-first differences needed in a model sharply decreases; when the series are 80 or more, the average number of lags converges on zero, even when there are four regressors in the model.

For the Monte Carlo simulation investigating Type II error in the main paper, I created a cointegrating relationship between the dependent and independent variables and examined if the cointegration tests could detect this relationship. The average number of lagged first-differences needed to minimize SBIC is remarkably similar across both the Type I and Type II simulations. A large number of lagged first-differences are necessary when the series contain only 35 observations; on average over seven are necessary when there are four independent variables. However, for 50 and 80 observations, the average number of lagged first-differences necessary to minimize SBIC is nearly always below one.

To conclude, I find that while lagged first-differences are necessary in order to proceed with the bounds cointegration testing procedure, they appear to be in danger of overfitting only in extremely short samples (i.e., $\boldsymbol{T} = 35$). Therefore, practitioners should use SBIC in conjunction with autocorrelation tests—as well as theory—to ensure that the residuals are white noise. If a model with five lags minimizes SBIC more than a model of two lags, but both appear to contain white noise residuals, users should pick the more parsimonious one in short series so as not to overfit their model. Users should also take care not to include too many regressors in short series; one suggestion put forth by Keele, Linn and Webb (2016, p. 40) suggests a minimum

of between 10 and 20 observations per parameter estimated.

3.2 Type II Error of the Cointegration Tests: Varying Adjustment Parameters and Long-Run Multipliers

Although the second Monte Carlo experiment in the main paper varied the number of observations and the number of cointegrating regressors, the size of the long-run effect, as well as the rate of adjustment, may also influence the performance of these cointegration tests, since these parameters determine the magnitude (the accumulated effect that a change in x_t has on y_t) and speed of the cointegrating relationship (how fast a move to disequilibrium corrects back to a stable equilibrium). To investigate this I created another Monte Carlo experiment to examine Type II error. I hold the number of observations ($T = 50$) and regressors ($k = 3$) fixed. The following data-generating process was used, which is somewhat similar to the second Monte Carlo experiment in the main paper:

$$x_{kt} = x_{kt-1} + \nu_{kt} \tag{1}$$

$$u_t = \alpha_1 u_{t-1} + \eta_t \tag{2}$$

$$y_t = \kappa_1 x_{1t} + \kappa_2 x_{2t} + \kappa_3 x_{3t} + u_t \tag{3}$$

Note that the errors ν_{kt} and η_t are independent. Using the same four cointegration tests as in the main paper (bounds, Engle-Granger, Johansen (rank), Johansen (SBIC)), I conducted 500 simulations across each of the following combinations:

- Varying the adjustment parameter: $\alpha_1 = 0.01, 0.30, 0.60, 0.90, 0.99$
- Varying the size of the long-run multiplier for each of the three series x_{kt}: $\kappa_k = 0.01, 3.0, 5.0$. Note that each series has the same magnitude for the long-run multiplier.

The results of the third Monte Carlo experiment are shown as contour plots in Figure 1. The long-run multiplier of the series, x_{kt}, is shown on the horizontal axis. The adjustment parameter is shown on the vertical axis, as it would appear in an error-correction model; adjustment parameters approaching zero indicate a slow rate of adjustment, while those approaching -1 indicate a fast rate of adjustment back to equilibrium. Lighter regions on the contour plot indicate a low proportion of simulations that find evidence of cointegration (high Type II error), while darker regions indicate that the test correctly finds evidence of cointegration at higher rates (low Type II error).

It is clear from Figure 1 that in terms of Type II error, the bounds test falls somewhere between the Engle-Granger and Johansen tests. This confirms the findings from the second Monte Carlo experiment in the main paper. All tests tend to find cointegration only about 10 to 20 percent of the time if the adjustment parameter is slow-moving, yet this increases to over 60 percent as the adjustment parameter converges on -1. The Engle-Granger approach still appears to be the best-performing cointegration test for Type II error, since it is able to find cointegration over 60 percent of the time, as long as the adjustment parameter is between -0.30 and -1. The bounds test finds cointegration over 60 percent of the time when the adjustment

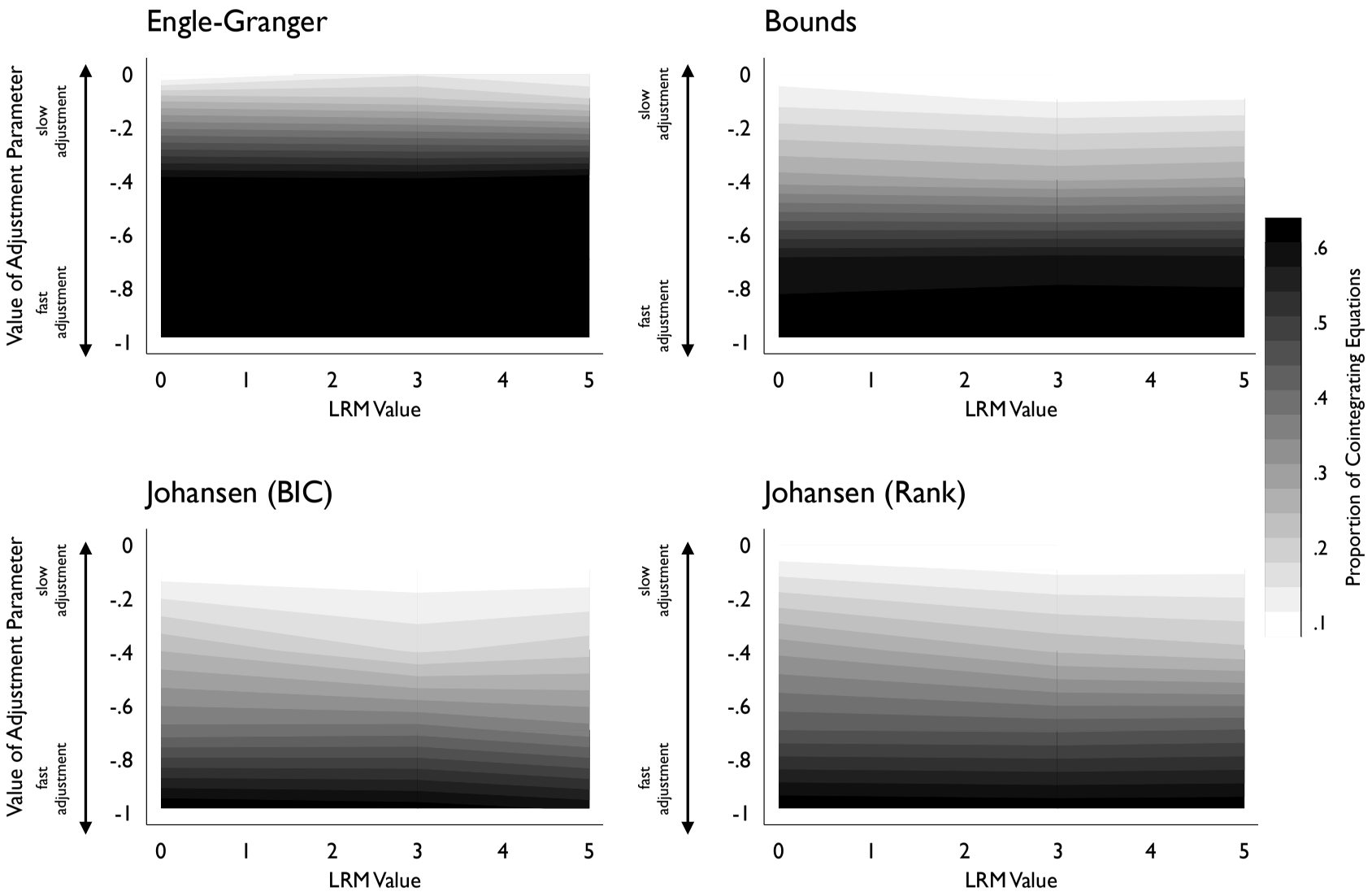

Figure 1: Contour Plots of Proportion of (correctly) Cointegrating Equations Found: $T = 50$, Three Independent Variables

Note: Higher proportion of (correctly) identified cointegrating equations shown in darker grayscale; lower proportion shown lighter grayscale. Adjustment parameter varies as follows: $0.01, 0.30, 0.60, 0.90, 0.99$, and long-run multiplier varies as follows: $0.01, 3.0, 5.0$.

parameter is between -0.6 and -1, and the Johansen tests (both rank and BIC), between -0.90 and -1. Although the bounds test has a higher rate of Type II error than the Engle-Granger procedure, it tends to perform better than either of the Johansen tests.

It is also clear from Figure 1 that the Type II error simulation in the main paper was an especially difficult experiment for the cointegration tests. In that experiment, the adjustment parameter was fixed at -0.25, a relatively slow rate. As shown in Figure 1, it is hard to detect cointegration when the adjustment parameter is slow.

Figure 1 also shows that the size of the long-run effect makes almost no difference in terms of performance, as evidenced by only slight variation across the horizontal axis. Thus, we can conclude from these results that all cointegration tests are better at picking up cointegration when the rate of adjustment is fast, and that the magnitude of the long-run multiplier has virtually no effect on Type II error rates.

3.3 Discordant Cointegration Tests

The two Monte Carlo experiment results in the main paper are just one way of evaluating cointegration tests. Since all four cointegration tests were run on the same data generated for each of the simulations, another way to evaluate them is to observe how often a cointegration test correctly (incorrectly) identifies cointegration when all other tests fail (succeed) in doing do so.[8] These "discordant" Monte Carlo results can be examined for both Type I and Type II error.

[8] I thank an anonymous reviewer for suggesting this.

For the Type I error, a "correct discordant" result is when a particular cointegration test correctly finds evidence of *no* cointegration when *all others* erroneously find cointegration. An "incorrect discordant" result is when a particular cointegration test incorrectly rejects the null hypothesis of no cointegration when all other tests fail to reject the null. The opposite is true for Type II error; a correct discordant result means the test identifies cointegration when all other tests do not, and an incorrect result means that a test fails to find cointegration when all other tests do:

- **Type I Error, Correct Discordant**: How often does a cointegration test find evidence of no cointegration when all other tests erroneously find cointegration?
- **Type I Error, Incorrect Discordant**: How often does a cointegration test find evidence of cointegration when all other tests correctly fail to find cointegration?
- **Type II Error, Correct Discordant**: How often does a cointegration test find evidence of cointegration when all other tests fail to find cointegration?
- **Type II Error, Incorrect Discordant**: How often does a cointegration test fail to find evidence of cointegration when all other tests correctly find cointegration?

3.3.1 Type I Error, Correct Discordant

The proportion of correct discordant simulations that avoid Type I error is shown in Figure 2.[9] Higher values are more desirable, since they correspond with a higher proportion of times that the cointegration test avoided Type I error when all other tests committed it. Across these results the bounds test is the clear outlier. When there are four independent variables, the bounds test avoids Type I error when the three other cointegration tests all commit Type I error—between 9 and 14 percent of the time—depending on the length of the series. The rates of discordant results for the bounds test gets substantially larger as the number of independent variables increases, yet only incrementally increases as the length of the series increases. In sum, these results suggest that the bounds test often correctly avoids a false conclusion of cointegration when all other tests erroneously do.

3.3.2 Type I Error, Incorrect Discordant

For the Type I Monte Carlo simulations, we can also examine how often a cointegration test commits Type I error when all other tests avoid it. This is shown in Figure 3. In this figure, lower values are preferred since they indicate that the cointegration test seldom (erroneously) diverges from the other tests to find evidence of cointegration when it does not exist.

As with Figure 2, there are very little differences in discordant rates across the number of observations; as expected, the likelihood of committing Type I error when

[9]These results use the same data as the ones discussed in the main paper.

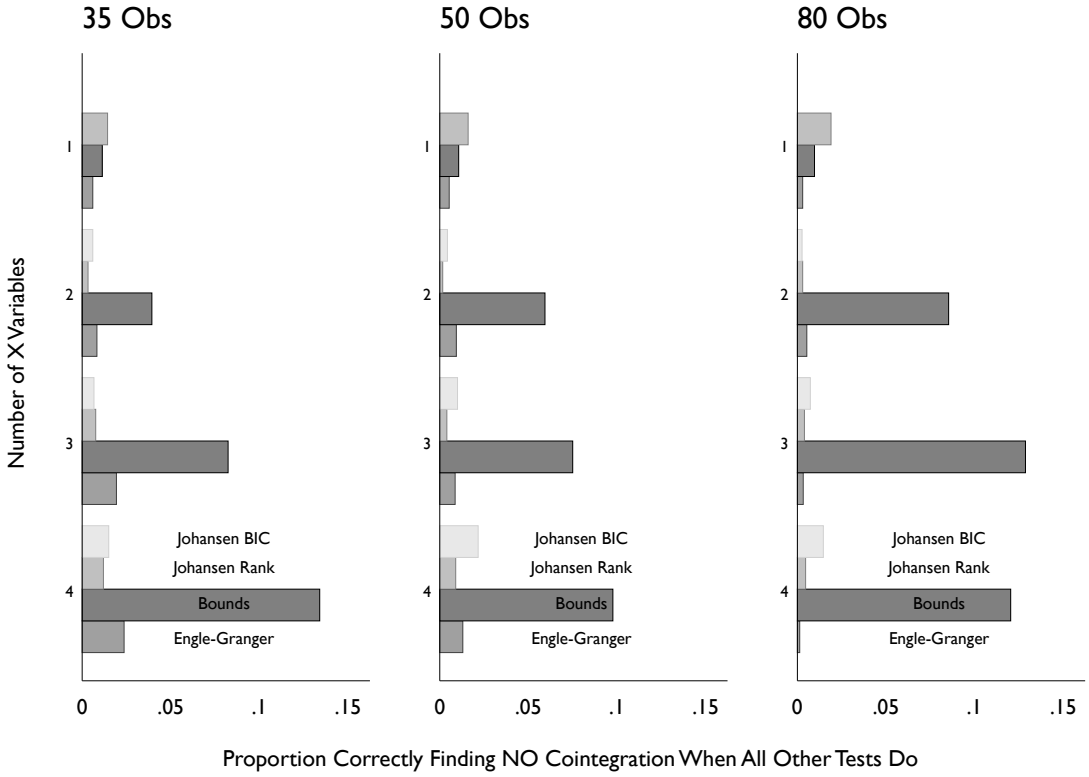

Figure 2: Discordant At Avoiding Type I Error Across the Four Cointegration Tests

Note: An instance of correct discordant means that cointegration test C does not find cointegration (avoids Type I error) when all other tests erroneously find cointegration (commit Type I error).

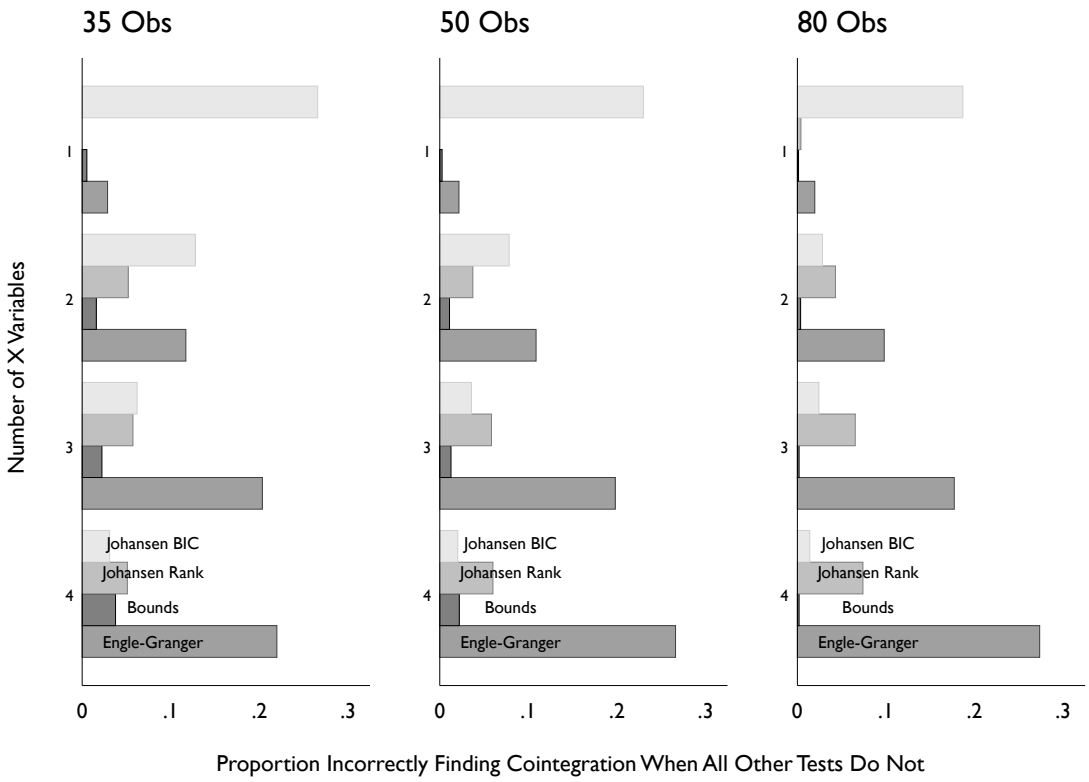

Figure 3: Discordant At Committing Type I Error Across the Four Cointegration Tests

Note: An instance of incorrect discordant means that cointegration test ***C*** erroneously finds cointegration (commits Type I error) when all other tests fail to find cointegration (avoid Type I error).

all other tests do not tends to decrease as the series get longer. However, there are substantial differences as the number of independent variables increase. When there is one independent variable, the Johansen BIC incorrectly finds evidence of cointegration between 20 and 30 percent of time, when *all* other tests do not. For two independent variables, this is sharply reduced. All of the tests—with the exception of the bounds test—commit Type I error when all other tests do not between 5 and 10 percent of the time. When the number of independent variables is three or larger, the Engle-Granger procedure stands out; it diverges from the other three tests to incorrectly find evidence of cointegration about 22 percent of the time. Moreover, this appears to get worse as the number of observations increase. This is an important finding since the Engle-Granger approach is by far the most common method used to test for cointegration (Gonzalo and Lee 1998). Therefore, users should employ more than one cointegration test to avoid an incorrect discordant finding using the Engle-Granger test. Last, once again the ability of the bounds test to avoid Type I error stands out. The bounds test will incorrectly conclude cointegration when all other tests do not less than five percent of the time when $\boldsymbol{T} = 35$. When $\boldsymbol{T} = 80$, this proportion is effectively zero.

3.3.3 Type II Error, Correct Discordant

The discordant plots also shed light on the relative performance of the cointegration tests in terms of avoiding Type II error. In Figure 4, I plot the proportion of simulations when a cointegration test correctly finds cointegration (avoids Type II error)

when all other tests fail to find cointegration (commit Type II error).[10] In contrast with the Type I results above, the Engle-Granger procedure now stands out as being able to identify true instances of cointegration when all other tests fail to do so. It correctly diverges from the other tests between 30 and 50 percent of the time. This is consistent with the evidence in the main paper that finds that the Engle-Granger test outperformed the other three cointegration tests at finding cointegration when it exists.[11]

3.3.4 Type II Error, Incorrect Discordant

For the last comparison, Figure 5 shows the proportion of times a particular cointegration test fails to find evidence of cointegration (commits Type II error) when all other tests find cointegration (avoid Type II error). In contrast with the other results in this section, the number of observations seems to matter substantially for this type of discordant result. For 35 observations, the bounds test stands out at failing to find cointegration when all others detect it; this rises from about 5 percent of the time for one independent variable to about 15 percent of the time when there are four independent variables. When there are 50 observations, the bounds test performs better, with rates almost always below five percent. In contrast, the Johansen Rank test for cointegration performs badly when there is only one independent variable. For 80 observations, the rank test appears to often fail to find cointegration when all others detect it when there is only a single independent variable. As the

[10] These data are from the second Monte Carlo experiment in the main paper.

[11] Although note that these results are from the case where the rate of adjustment is very slow-moving; as Section 2.2 shows, all Type II error rates fall as α_1 converges on -1.

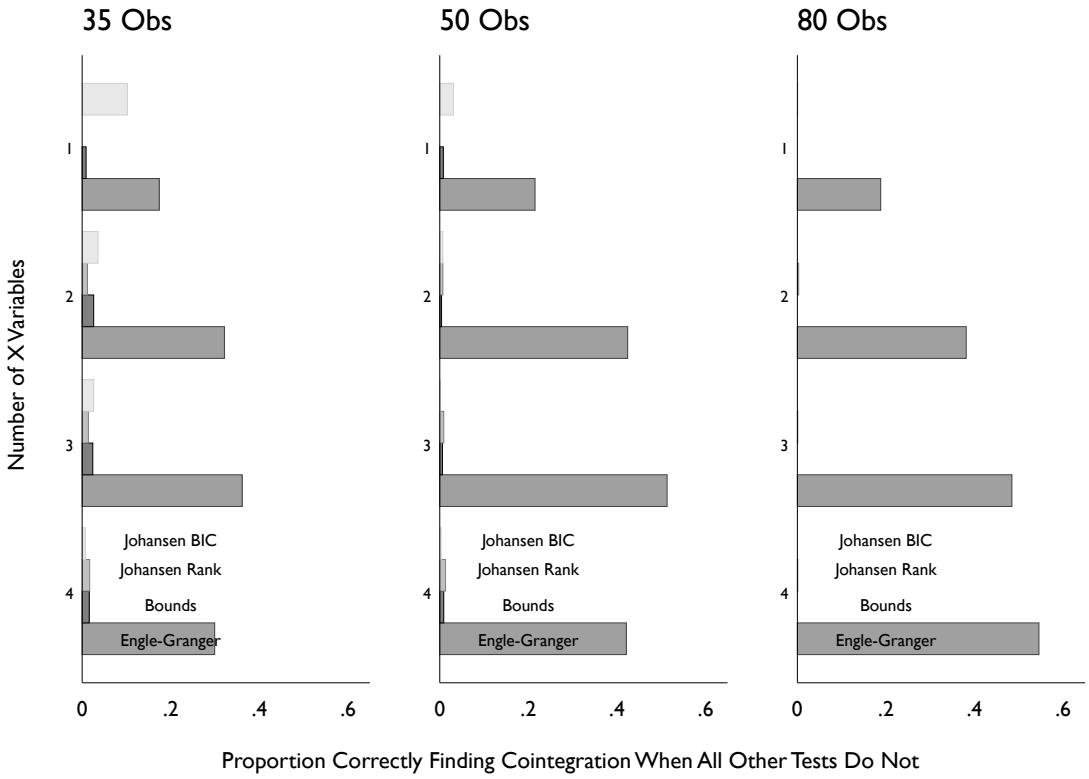

Figure 4: Discordant At Avoiding Type II Error Across the Four Cointegration Tests

Note: An instance of correct discordant means that cointegration test C finds cointegration (avoids Type II error) when all other tests fail to find cointegration (commit Type II error).

number of independent variables increase, the Johansen BIC often fails to find cointegration when all other tests detect it. In contrast, when there are 80 observations, the bounds test performs better relative to the two Johansen tests as the number of independent variables increase.

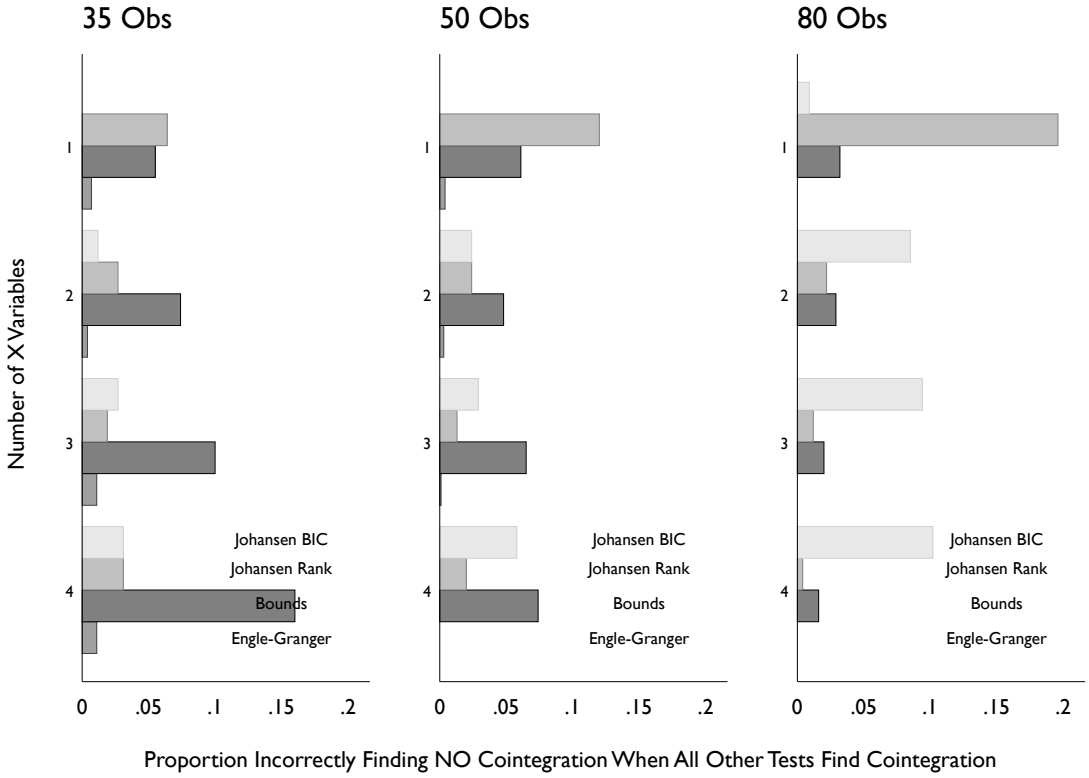

Figure 5: Discordant At Committing Type II Error Across the Four Cointegration Tests

Note: An instance of incorrect discordant means that cointegration test *C* fails to find cointegration (commits Type II error) when all other tests find cointegration (avoid Type II error).

3.3.5 Suggestions for Practitioners for Cointegration Testing

The findings of these Monte Carlo results are largely consistent with the ones in the main paper. Below are several recommendations for practitioners:

- In general, the bounds test excels at avoiding Type I error. It appears to also handle multiple independent variables well, performs well in small samples, and, as shown in the main paper, it remains robust to the accidental inclusion of a stationary regressor. This may be an advantage when using short series (which are common in political science), since unit-root testing is difficult on such series—and, with multiple regressors—we increase the likelihood of accidentally mis-diagnosing at least one as I(0) or I(1). However, the bounds test seems to lack performance in extremely short series, as the results for Type II error for 35 observations showed. Moreover, while it appears to avoid Type II error at higher rates than the Johansen tests, it is still much lower than for the Engle-Granger procedure.

- As shown in the previous section, all cointegration tests are better at detecting cointegration with a fast rate of error correction (adjustment parameter near -1.0) than a slow rate (near -0.0). Therefore, it makes sense to rely more on the bounds test when the adjustment parameter is near -1.0, since it is likely to avoid Type I error yet still have the power to pick up true instances of cointegration. If the adjustment parameter is near -0.0, the Engle-Granger results may be more believable since it is hard for any of the cointegration tests to pick up true instances of cointegration with a slow adjustment parameter.

In addition, I find that the size of the long-run multiplier appears to make no difference in terms of Type II error.

- It is often the case that practitioners utilize multiple statistical tests in order to establish whether or not their conclusion remains robust. However, as the discordant plots have shown, there are often substantial differences among cointegration tests. As shown in the Monte Carlo results in this section, the bounds test often avoids a spurious conclusion of cointegration when all other tests fail to do so. Likewise, the Engle-Granger procedure tends to find truly cointegrating series when all other tests fail to do so. Therefore, there are many times in which going with a "majority" of test results might lead to incorrect conclusions about the nature of the data. Ultimately, the best solution may be reporting results from a large number of unit root and cointegration tests, and—assuming these tests diverge—probing the robustness of the results when different pathways for model specification are considered, as shown by the schematic figure in the main paper. I provide an example of this strategy in the section, "Different conclusions about the time series properties of welfare mood" below.

3.4 Fractional Integration and the ARDL Procedure

Although the method of Pesaran, Shin and Smith (2001) is designed to be run *only* when the dependent variable is I(1), to the best of my knowledge there has been no analysis of the performance of the bounds test for cointegration when series are

fractionally cointegrated in small samples.[12] This section does not advocate using the bounds procedure over methods designed to handle fractional integration (c.f. Box-Steffensmeier and Smith 1998; Grant and Lebo 2016). Nor does it discuss the performance of approaches designed to model fractional integration; this has recently been investigated in small series like the ones discussed in the main paper (Helgason 2016; Esarey 2016; Keele, Linn and Webb 2016). However, since determining if a variable is I(0), I(1), or some other I(d) can be quite difficult in short series, it is worth examining the performance when the user erroneously concludes that the data are I(1)—and proceeds to run an ARDL-bounds model—when in fact they are fractionally integrated and either spuriously related or fractionally cointegrating.

In this section, I examine how the bounds test performs when two independently generated fractionally integrated series are regressed on one another. I next examine how often we can detect a relationship between a weakly exogenous x_t that is in a fractionally cointegrating relationship with y_t. I largely draw from the setup and R code used by Helgason (2016, p. 60-62), who specifies the following data-generating process:[13]

$$x_t = (1-L)^{-d}\varepsilon_{2t} \tag{4}$$

[12] I thank an anonymous reviewer for suggesting to examine the performance of the bounds test in regards to fractional integration.

[13] While fractionally integrated series are easily generated in `R` using the package `ARFIMA` (Fraley et al. 2006), I estimated the ARDL model on the saved `R` series using Stata, since it was the program used for all other simulations in this paper.

$$y_t - \beta x_t = (1-L)^{-(d-b)}\varepsilon_{1t} \tag{5}$$

In the setup above, let x_t be a weakly exogenous series, since the two error terms, ε_{1t} and ε_{2t}, are independently generated from one another. Let L denote the lag operator, which provides for a fractionally integrated x_t of order $I(d)$, and a fractionally integrated y_t of order $I(d-b)$, where $d > b > 0$. I then generated the following series for $T = 35, 50, 80$, the same number of observations as examined in the Monte Carlos in the main paper:[14]

- **No fractional integration, finite variance**: In this scenario, there is no fractional integration since $d = 0.4$, $b = 0$, and $\beta = 0$. Each of the series is $I(0.4)$—and, as discussed by Helgason (2016, p. 61)—the series have finite variance and are mean reverting, since $d < 0.5$.

- **No fractional integration, infinite variance**: There is no fractional integration in this scenario since $d = 0.8$, $b = 0$, and $\beta = 0$. Each of the series is $I(0.8)$, and so has infinite variance since $d > 0.5$.[15]

- **Fractional integration, finite variance**: In this scenario, the two series are fractionally integrated since $d, b = 0.4$ and $\beta = 0.5$. The linear combination of x_t and y_t is stationary.

- **Fractional integration, infinite variance**: In this scenario, the two series

[14] To mitigate issues involving initial conditions (Balke and Fomby 1997), I first created a burn-in period of $T = 100$ for all scenarios.

[15] As noted by Keele, Linn and Webb (2016), values of $d > 0.5$ are integer-differenced before simulating in the `R` package `ARFIMA`, so $d = -0.2$ in the data-generating process.

are fractionally integrated since $d = 0.8$, $b = 0.6$, and $\beta = 0.5$. However, the variance of the series is infinite. The linear combination of x_t and y_t will be integrated of order $(d-b) = I(0.2)$, which is still mean-reverting and evidence of cointegration, even if the residuals are not I(0) (Cheung and Lai 1993).[16]

I conducted 1000 simulations for each of the four data-generation processes described above for $T = 35, 50, 80$. I then ran the following ARDL model:

$$\Delta y_t = \alpha_0^* + \theta_0 y_{t-1} + \theta_1 x_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \sum_{j=0}^{q} \beta_j \Delta x_{t-j} + \varepsilon_t \tag{6}$$

and used the bounds F-test with the appropriate critical values given by Philips (2016*b*) as a test for cointegration.[17] As with the other Monte Carlo experiments, SBIC was used to determine the number of lagged first-differences of x_t and y_t to include. For comparison, I also ran the Engle-Granger two step procedure in order to test if the residuals of the first-stage were I(0), which would indicate cointegration.[18]

3.4.1 No Fractional Integration, Finite vs. Infinite Variance

The results from the spuriously-related series are shown in Figure 6. The bar-graph on the left shows the case of two unrelated series that are both integrated of order $I(0.4)$. The vertical axis shows the length of the series, while the horizontal axis shows the proportion of simulations that find evidence of cointegration. Since the

[16] As Cheung and Lai (1993) discuss, the cointegrating residuals must be $I(d)$, where $0 \leq d < 0.5$.

[17] A restriction of $p, q \leq 3$ was placed on the maximum number of lag lengths $T = 35$, and 4 for $T = 50,\ 80$.

[18] The number of augmenting lags to include was determined by SBIC. A restriction of $p, q \leq 3$ was placed on the maximum number of lag lengths $T = 35$, and 4 for $T = 50,\ 80$.

two series were created to be independent from one another, this is a form of Type I error. It is clear that both the Engle-Granger and bounds procedures find evidence of cointegration when it does not exist at extremely high rates when both the dependent and independent variable are fractionally integrated and have finite variance. Although this rate decreases by about half for the bounds test when there are 35 observations, it still remains very high.

In contrast, when two unrelated fractionally integrated series have infinite variance (when $I(0.8) > 0.5$), the Engle-Granger and bounds procedures perform much better, as shown by the bar-graph on the right in Figure 6. The bounds test correctly fails to reject the null of no cointegration between 90 and 95 percent of the time when there are only 35 observations, and just below 80 percent of the time when there are 80 observations. The Engle-Granger procedure has a rate of Type I error over twice as much as the bounds test. Interestingly, both the Engle-Granger and bounds procedures tend to have slightly higher rates of Type I error as the length of the fractionally integrated series increases.

3.4.2 No Fractional Integration, Finite vs. Infinite Variance

In Figure 7, I examine how well the Engle-Granger and bounds procedures are able to detect a fractionally cointegrating relationship when it does exist; in other words, the ability of these tests to avoid Type II error when the data-generating process is fractionally cointegrating. Since the two series are fractionally cointegrating, higher proportions indicate the the procedure is better at detecting cointegration.

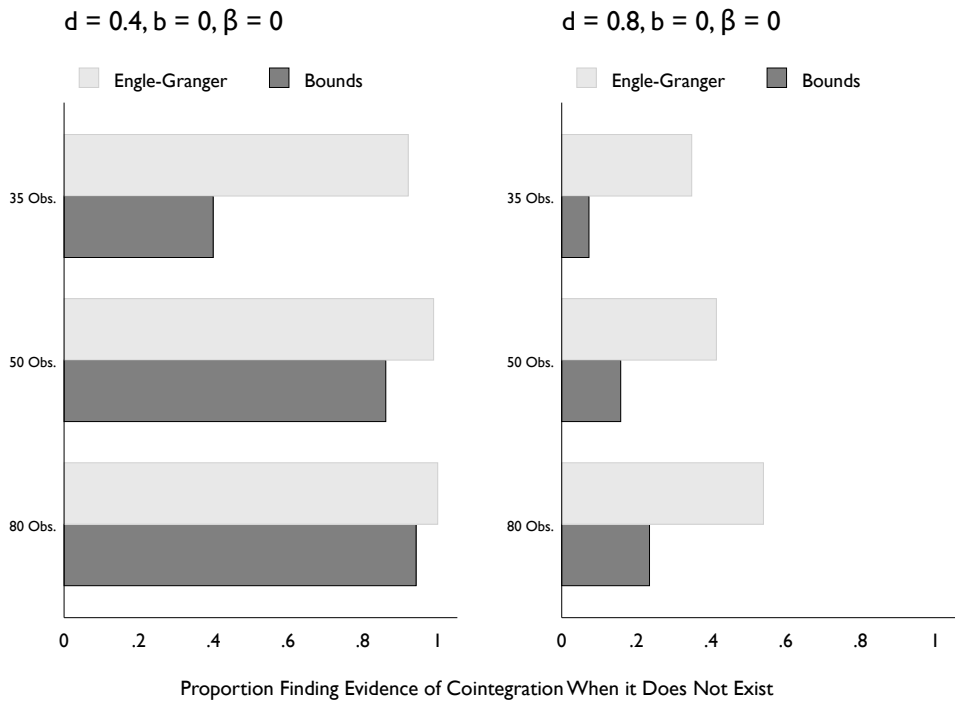

Figure 6: Performance of the Engle-Granger and Bounds Test in the Face of Type I Error

Note: Each bar-graph shows the proportion of simulations finding (at $p < 0.05$) evidence of one cointegrating relationship with one regressor. The bar-graph on the left uses a DGP with finite variance, the right does not. The Engle-Granger procedure is an augmented Dickey-Fuller unit-root test with the hypothesis $H_0 = z_t \sim I(1)$ from $y_t = \hat{\kappa}_0 + \hat{\kappa}_1 x_{1t} + z_t$, with augmenting lags determined by SBIC. Critical values for Bounds test determined by k regressors and number of in-sample observations, with assumption of no trend and unrestricted constant, and lag lengths determined via SBIC.

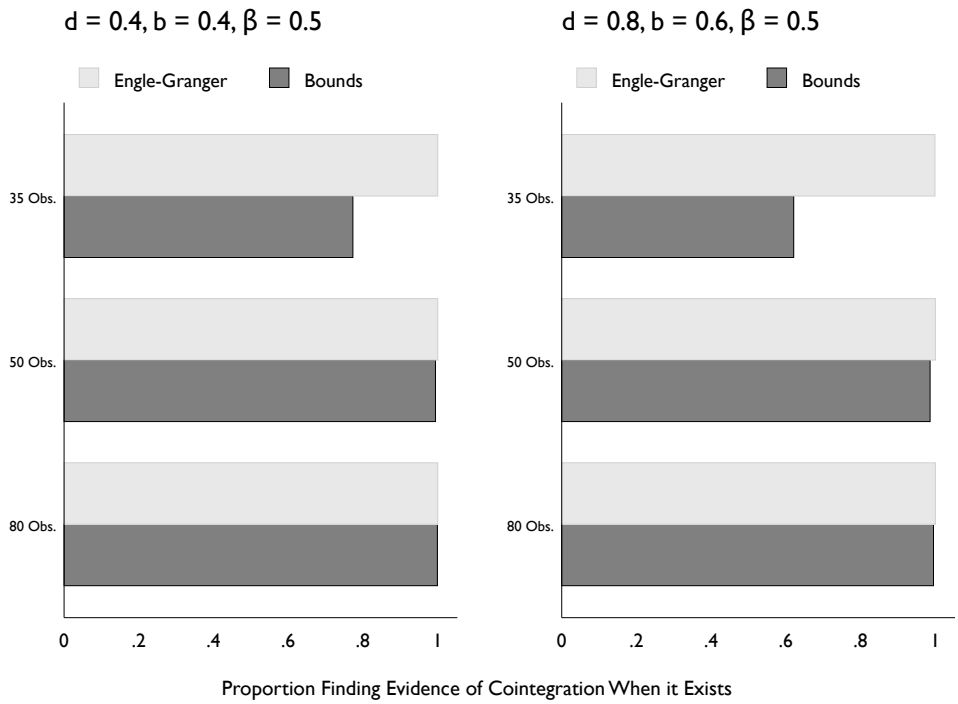

Figure 7: Ability of the Engle-Granger and Bounds Procedures to Find Fractionally Cointegrating Relationships

Note: Each bar-graph shows the proportion of simulations finding (at $p < 0.05$) evidence of one cointegrating relationship with one regressor. The bar-graph on the left uses a DGP with finite variance, the right does not. The Engle-Granger procedure is an augmented Dickey-Fuller unit-root test with the hypothesis $H_0 = z_t \sim I(1)$ from $y_t = \hat{\kappa}_0 + \hat{\kappa}_1 x_{1t} + \cdot + \hat{\kappa}_k x_{kt} + z_t$, with augmenting lags determined by SBIC. Critical values for Bounds test determined by k regressors and number of in-sample observations, with assumption of no trend and unrestricted constant, and lag lengths determined via SBIC.

As shown by the bar-graph on the left of Figure 7, it is clear that when the series have finite variance, both the Engle-Granger and bounds procedures are very good at detecting fractionally cointegrated series that result in an I(0) residual; detection rates approach 100 percent for all cases, except the bounds test for 35 observations.

The bar-graph on the right in Figure 7 shows the results of the cointegration tests from a DGP containing series with infinite variance, which results in an I(0.2) cointegrating residual. The results are very similar to when the cointegrating residual is I(0) (the bar-graph on the left in Figure 7), albeit slightly lower. Thus, under the bivariate DGPs considered in this Monte Carlo experiment, it looks like both the bounds and Engle-Granger procedures are nearly always able to pick up fractionally cointegrating relationships when they exist.

3.4.3 Suggestions for Practitioners

There are a number of suggestions for practitioners based on the results in this section:

- The bounds test is only designed for an I(1) dependent variable. It is crucial that this is tested. Users that find that their dependent variable is I(0) should consult the schematic in the main paper on the appropriate lagged-dependent variable model to employ. If practitioners find evidence that their dependent variable is fractionally integrated, or have strong theoretical reasons for believing it is, they should estimate a model using fractional integration techniques, which typically involve fractionally differencing all series by their $I(d)$ esti-

mate and then estimating an error-correction model (known as a fractional error-correction model) (e.g. Grant and Lebo 2016; Clarke and Lebo 2003). Of course, they should be aware that these techniques may be questionable in small series, since overfitting is possible using ARFIMA models (Helgason 2016; Esarey 2016; Keele, Linn and Webb 2016).[19]

- Procedures used to detect cointegration tend to find fractional integration, even when it does not exist. Although I only considered four different data-generation processes, it is clear that both the bounds and Engle-Granger procedures are able to detect fractionally cointegrating relationships at rates approaching 100 percent. However, they also tend to commit Type I error at extremely high rates if the series have finite variance $d < 0.5$; this rate is much smaller, though still substantial (especially for the Engle-Granger procedure) when the series have infinite variance $d > 0.5$.

3.5 How Well Can the ARDL Procedure Recover Cointegrating Effects?

In the main paper, as well as in the sections above, I investigated Type I and Type II error in terms of cointegration; were we able to avoid spurious conclusions of cointegration, and were we able to detect cointegration when it exists? While this investigated a crucial component in time series that fits in with recent discussions on

[19]For instance, Esarey (2016) suggests that series with less than 100 time points may be difficult to estimate using fractional integration techniques, while Helgason (2016, p. 60) suggests anything less than 250.

spurious regressions in time series (c.f. Grant and Lebo 2016; Keele, Linn and Webb 2016), so too is the ability of our models to correctly recover the effect sizes we seek to test. That is to say, are we able to come to the correct conclusion our hypotheses seek to test? Such an approach has been done in the context of the GECM (Enns et al. 2016), but not with the ARDL-bounds model.[20]

To investigate how well various effect sizes are recovered in the context of true integration using the ARDL-bounds model, and how this compares to the standard GECM, I used a setup similar to the second Monte Carlo experiment in the main paper. The following data-generating process was used to create a relationship between a dependent variable, y_t, and two regressors, x_{1t} and x_{2t}:[21]

$$x_{1t} = x_{1t-1} + \nu_{1t} \tag{7}$$

$$x_{2t} = x_{2t-1} + \nu_{2t} \tag{8}$$

$$u_t = 0.75u_{t-1} + \eta_t + \rho\eta_{t-1} \tag{9}$$

$$y_t = 0.25x_{1t} + 0.25x_{2t} + u_t \tag{10}$$

Note that the errors ν_{1t}, ν_{2t} and η_t are independently generated from one another, thus ensuring weak exogeneity. This DGP yields an adjustment parameter of -0.25, a contemporaneous effect of 0.25, a long run-multiplier of 0.25, and a coefficient on the lagged x_{kt} variable of 0.0625. For a proof of the calculations of these various effects, see Section 3 in the Supplemental Materials. Note also that there will be

[20] I thank an anonymous reviewer for suggesting this section.

[21] To mitigate issues involving initial conditions (Balke and Fomby 1997), I first created a burn-in period of $T = 100$.

autocorrelation in the residuals if $\rho \neq 0$. I explore this possibility since, in short series, models are likely to have some amount of noise or suffer from model misspecification. I examined all of the possible combinations below to see how well we can obtain different effects from a cointegrating data-generating process:

1. Varying the number of observations: $T = 35, 50, 80$.

2. Varying the level of autocorrelation: $\rho = 0.0, 0.2, 0.5$.

The following models were run for the ARDL-bounds and the ECM models, respectively:

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \sum_{j_1=0}^{q_1} \beta_{j_1} \Delta x_{1t-j_1} + \sum_{j_2=0}^{q_2} \beta_{j_2} \Delta x_{2t-j_2} + \varepsilon_t \tag{11}$$

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \Delta\beta_1 x_{1t} + \Delta\beta_2 x_{2t} + \varepsilon_t \tag{12}$$

Augmenting lags of the first-difference of x_{1t}, x_{2t} and y_t were determined by SBIC for the ARDL-bounds model.[22] To compare the performance of the ARDL-bounds approach and the standard ECM, I use two different approaches:

- **Coverage Rates**: What percentage of the time do constructed 95 percent confidence intervals for each of the simulations fail to encompass the true effect size?

[22] A maximum lag restriction of $p, q_1, q_2 \leq 4$ was used for $T = 50, 80$ and a restriction of $p, q_1, q_2 \leq 3$ for $T = 35$.

- **Empirical Distribution**: Taking all simulation parameter estimates together and constructing a single mean and 95 percent confidence interval, what is the likely value (and sampling variability) of the estimate?

Both are important, since they since they help show how well the models get our substantive hypotheses correct. The former is what we would get if we ran a single model and calculated the 95 confidence intervals (or the percent of simulations whose effect size is statistically significantly different from zero at the five percent level). The latter is what we mean (in terms of interpretation) when we construct confidence intervals.

A stylized example of this approach is shown in Figure 8. Each density plot represents estimated coefficients using a different model. The "true" parameter value, shown by the vertical black line, was specified in the hypothetical data generating process. If we were to construct 95 percent confidence intervals for each simulation conducted for each of the three models, and then test the null hypothesis that the parameter was at "True", we would create a coverage rate, or the likelihood that a parameter estimate from a single model fails to model the underlying data-generating process. In Figure 8, we would find that the coverage rate of Model 1 was zero (its constructed 95 percent confidence intervals *never* contained the true value of the parameter; equivalently, we would be able to reject the null hypothesis), the coverage rate of Model 2 would be very high (the constructed 95 percent confidence intervals nearly always contained the true parameter estimate), and the coverage rate of Model 3 would be very low. In terms of our substantive hypotheses, we would never gain correct inferences if we were using Model 1, we would almost always be correct using

Model 2, and would be almost never correct using Model 3.

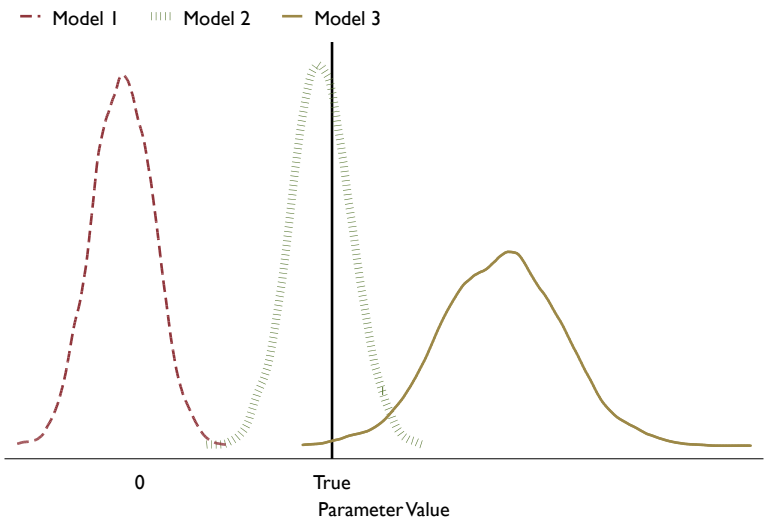

Figure 8: Stylized Example of How Effect Coverage and Empirical Distribution Inform About Inference

For the empirical distribution, we simple calculate the mean and upper- and lower- percentile confidence intervals of the estimated parameters. We can then examine a number of factors, such as how far away the estimated mean is from the mean of the data-generating process, the variability of the distribution of parameters, and if the constructed confidence interval encompasses the true parameter. In Figure 8, this would involve taking the mean and calculating percentiles of each of the three distributions of parameter estimates. We would find that Model 2 had a mean estimate closest to the true parameter, that Model 3 had the largest sampling variability, and that Model 1 had a constructed confidence interval that never approached the parameter value of the data-generating process.

The coverage probabilities for the short-run effect from the Monte Carlo experiment are shown in Figure 9. The horizontal axis shows the proportion of simulations whose constructed 95 percent confidence intervals *did not* contain the parameter value of the data-generating process (0.25 in this case); thus, lower values indicate the method is more likely to recover the true effect size. At conventional standards, we would expect to reject the null hypothesis of the true effects about five percent of the time. This is shown in the figures by the red vertical line at 0.05.

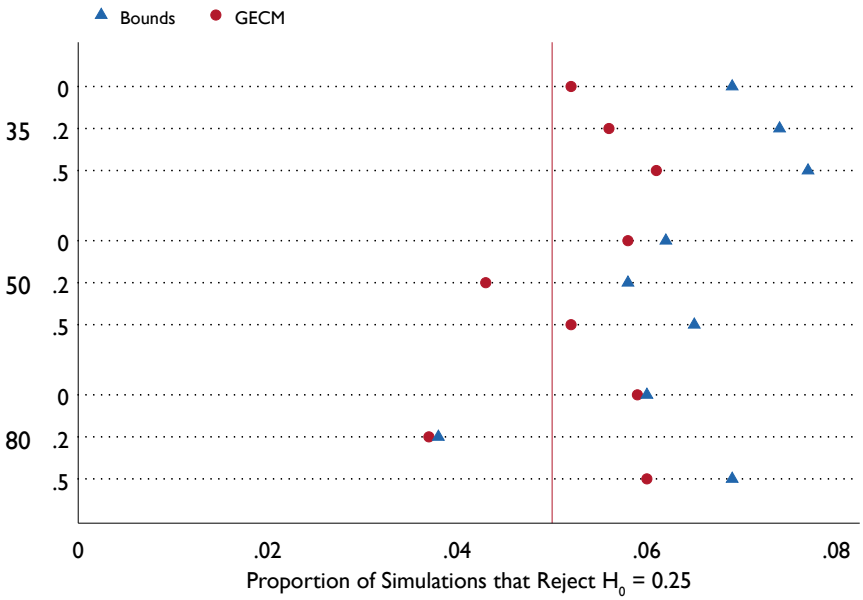

Figure 9: Coverage of the Short-Run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

There are a number of interesting findings in Figure 9. Only a small proportion of simulations had constructed 95 percent confidence intervals that did not overlap

with the short-run effect used in the data-generating process, as evidenced by the clustering around 0.05. The GECM slightly outperforms the ARDL-bounds model, though this difference appears to shrink as the sample size increases. In addition, while we saw above that increased autocorrelation increased absolute bias (especially in the GECM), it actually tends to increase the coverage probability for the GECM, and less so for the ARDL-bounds model. In sum, both the ARDL-bounds and GECM are very good at recovering the true estimate of the contemporaneous effect of x_t on y_t, even in short series, and even when serial correlation of the errors is present.

I next examine the coverage probability of the long-run effect in Figure 10. The largest difference between these results and the coverage of the short-run effect is the large increase in the proportion of simulations that could not recover the long-run effect. Also interesting is that increases in autocorrelation tend to lead to better recovery of the effect size. In addition, the GECM seems to outperform the ARDL-bounds model across the size of the sample and level of residual autocorrelation; this difference is greatest when autocorrelation is high. Overall, it appears that both models have a difficult time accurately recovering the true effect size of the long-run effect.

In Figure 11, I plot the proportion of simulations where the constructed 95 percent confidence intervals did not encompass the coefficient of the lagged independent variable. The results are more similar to the short-run effect than to the long-run effect shown in Figure 10; both models tend to be unable to find the true coefficient of the lagged x_t variable between 7 and 18 percent of the time, depending on the number of observations (as the series lengthens, the models recover the true parameter more

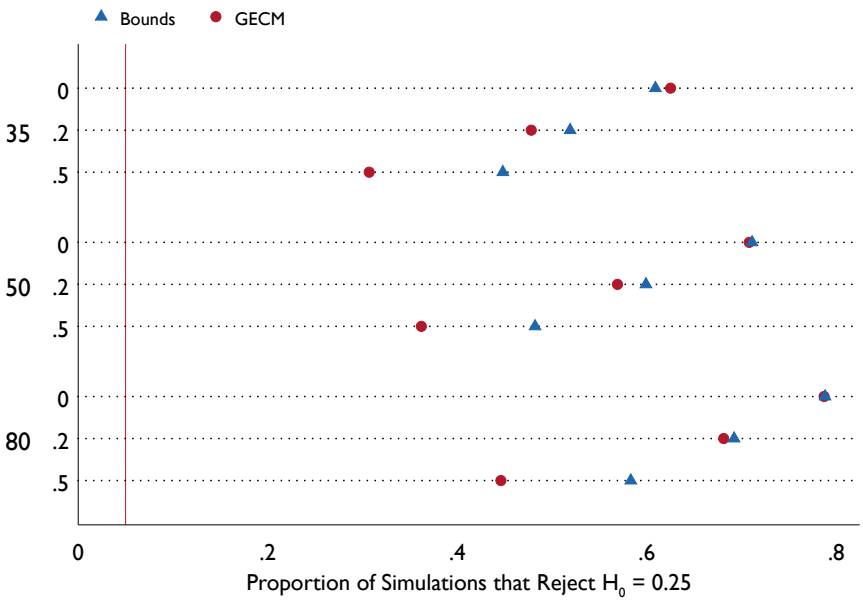

Figure 10: Coverage of the Long-Run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

frequently) and the level of residual autocorrelation (more autocorrelation makes recovery less frequent). Overall, the GECM seems to perform slightly better than the ARDL-bounds model.

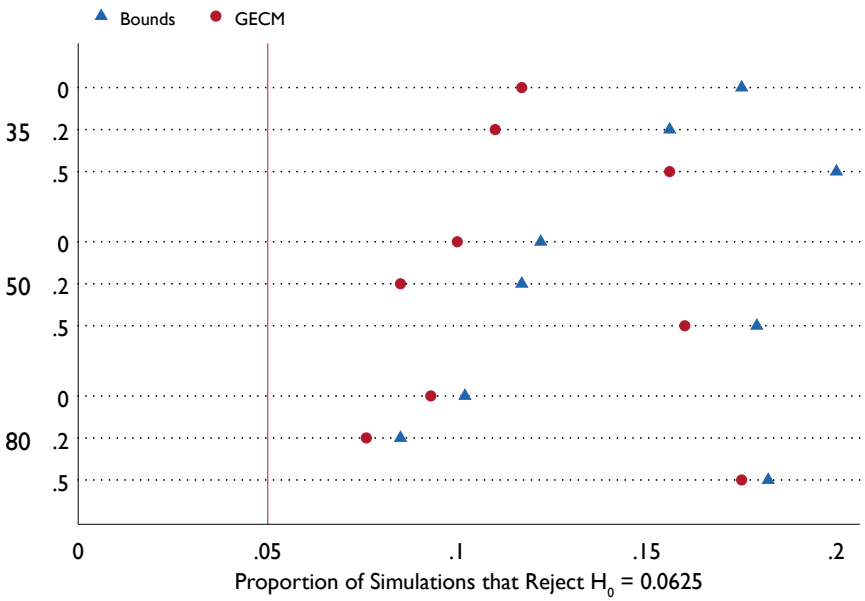

Figure 11: Coverage of the Lagged Independent Variable

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

Last, I plot the coverage probability of the adjustment parameter in Figure 12. A number of interesting results stand out. First, increasing sample size does not improve coverage by much, in contrast to the findings about absolute bias above. Second, the GECM tends to have higher rates of coverage, unless residual autocorrelation is high. Note that when there are 80 observations and $\rho = 0.5$, using the ARDL-bounds model over the GECM results in an improvement in coverage of about

30 percent. Last, it appears that as sample size increases, the difference in coverage between the two methods shrinks, as long as autocorrelation is low (i.e., $\rho = 0.0, 0.2$).

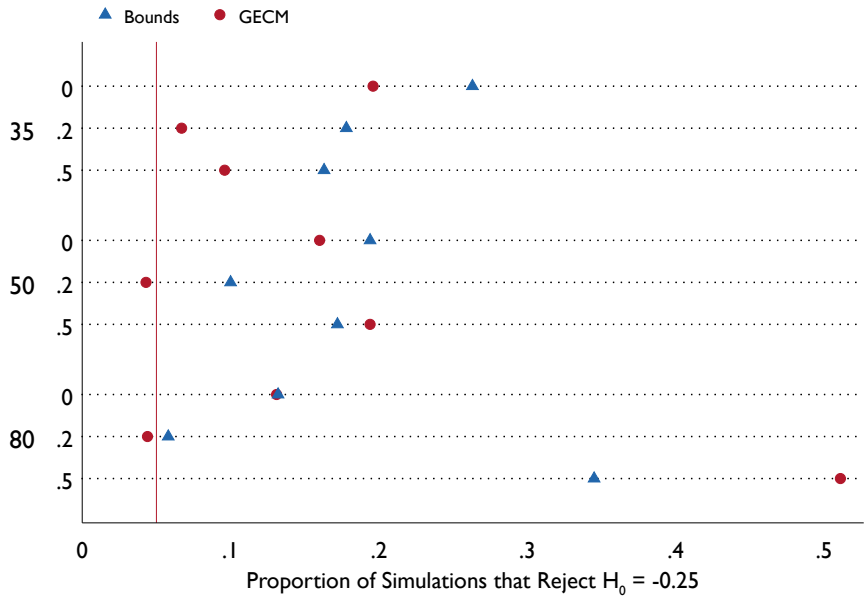

Figure 12: Coverage of the Adjustment Parameter

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

There are a number of conclusions to draw from this simulation experiment. First, both the ARDL-bounds and GECM tend to be very good at recovering the short-run effect, the coefficient on the lagged independent variable, and adjustment parameter. Both models tend to be unable to recover the long-run effect, possibly because this involves a combination of two estimates: the coefficient on the lagged x_t, and the coefficient on the lagged dependent variable. Second, although bias increases and coverage is worse for the smallest sample size, $T = 35$, it is less drastic

than one might expect. In fact in some cases (such as when autocorrelation is high), shorter series may result in better estimates. Third, the ARDL-bounds method offers lower bias than the GECM when residual autocorrelation is present, and possibly an improvement in coverage, though it was less clear for the latter.[23] Fourth, these findings hold for a model of two weakly exogenous regressors. The findings may be quite different for multiple regressors, as many of the Monte Carlo simulations in the main paper found. Last, this Monte Carlo experiment used an lagged dependent variable parameter with a slow rate of adjustment (-0.25); if coverage rates and bias are affected in a similar way to the Type II error performance of the cointegration tests (see the contour plot results in Figure 1), bias may decrease and coverage rates may be much higher with a faster rate of adjustment. Therefore, this represents what is likely to be a difficult test for both models.

The empirical distribution of the short run effect from the Monte Carlo experiment is shown in Figure 13. The vertical axis shows the length of the series as well as the level of residual autocorrelation, while the horizontal axis shows the size of the distribution of the estimated coefficients. Combining all simulations together, there appears to be very little difference between the average parameter estimate and the actual short-run effect from the data-generating process (shown by the vertical line at 0.25). As expected, the constructed 95 percent confidence intervals tend to shrink (more simulations have an estimate that is closer to the mean) as the length of the series increases. Residual autocorrelation appears to increase the variability of parameter estimates somewhat, and the variability of estimates is slightly wider

[23]The GECM had better coverage than the ARDL-bounds for the short-run and long-run effects in the face of autocorrelation, but not for the adjustment parameter.

for the ARDL-bounds than for the GECM. Overall, both models appear to be very good at estimating the actual short-run effect of a cointegrating relationship, with only slight variability in the estimates from sample to sample.

The empirical distribution of the long-run effect is shown in Figure 14. In contrast to the short-run effect, there is a large spread in parameter estimates when autocorrelation is high, and much less when it is low. Note too that although the actual value of the long-run effect is 0.25, the upper- and lower- confidence intervals are much higher or lower in many cases; only when autocorrelation is low—and the length of the series is $T = 50$ or more—do most estimates fall right around 0.25. Last, note that, especially when autocorrelation is high, the ARDL-bounds model has a slightly lower spread of parameter estimates than the GECM.

I plot the empirical distribution of the lagged independent variable, x_{1t-1}, in Figure 15. In contrast to the long-run effect, the spread of parameter estimates lie much closer to the actual parameter value, 0.0625. However, unlike the short-run effect, there appears to be bias in the estimates when there is no autocorrelation; that is to say, taken together, we can expect both models to slightly over-estimate the size of the lagged independent variable (on average). Surprisingly, much of this bias goes away at low levels of autocorrelation (i.e., when $\rho = 0.2$), and models tend to under-estimate the size of the lagged independent variable when autocorrelation is high. Last, unless the length of the series is about 80, using the ARDL-bounds model leads to a slightly greater spread of parameter estimates.

Last, I plot the empirical distribution of the adjustment parameter in Figure

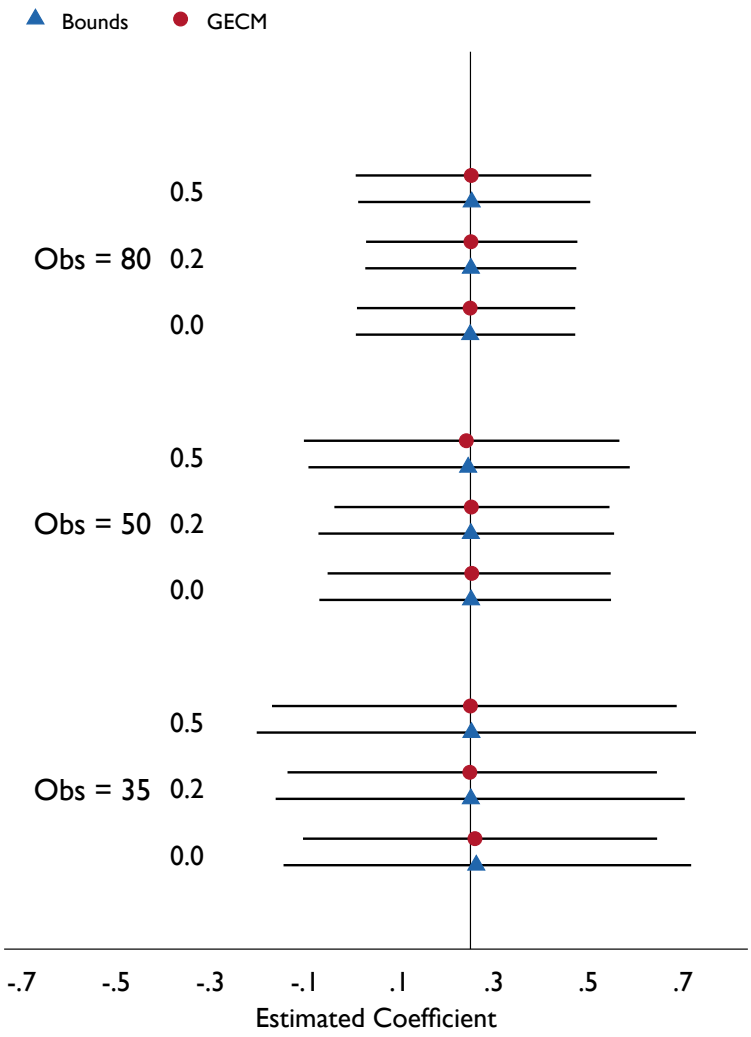

Figure 13: Empirical Distribution of the Short-Run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

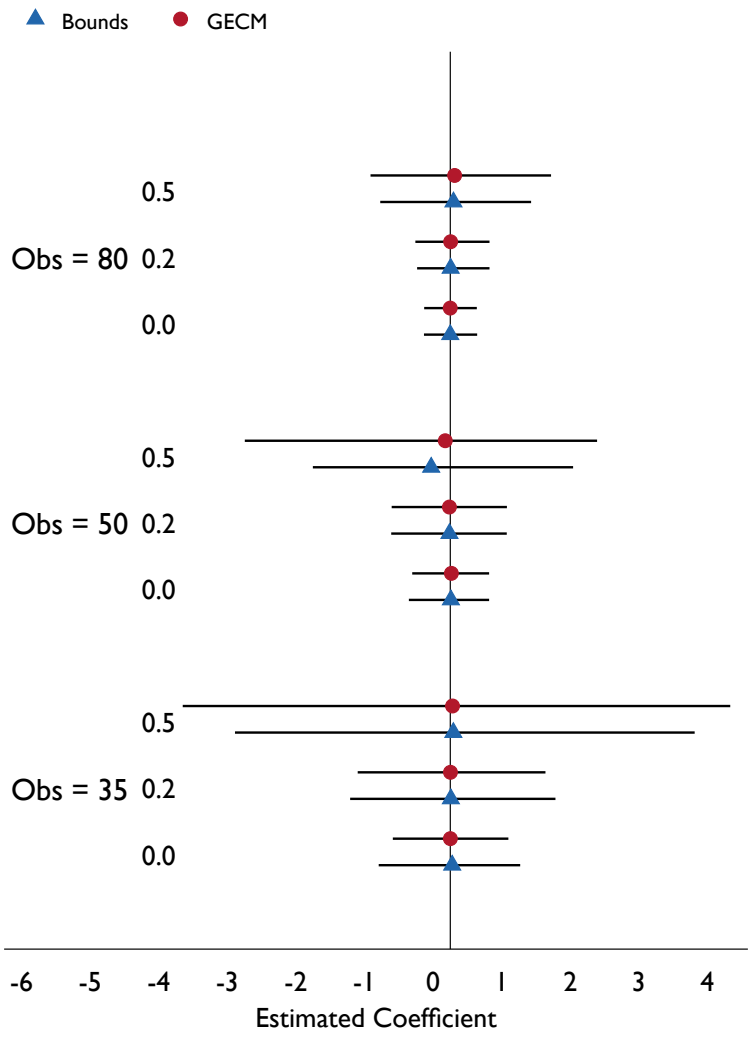

Figure 14: Empirical Distribution of the Long-Run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

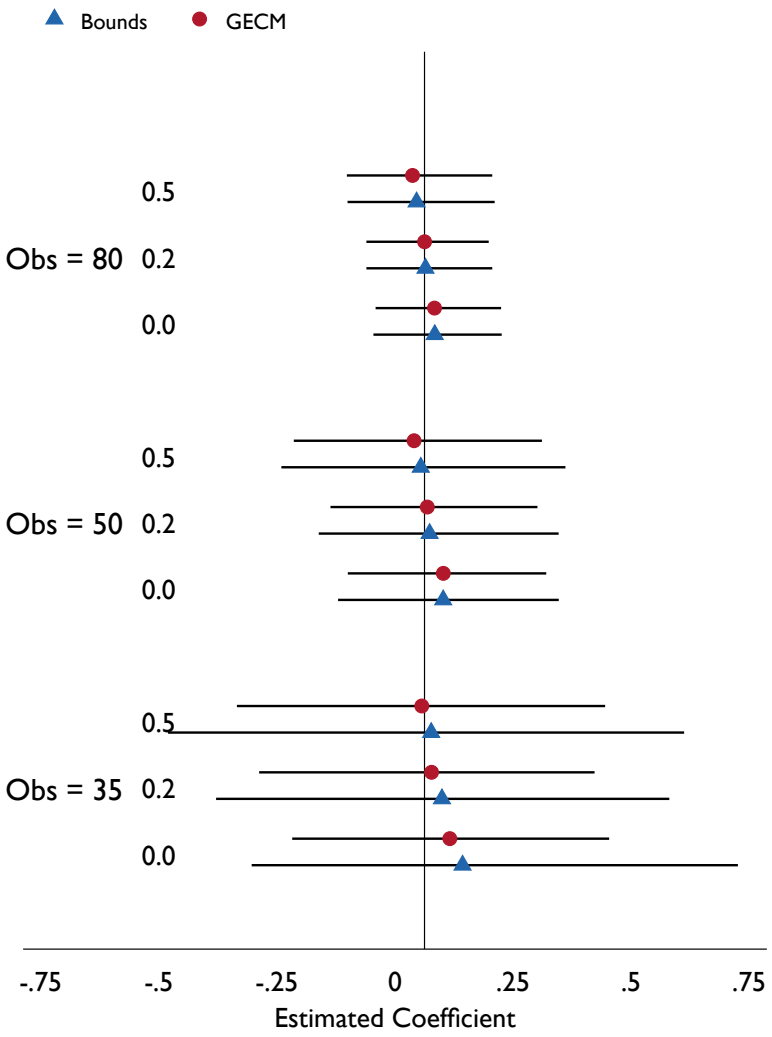

Figure 15: Empirical Distribution of the Lagged Independent Variable

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

16. There are many important findings about the ability of the ARDL-bounds and GECM to recover the correct coefficient on the lagged dependent variable. First, estimates tend to be more negative at low levels of autocorrelation, and less negative as autocorrelation increases. Surprisingly, these shifts are present across all numbers of observations. This suggests that without autocorrelation, estimates are likely to conclude that the adjustment parameter has a faster rate of adjustment (i.e., approaches -1) than it actually does, while with autocorrelation, estimates are likely to conclude a slower rate of adjustment (closer towards zero) than is actually the case. In fact, when autocorrelation is high and when $\boldsymbol{T} = 80$, using the GECM leads to a situation where 95 percent of the parameter estimates are less negative than the actual adjustment parameter. Overall however, the GECM has relatively less variability in parameter estimates, especially when $\boldsymbol{T} = 35$.

Another interesting characteristic of the estimates in Figure 16 is the skewness of the parameter estimates.[24] Nearly every set of simulation results, especially those when the length of the series is $\boldsymbol{T} = 35$, have a left-skew, suggesting that while many estimates are clustered tightly near the actual value of the adjustment parameter, those to the left of the distribution are much more spread out. This may be a factor of the data-generating process; were we to create a cointegrating relationship where the value of the adjustment parameter was -0.75, we might expect a right-skew instead.

[24]Skewness appears to exist since the mean values are not centered on the 95 percent confidence intervals.

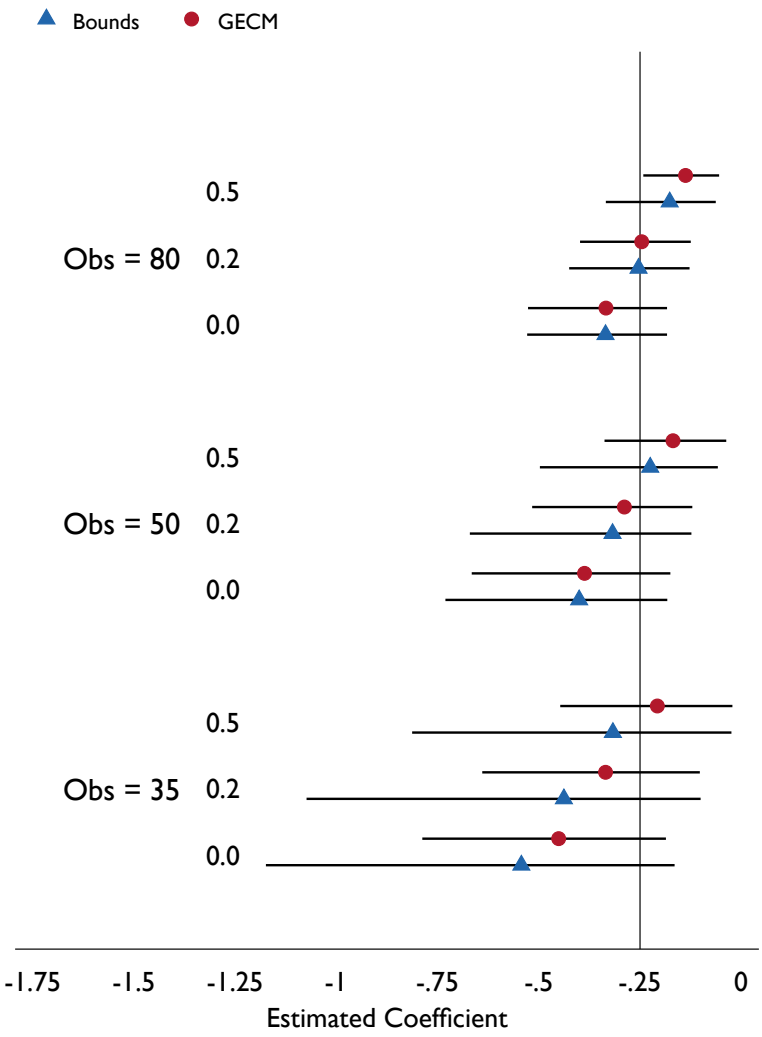

Figure 16: Empirical Distribution of the Adjustment Parameter

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

3.6 Can the ARDL-Bounds Procedure Avoid Spurious Cointegrating Effects?

The section above investigated how well we are able to recover the parameter estimates and effects of interest for a cointegrating data-generating process. How well do the ARDL-bounds and GECM avoid spurious conclusions about parameters and effects of interest when all series are I(1) but unrelated to one another? As discussed in the main paper, the ARDL-bounds model and the GECM are inappropriate to use if all series are I(1) and not cointegrating. Yet deciding whether a series is I(0) or I(1) is extremely difficult in short series. Therefore, it is worth examining how well the ARDL-bounds model avoids making incorrect inferences about the effect of an unrelated series on another, and how this compares to the GECM.

The following data-generating process was used:[25]

$$x_{1t} = x_{1t-1} + \nu_{1t} \tag{13}$$

$$x_{2t} = x_{2t-1} + \nu_{2t} \tag{14}$$

$$u_t = 0.75u_{t-1} + \eta_t + \rho\eta_{t-1} \tag{15}$$

$$y_t = u_t \tag{16}$$

Note that the errors ν_{1t}, ν_{2t} and η_t are independently generated from one another, and y_t is unrelated to the two I(1) regressors, x_{1t} and x_{2t}. Because of this, the contemporaneous effect of the regressors is zero, the coefficient on the lagged regressors

[25] To mitigate issues involving initial conditions (Balke and Fomby 1997), I first created a burn-in period of $T = 100$.

is zero, and the long-run multiplier is zero. Since y_t still depends on past values, the adjustment parameter is -0.25. As with the previous section, there is autocorrelation in the residuals if $\rho \neq 0$. 1000 simulations were conducted for each of the following combinations:

1. Varying the number of observations: $T = 35, 50, 80$.
2. Varying the level of autocorrelation: $\rho = 0.0, 0.2, 0.5$.

Then the following models were run for the ARDL and the ECM models, respectively:

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \sum_{j_1=0}^{q_1} \beta_{j_1} \Delta x_{1t-j_1} + \sum_{j_2=0}^{q_2} \beta_{j_2} \Delta x_{2t-j_2} + \varepsilon_t \tag{17}$$

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \Delta\beta_1 x_{1t} + \Delta\beta_2 x_{2t} + \varepsilon_t \tag{18}$$

Since both x_{kt} series were constructed to have no relationship to y_t, we should only reject the null hypothesis that the short- and long-run effects, as well as the coefficient on the lagged x_{1t}, are equal to zero about five percent of the time.

As shown in Figure 17, both the GECM and the ARDL-bounds methods do a good job at estimating the true effect of Δx_{1t} when it is zero. The rates of rejection of $H_0 = 0$ tend to fall slightly as the length of the series increase. In addition, the GECM more frequently estimates short-run parameters whose constructed 95 percent confidence intervals overlap zero. Finally, residual autocorrelation has only a small effect on coverage rates of the short-run effect; it is not clear whether higher levels

of autocorrelation decrease or increase coverage.

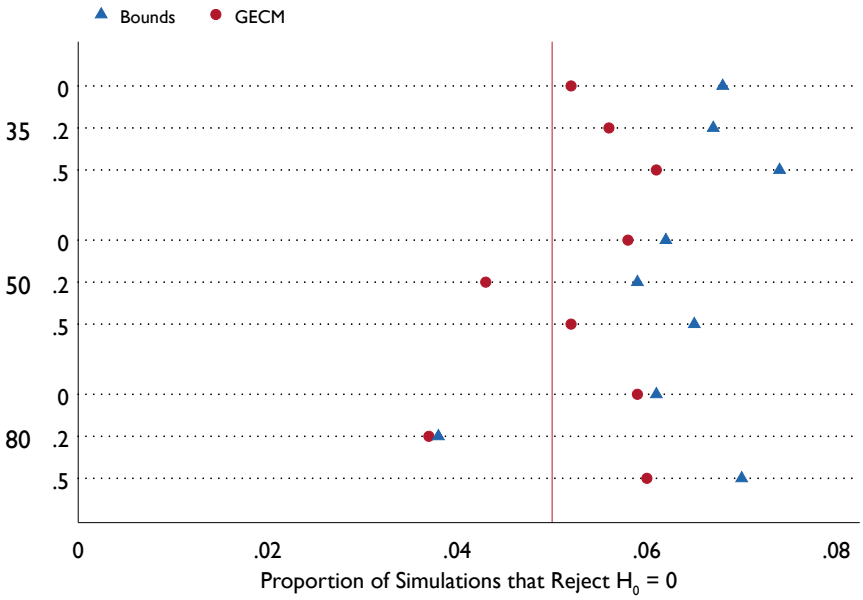

Figure 17: Coverage of the Short-run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

Coverage rates of the long-run effect are shown in Figure 18. Compared to the short-run effect, there are substantially more simulations that incorrectly reject the null hypothesis that the long-run effect is equal to zero. Counterintuitively, the proportion of simulations whose 95 percent confidence intervals do not cover zero tends to increase as the length of the series increase. Moreover, greater levels of autocorrelation tend to make it less likely that either the ARDL-bounds or GECM find spurious evidence of a long-run effect. Overall, these findings suggest that when using all-I(1) series, we are highly susceptible to finding evidence of a long-run effect

when it does not exist. Thus, testing for cointegration is crucial.

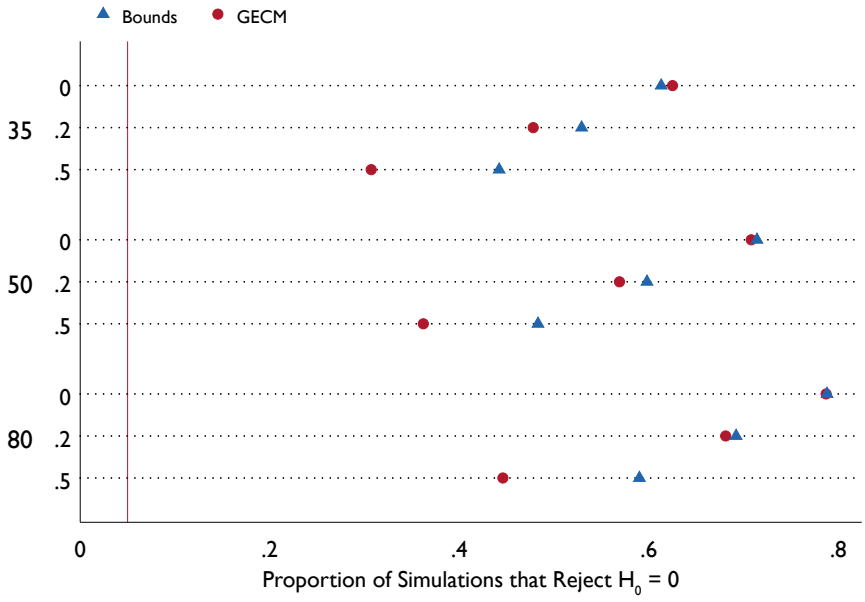

Figure 18: Coverage of the Long-run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In Figure 19, I show the proportion of simulations whose constructed 95 percent confidence intervals of the lagged independent variable do not overlap with zero. Relative to the long-run multiplier, coverage is much better; when the series is small, one finds evidence of a statistically significant (i.e., not equal to zero) coefficient on the lag of x_t about 15 percent of the time. As the series increases to about $T = 80$, this rate drops to less than 10 percent, unless autocorrelation is high.

Last, I examine the coverage rates of the adjustment parameter in Figure 20. Recall that although the two I(1) regressors in the data-generating process were

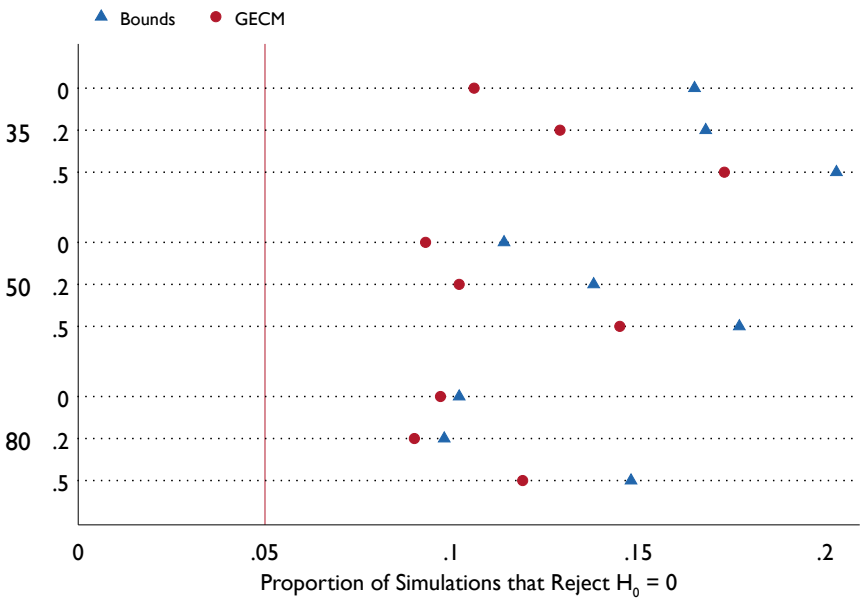

Figure 19: Coverage of the x_{t-1} Parameter

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

constructed so that they are unrelated to y_t, in the resulting GECM and ARDL-bounds models we would still expect an adjustment parameter of -0.25, since current values of y_t are related to its past values. The proportion of simulations whose constructed 95 percent confidence intervals did not include -0.25 are shown in Figure 20.

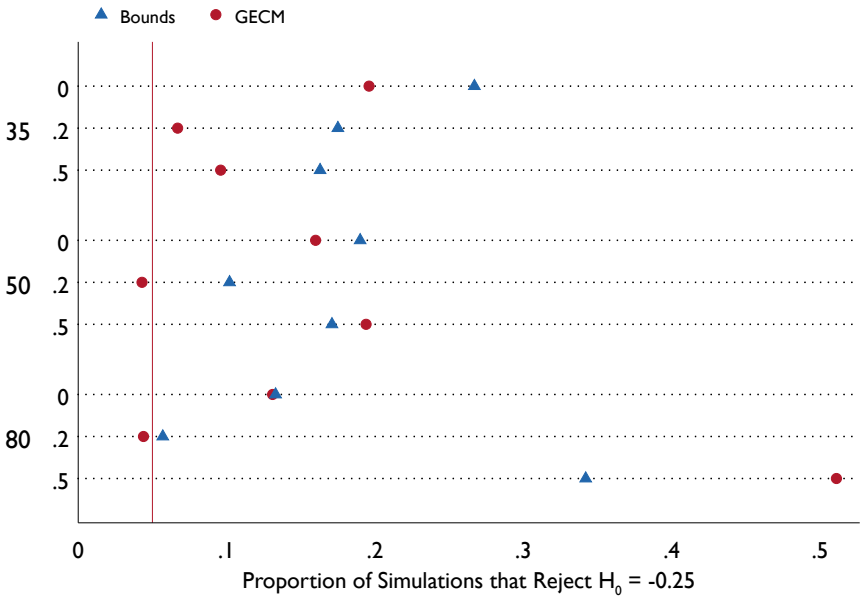

Figure 20: Coverage of the Adjustment Parameter

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

A number of important points stand out. First, when autocorrelation is low, both the GECM and ARDL-bounds models tend to more often model the true adjustment parameter as the number of observations increase. In contrast, when autocorrelation is high, our estimates of the adjustment parameters are less likely to include the

actual estimate at the length of the series increases. In fact, when $T = 80$ and autocorrelation is $\rho = 0.5$, the ARDL-bounds model results in an estimate whose confidence interval does not include the true value of the adjustment parameter about 35 percent of the time. For the GECM, this is over 50 percent. Last, when there were moderate levels of autocorrelation ($\rho = 0.2$), both models, especially the GECM, were most likely to recover the true value of the adjustment parameter.

The empirical distribution of the short-run effect is shown in Figure 21. Recall from the previous section that the empirical distribution describes the spread of the estimated effects using the mean of the 1000 simulations. In addition, the constructed 95 percent confidence intervals (where do 95 percent of the estimates fall?) help show the spread of the estimates. As is clear from Figure 21, the estimated short-run effects are clustered around the true effect size of zero. There are slight increases in variability of the estimates as residual autocorrelation increases, and the ARDL-bounds model leads to a slightly larger spread of empirical estimates when $T = 35$.

In Figure 22, I plot the empirical distribution of the long-run effect. Since this effect is calculated from two estimates (the lagged dependent variable and the coefficient on the lagged independent variable), it is not surprising that the spread of parameter estimates is much greater than for the short-run effect. Overall, the average effect estimate is centered on its true value of zero, except when autocorrelation is high ($\rho = 0.5$) and the series is extremely short ($T = 35$). The variability in estimates tends to be much smaller when autocorrelation is low. Moreover, the ARDL-bounds model tends to have smaller variability than the GECM when autocorrelation is high.

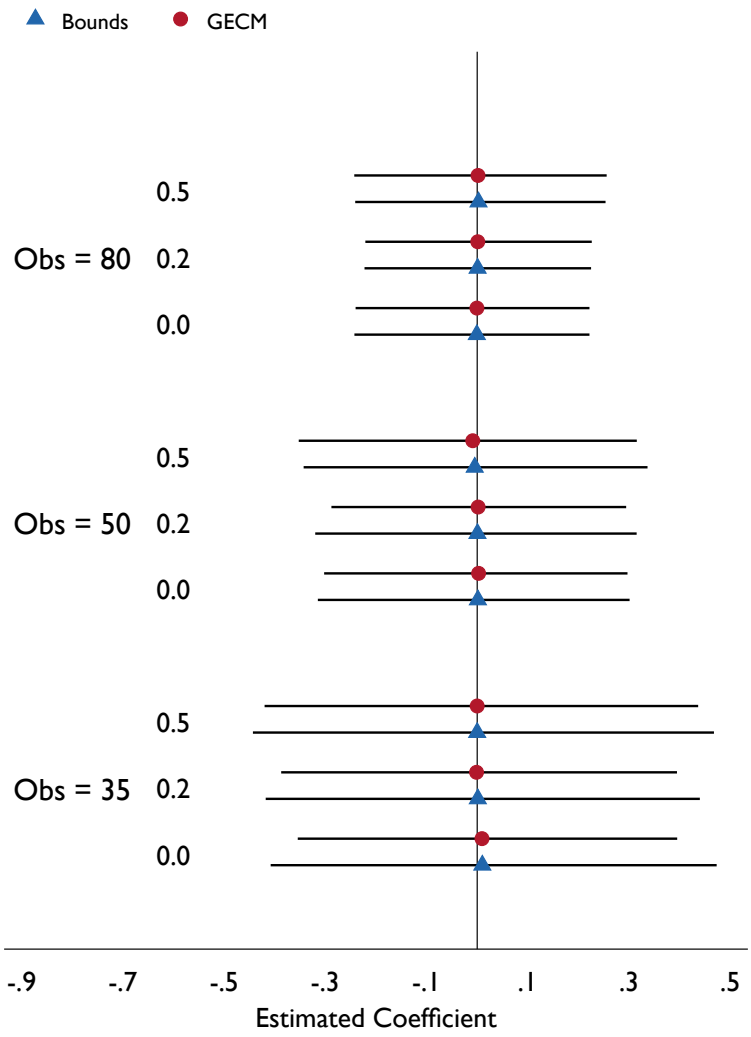

Figure 21: Empirical Distribution of the Short-Run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

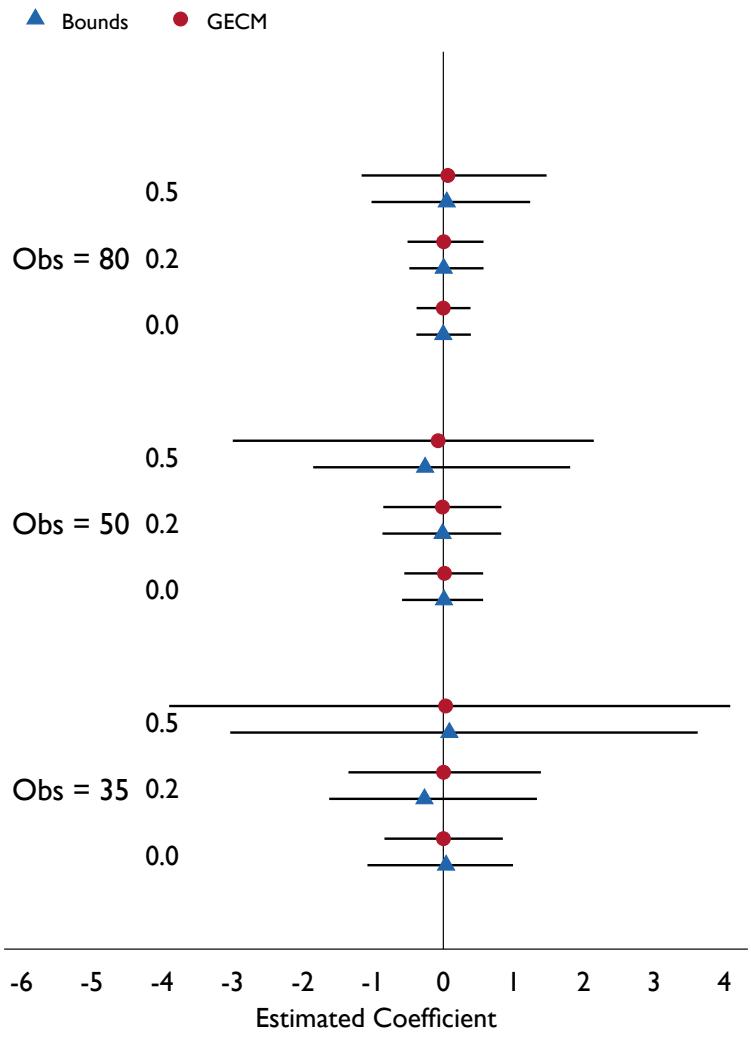

Figure 22: Empirical Distribution of the Long-Run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

The empirical distribution of the estimated coefficient of the lagged independent variable is shown in Figure 23. Nearly all estimated parameters have a mean of about zero. While the variability of estimates remains small when residual autocorrelation is low, there is a marked increase when it increases. For example, when autocorrelation is high ($\rho = 0.5$) and the number of observations small ($T = 35$), 95 percent of the 1000 estimates fell between -4 and 4 for the GECM, and between -3 and 3 for the ARDL-bounds model; when $\rho = 0.0$, this spread falls to around -0.8 and 8, and -1 and 1, respectively.

I last examine the empirical distribution of the adjustment parameter in Figure 24. In this experiment, even though the short-run effect, long-run effect, and coefficient on the lagged independent variable should be zero, the adjustment parameter still has a value of -0.25. As shown by Figure 24, both models tend to estimate parameters that are more negative than -0.25 at low levels of autocorrelation, and actually get closest to the true adjustment parameter only when autocorrelation is $\rho = 0.5$. Overall, estimates tend to cluster around the true value of the adjustment parameter as the length of the series increase, although there is a tendency for the estimates to attenuate towards zero when autocorrelation is high.

To summarize this section, I find that on average, both the GECM and ARDL-bounds model obtain the correct estimate of null effects using an all I(1), spuriously related data-generating process. However, a moderate proportion of estimates tend to indicate a significant effect when there is none, especially for the long-run effect. This underscores the need to test for cointegration before estimating error-correction models with non-stationary data.

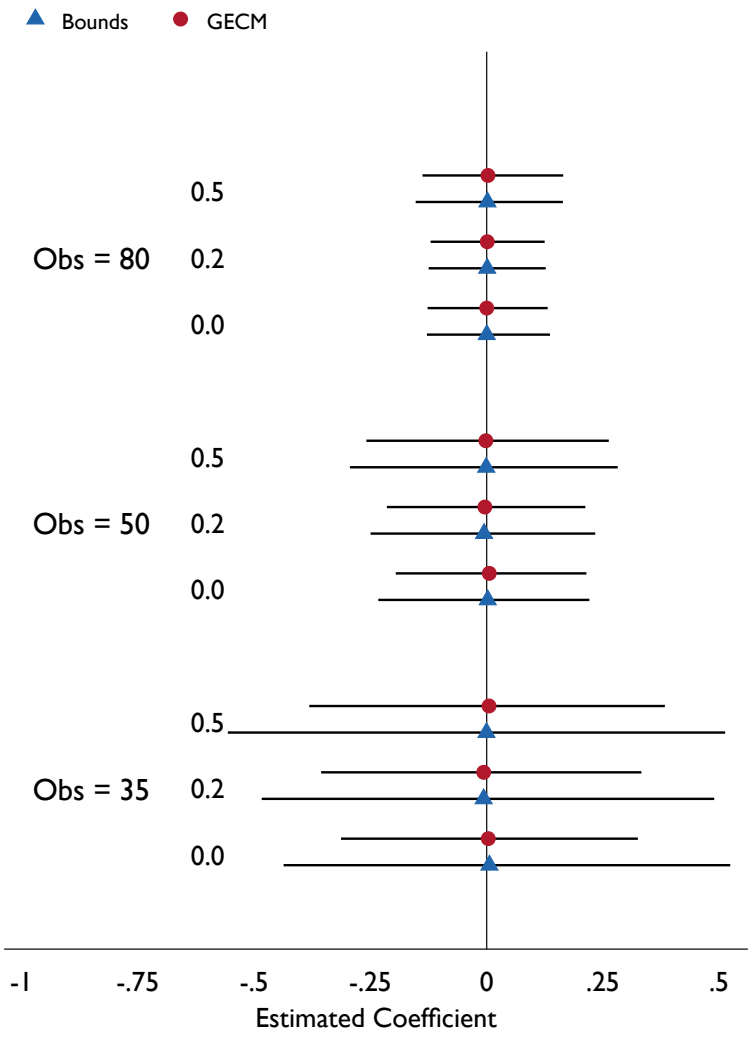

Figure 23: Empirical Distribution of the Lagged Independent Variable

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

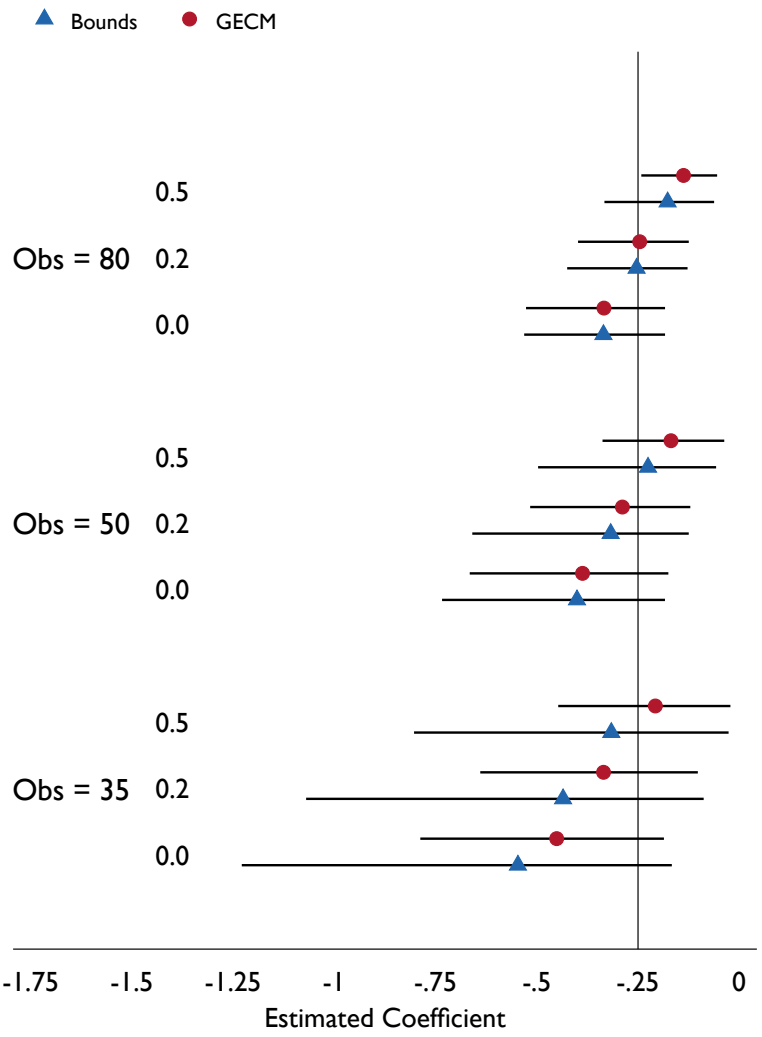

Figure 24: Empirical Distribution of the Adjustment Parameter

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

3.7 How Well Can the ARDL Procedure Recover Stationary Relationships?

In the sections above I investigated the performance of the ARDL model at recovering effect sizes and parameter estimates under two scenarios: when the series were all I(1) and cointegrating, and when the series were all I(1) and unrelated to one another. In this section I examine how well the ARDL-bounds model can recover effects of interest from a stationary, autoregressive relationship. As noted by many scholars (De Boef and Keele 2008; Keele and Kelly 2006), the general error-correction model can be used in the all-I(0) case—although the standard lagged-dependent variable model is a simpler alternative (Enns et al. 2016).[26] Therefore, in this section I examine the performance of the ARDL-bounds model as well as the GECM.

Since the ARDL modeling procedure in the main paper simply involves estimating a standard error-correction model (the first-difference of the dependent variable is regressed on its lag and the lag and first difference of the independent variables) with additional lagged first differences to remove residual autocorrelation, *a priori*, I expect the recovery rate of relationships to be very similar to the standard GECM. However, since models often contain some amount of mis-specification, particularly in short samples, I investigate the consequences of varying amounts of autocorrelation

[26]Moreover, the GECM requires at least two parameters for each regressor (i.e., Δx_t and x_{t-1} in order to recover short- and long-run effects, while the lagged-dependent variable model requires only one (typically x_t).

in the residuals. I used the following data-generating process:[27]

$$x_{1t} = 0.5x_{1t-1} + \nu_{1t} \tag{19}$$

$$x_{2t} = 0.5x_{2t-1} + \nu_{2t} \tag{20}$$

$$y_t = 0.5y_{t-1} + 2x_{1t} + 2x_{2t} + \varepsilon_t + \rho\varepsilon_{t-1} \tag{21}$$

where the errors ν_{1t}, ν_{2t}, and ε_t are independently generated from one another. Note also that there exists autocorrelation in the error term for the data-generating process of y_t, as long at $\rho \neq 0$. I then ran the following models for the ARDL-bounds and the GECM, respectively:

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \sum_{j_1=0}^{q_1} \beta_{j_1} \Delta x_{1t-j_1} + \sum_{j_2=0}^{q_2} \beta_{j_2} \Delta x_{2t-j_2} + \varepsilon_t \tag{22}$$

$$\Delta y_t = \alpha_0 + \theta_0 y_{t-1} + \theta_1 x_{1t-1} + \theta_2 x_{2t-1} + \Delta\beta_1 x_{1t} + \Delta\beta_2 x_{2t} + \varepsilon_t \tag{23}$$

Augmenting lags of the first-difference of x_{1t}, x_{2t}, and y_t were determined by SBIC for the bounds-ARDL model.[28] I then tested how well both the ARDL-bounds model and the GECM recovered the short-run effect of $\beta_1 = 2$ for x_{1t}, the long-run effect, $\frac{2}{(1-0.5)} = 4$, the coefficient on the lagged regressor, $\theta_1 = 2$, and the adjustment parameter, which should have a coefficient of $\theta_0 = (0.5-1) = -0.5$ in the GECM and

[27] To mitigate issues involving initial conditions (Balke and Fomby 1997), I first created a burn-in period of $T = 100$.

[28] A maximum lag restriction of $p, q_1, q_2 \leq 4$ was used for $T = 50, 80$ and a restriction of $p, q \leq 3$ for $T = 35$.

ARDL-bounds models.[29] Since x_{1t} and x_{2t} had identical data-generating processes, I only examine x_{1t} below. I generated 1000 simulations across each of the following combinations:

- Varying the number of observations: $T = 35, 50, 80$.
- Varying the level of autocorrelation: $\rho = 0.0, 0.2, 0.5$.

As with the cointegrating Monte Carlo experiment, I calculated coverage probabilities for the stationary relationships simulated in this section. This series of figures helps to answer whether or not we get our substantive hypotheses correct (see Figure 8 for a graphical depiction of this); if a high proportion of simulations find that—after constructing 95 percent confidence intervals, that the true parameter lies within the estimated interval—then we can be relatively confident that these models are able to recover the actual parameter estimates of the underlying process we seek to model. If instead, we find that the constructed intervals are wildly off the mark, it is less likely that we will get our substantive hypotheses correct.[30]

The results for the short-run effects are shown in Figure 25. The horizontal axis shows the proportion of simulations whose constructed 95 percent confidence intervals *did not* contain the parameter value of the data-generating process (2.0 in this case); thus, lower values indicate the method is more likely to recover the true effect size. As shown in Figure 25, only a small proportion of simulations have

[29] On the equivalence and interpretation of the ADL and GECM, see De Boef and Keele (2008).

[30] By hypotheses, I mean both direction (is the effect positive or negative) and magnitude (the size of the effect). This section does not address cases where we only get one of these two components correct.

constructed coverage probabilities that do not cover the parameter from the data-generating process. The GECM appears to be slightly better at recovering the short-run effect sizes, but this difference becomes negligible as the sample size increases. As expected, an increase in autocorrelation tends to lead to poorer coverage.

Figure 25: Coverage of the Short-Run Effect

Note: Dot plot shows the proportion of the time that the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

I examine the coverage probabilities of the long-run effect in Figure 26. Neither the ARDL-bounds or GECM have constructed coverage probabilities that tend to include the long-run effect at conventional rates. Interestingly, coverage tends to be better in both models as autocorrelation *increases*. Surprisingly, coverage does not seem to improve, and in fact gets worse, as the number of observations increase. Last,

the ARDL-bounds model appears to outperform the GECM in terms of coverage; this indicates that users are more likely to recover the actual effect of interest when using the ARDL-bounds.

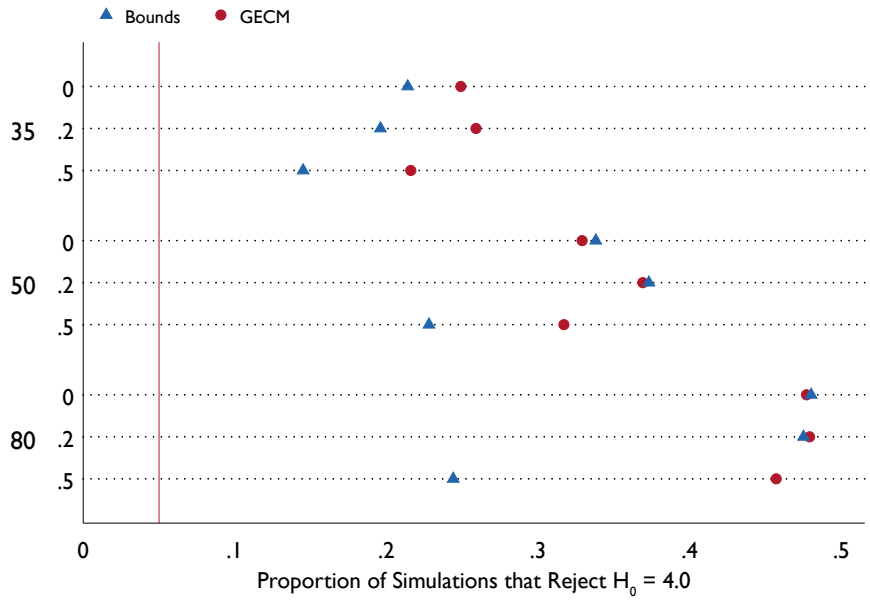

Figure 26: Coverage of the Long-Run Effect

Note: Dot plot shows the proportion of the time that the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In Figure 28, I examine the coverage of the adjustment parameter. Recall that since we created a stationary data-generating process where the value of the lagged dependent variable was 0.5, this DGP would yield an adjustment parameter of -0.50 in the GECM and bounds model. It appears as though the largest determinant of coverage for the adjustment parameter is the amount of residual autocorrelation. Although coverage improves as sample size increases when there is little to no auto-

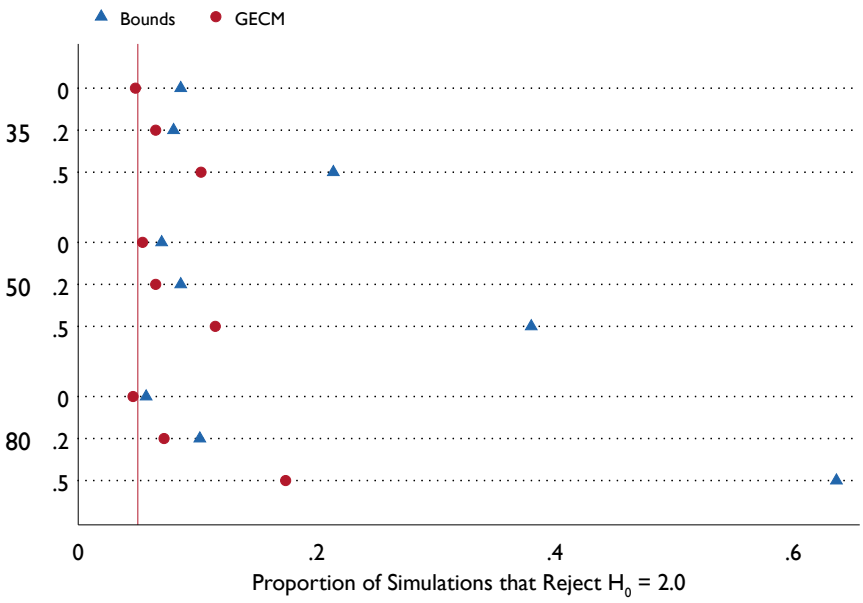

Figure 27: Coverage of the x_{t-1} Parameter

Note: Dot plot shows the proportion of the time that the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

correlation (i.e. $\rho = 0.0, 0.2$), when autocorrelation is equal to $\rho = 0.5$, the calculated coverage probability of the adjustment parameter is *less* likely to contain the DGP parameter as sample size increases. This holds for both the bounds and GECM models. Therefore, it appears as though residual autocorrelation can be problematic for obtaining the correct adjustment parameter, especially as the number of observations increases. In addition, as the number of observations increase, the bounds test appears to perform slightly better than the GECM if autocorrelation is high.

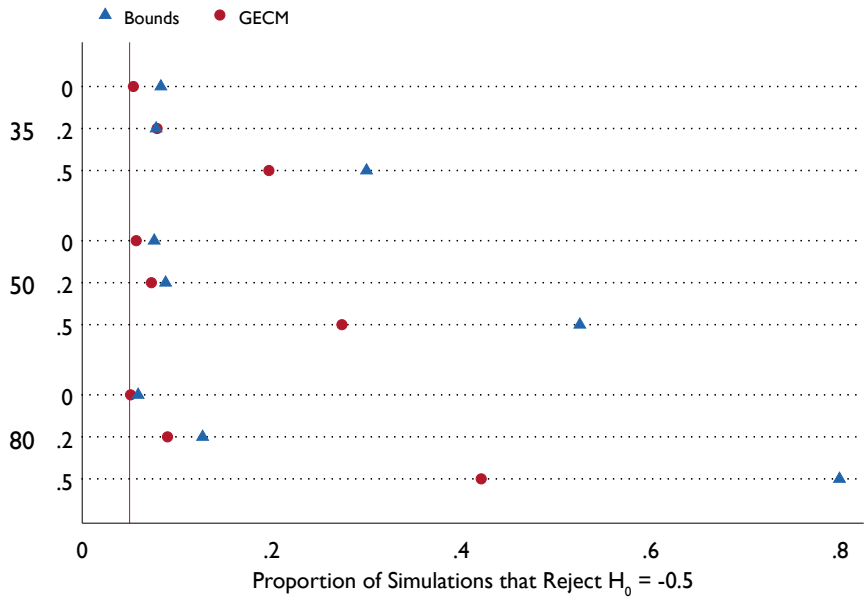

Figure 28: Coverage of the Adjustment Parameter

Note: Dot plot shows the proportion of the time that the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In addition to examining coverage probabilities of the stationary DGP, I also examine the empirical distribution of each of the parameters or effects. This is

shown for the short-run effect in Figure 29. As with all other empirical distributions of the short-run effect, this one shows that all parameter estimates are relatively tightly clustered around the actual short-run effect size of 2.0. There is very little difference in estimated parameter variability between the two models.

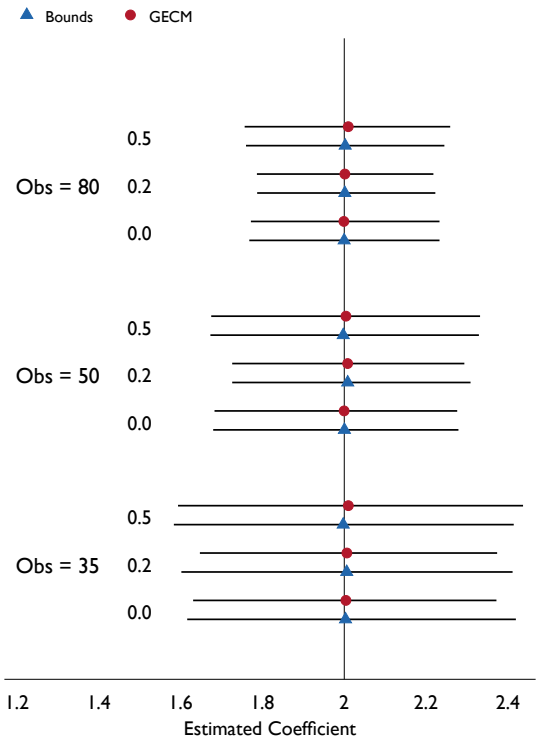

Figure 29: Empirical Distribution of the Short-run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In Figure 30 I plot the empirical distribution of the long-run effect. Overall, the GECM has a tighter spread of parameter estimates. Mean estimates always tend to be the same across both models, except when autocorrelation is high; in these

instances, the ARDL-Bounds model tends to lead to estimates that are attenuated towards zero. Last, as seen for the other effects, the empirical distribution tends to concentrate towards the true value of four as the length of the series increase.

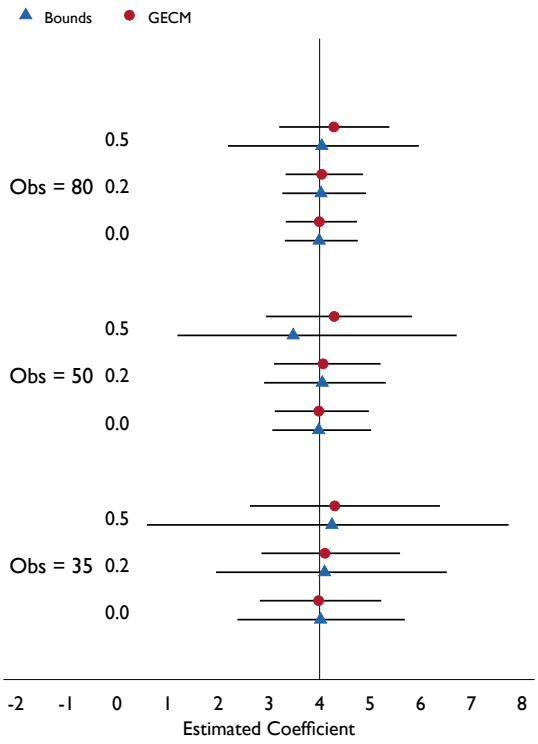

Figure 30: Empirical Distribution of the Long-run Effect

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

The empirical distribution of the estimated parameter on the lagged independent variable is shown in Figure 31. As with the long-run effect, differences between the two models—in terms of average parameter estimates—tend to appear only when autocorrelation is high. Unlike the long-run effect, average parameter estimates

tend to attenuate towards zero as the length of the series increase, but only when autocorrelation is high.[31] Overall, the GECM tends to have parameter estimates more concentrated around the actual value of the lagged independent parameter.

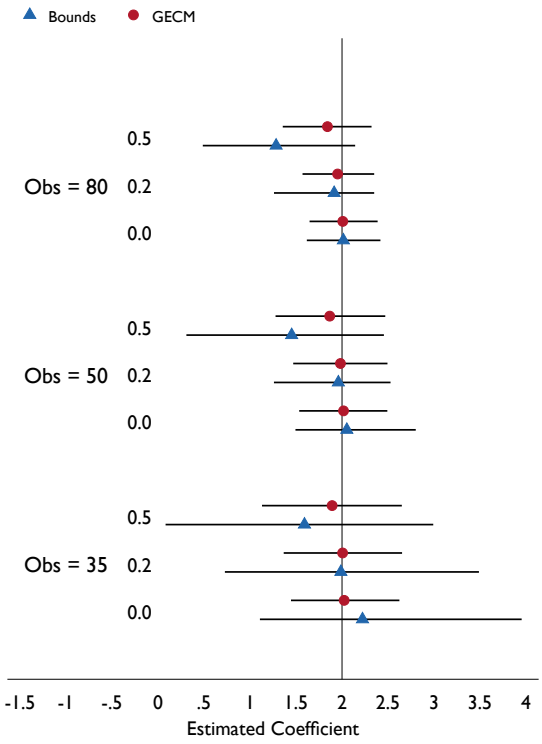

Figure 31: Empirical Distribution of the Lagged Independent Variable Parameter

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

[31] As shown in Figure 31, this may be because the estimated coefficient on the adjustment parameter tends to attenuate towards zero (from the negative side) when autocorrelation is high. Interestingly, since both parameter estimates are used to calculate the long-run effect in Figure 30 (which did not have increased bias as the length of the series increased under high autocorrelation, and since both estimates are moving towards zero, the two essentially cancel each other out, leading to very little bias shown in Figure 30.)

Last, I examine the empirical distribution of the adjustment parameter in Figure 32. As with the lagged independent variable estimates, estimates of the adjustment parameter tend to attenuate towards zero when autocorrelation is high; this gets *worse* as the length of the series increase. Overall however, at low levels of autocorrelation, both the ARDL-bounds and GECM are quite good at providing estimates near the true value of the adjustment parameter.

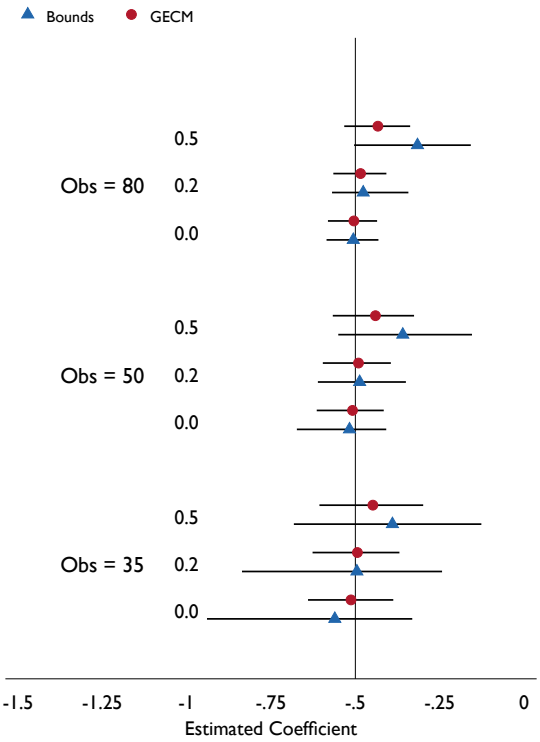

Figure 32: Empirical Distribution of the Adjustment Parameter

Note: Plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

There are a number of important findings in regards to the use of the bounds

and GECM models to model a stationary DGP.[32] As with the cointegrating case (see Section 2.4), I find that the bounds test tends to reduce bias to a greater extent than the GECM, while the GECM tends to have calculated coverage probabilities that include the parameters of the DGP more often than the ARDL-bounds model. Coverage appears to be quite good for the short-run and long-run effects across both models, and less so for the adjustment parameter when there is substantial residual autocorrelation. Bias appears to be especially large only for the long-run effect. Also interesting is that there appear to be no large changes as sample size increases; coverage and bias is not poor in short samples, and even may lead to better coverage probabilities in the face of high autocorrelation, as seen with the adjustment parameter (see Figure 28).

3.8 Can the ARDL Procedure Avoid Spurious Stationary Relationships?

In the previous section I investigated the performance of the ARDL-bounds model and the GECM when modeling stationary series. In this section, I explore how well each model avoids spurious conclusions when there is *no* relationship between the dependent and independent variables. The data-generating process is the same as in

[32] As with Section 2.4, these findings may hold only for the case of a single, weakly exogenous regressor. As I found in the main paper, the performance of models and test statistics appears to vary substantially as the number of regressors change; tests that perform well under a single regressor may not perform well when using three or four regressors.

the above example, except now the two regressors are unrelated to y_t:

$$x_{1t} = 0.5x_{1t-1} + \nu_{1t} \tag{24}$$

$$x_{2t} = 0.5x_{2t-1} + \nu_{2t} \tag{25}$$

$$y_t = 0.5y_{t-1} + \varepsilon_t + \rho\varepsilon_{t-1} \tag{26}$$

Since neither of the stationary independent variables are related to the dependent variable, the short-run and long-run effects, as well as the coefficient on the lagged independent variable, should be zero. Since past values of y_t are related to current ones, we should still expect to recover an adjustment parameter of -0.5.

I first examine coverage rates. The coverage rate of the short-run effect is shown in Figure 36. Since the DGP specified that x_{1t} is unrelated to y_t, the figure shows the proportion of simulations whose 95 percent confidence intervals *did not include zero*. Across different lengths of the series, and varying levels of autocorrelation, we find evidence of an effect that is statistically significantly different from zero between four and seven percent of the time. Surprisingly, there appears to be no relationship between the length of the series, or the amount of autocorrelation, and the coverage rates. Overall, the GECM is slightly better at estimating parameters whose confidence intervals correctly include zero.

In Figure 34 I examine the coverage rates of the long-run effect of a spuriously-related I(0) series. Counterintuitively, we are less likely to conclude evidence of a long-run effect (when the true DGP is a spurious relationship), in shorter series and under high levels of residual autocorrelation; for $T = 80$, moving from no autocorre-

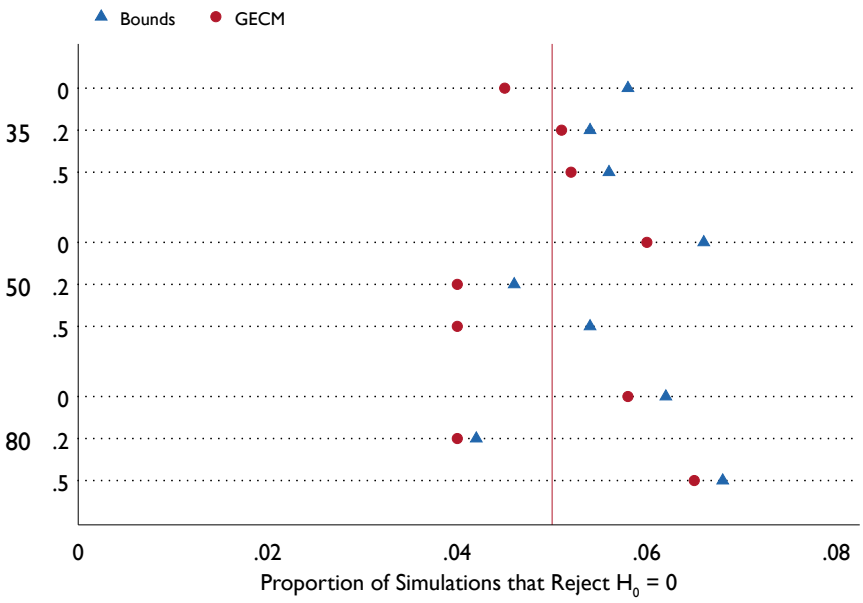

Figure 33: Coverage of the Short-run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

lation ($\rho = 0.0$) to high correlation ($\rho = 0.5$) decreases the proportion of simulations that reject the (true) null hypothesis that the long-run effect equals zero from about 50 percent of the time to just over 10 percent. Coverage rates between the two models are generally similar.

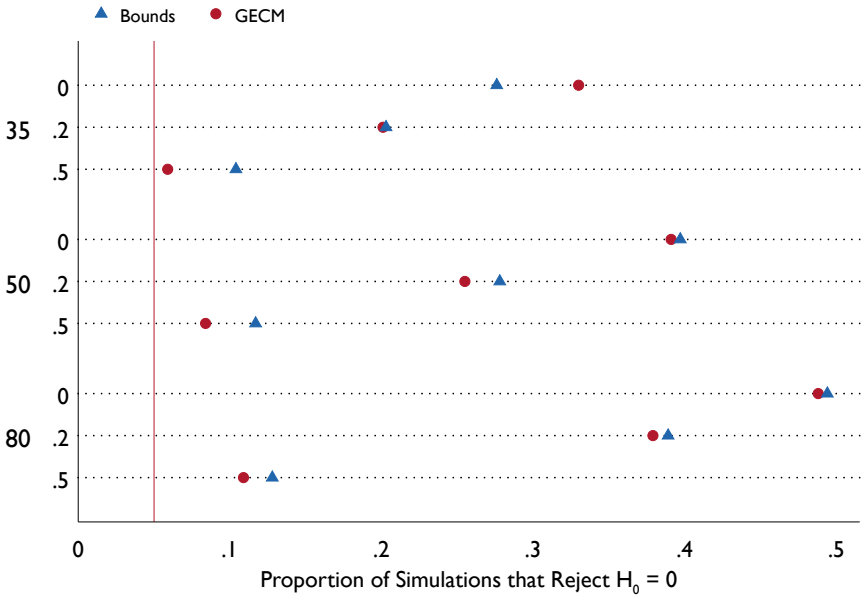

Figure 34: Coverage of the Long-run Effect

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

While coverage rates of the long-run effect were quite high in the case of spuriously-related, I(0) series, the coverage rates of the parameter estimate on the lagged independent variable are much closer to conventional levels, as shown in Figure 35. In general, lower levels of autocorrelation are associated with coverage rates closer to conventional levels of acceptance. In addition, increasing the length of the series

leads to much closer coverage rates between the ARDL-bounds and GECM.

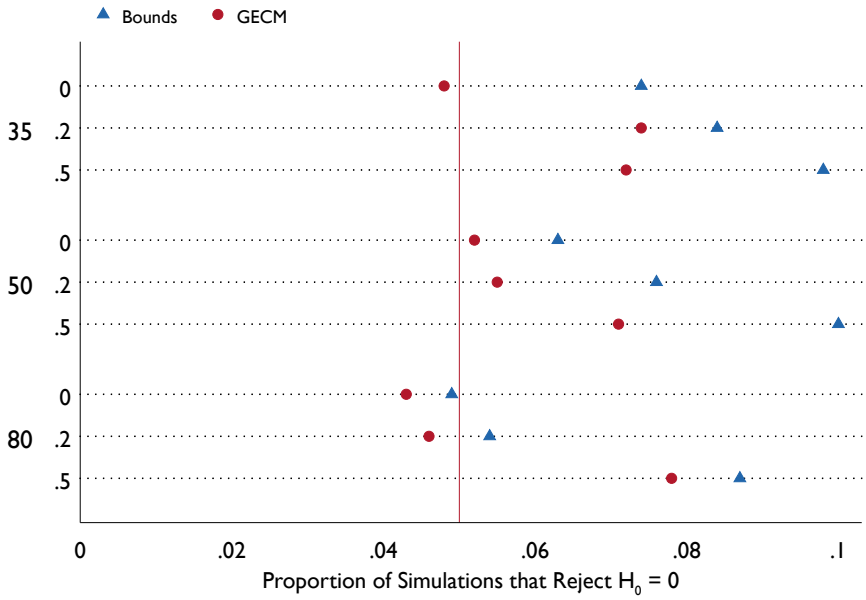

Figure 35: Coverage of the x_{t-1} Parameter

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In Figure 36, I show the coverage rate of the adjustment parameter. Recall that even though the independent variables are unrelated to y_t, we still expect the coefficient on the adjustment parameter to be $(0.5 - 1) = -0.5$. As is clear from the figure, a large number of simulations result in 95 percent confidence intervals that do not include -0.5. When autcorrelation is low, the GECM tends to perform better than the ARDL-bounds, but only when the length of the series is large ($T = 80$). In contrast, if there is high autocorrelation in sizable series (again $T = 80$), the ARDL-bounds performs much better than the GECM. Overall, both models tend to

perform better in short series, and when autocorrelation is low. In sum, this suggests that when autocorrelation and the length of the series is large, the stationary (and unrelated) properties of the regressors may make it hard to estimate the correct value of the adjustment parameter.

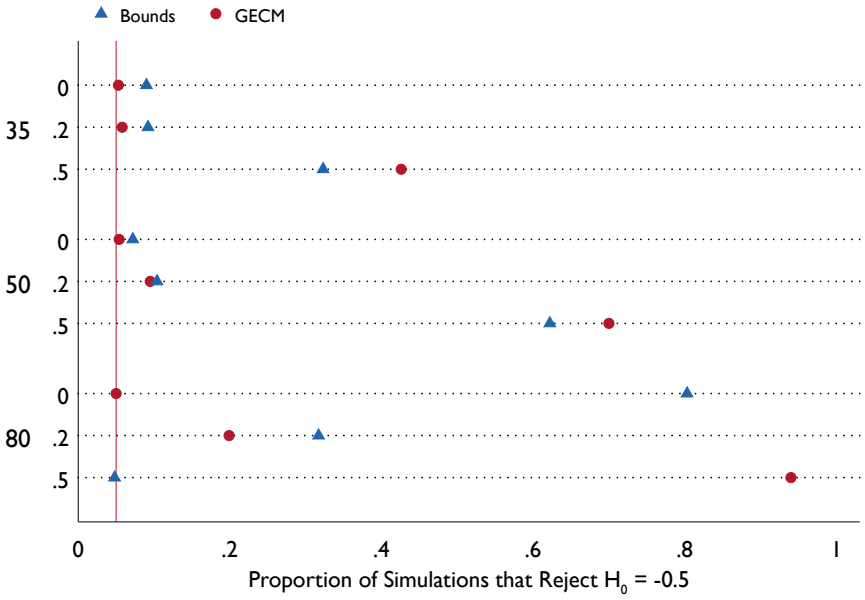

Figure 36: Coverage of the Adjustment Parameter

Note: Dot plot shows the proportion of simulations where the DGP parameter fell outside of the calculated 95 percent confidence confidence interval for a simulation, across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

As with all other Monte Carlo experiments in this section, I also investigate the empirical distribution of the estimates to get a sense of how clustered the estimates are, and if the average estimate lies near the DGP value. The first of these is shown in Figure 37, which is of the short-run empirical distribution of x_t. As has been clear from all simulation results so far, both the ARDL-bounds and GECM are quite good

at estimating the short-run effect under a variety of conditions. Even under short series and high autocorrelation, both models result in parameter estimates that are centered on zero and 95 percent of estimates range from -.4 to .4.

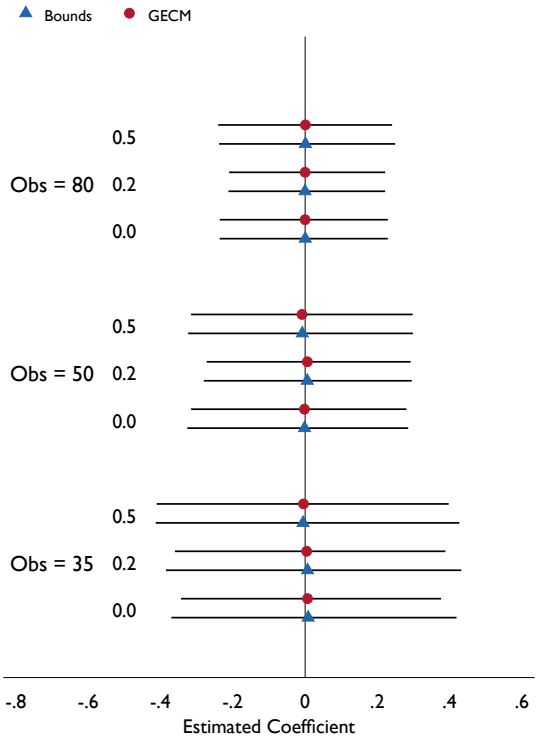

Figure 37: Empirical Distribution of the Short-Run Effect

Note: Dot plot shows the average estimated from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

Unlike previous results, the empirical distribution of the long-run effect under a stationary, yet unrelated process, tends to center around the actual effect of zero, as shown in Figure 38. While the variability of parameter estimates tends to increase when autocorrelation is very high ($\rho = 0.5$), this only seems to be a problem in short

series; when $T = 80$ and $\rho = 0.5$, 95 percent of parameter estimates lie between -2 and 2 for both models.

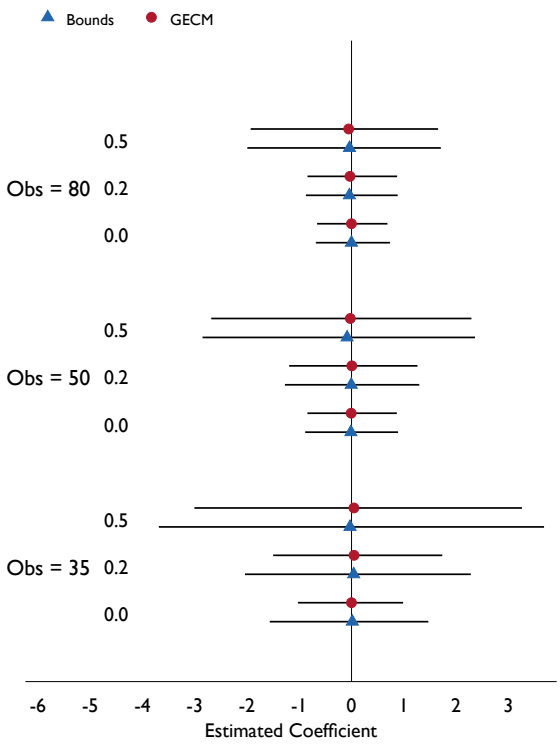

Figure 38: Empirical Distribution of the Long-Run Effect

Note: Dot plot shows the average estimated from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In Figure 39 I show the empirical distribution of the lagged coefficient on x_t. Although parameter estimates are slightly larger under the ARDL-bounds in short series, for the most part both models tend to accurately find that the true DGP is zero, given enough repeated samples. In addition, estimates tend to improve as the length of the series increases.

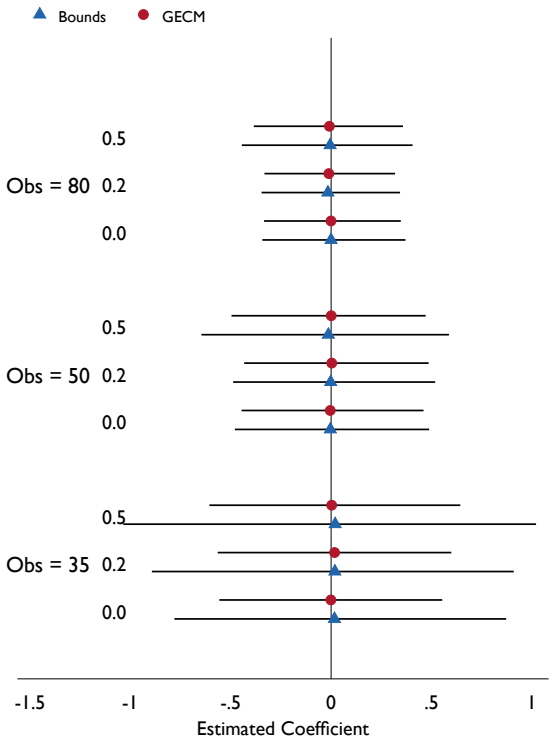

Figure 39: Empirical Distribution of the x_{t-1} Parameter

Note: Dot plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

Last, I show the empirical distribution of the adjustment parameter in Figure 40. While all other empirical distributions for a spuriously-related I(0) process were centered around zero, the adjustment parameter is much different. As is clear based on the other simulations in this section, adjustment parameters tend to slightly overestimate coefficients (i.e., make them more negative) when autocorrelation is low; estimates tend to attenuate towards zero when autocorrelation is high. In the case of high autocorrelation and long series ($\boldsymbol{T} = 80$), 95 percent of the simulated estimates do not fall anywhere near the actual adjustment parameter of -0.5. This helps explain the surprising lack of coverage in Figure 36 for the GECM; since the ARDL-bounds model tends to have slightly wider confidence intervals, the skewed empirical distribution of the adjustment parameter means that very few of the GECM estimates have confidence intervals that overlap with -0.5.

3.9 A More Conservative Assessment of Type I Error for the Cointegration Tests

In the main paper, I investigated rates of Type I error across four different cointegration tests in the first Monte Carlo experiment. The bounds test was treated as having avoided a spurious conclusion of cointegration if the resulting test statistic was below the upper I(1) critical value. This was justified since—by using this cut-off point—the bounds test provides a result that can be interpreted just like the other cointegration tests: an F-statistic above the I(1) critical value suggests that all regressors appear to be cointegrated with the dependent variable.

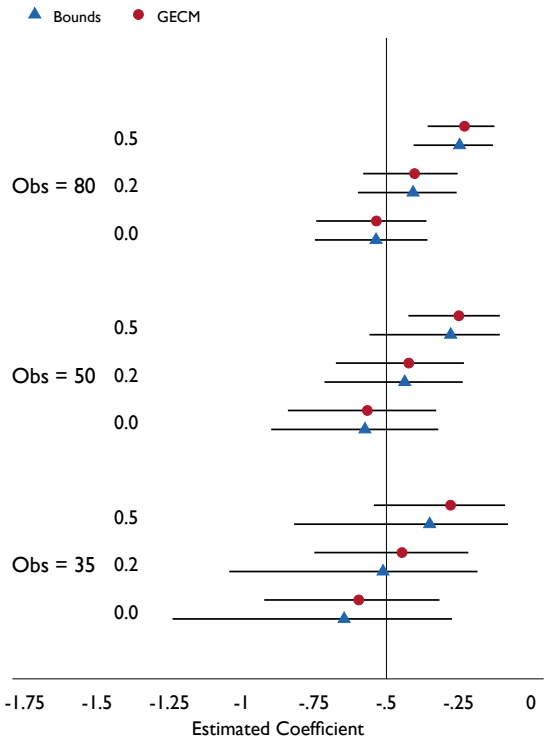

Figure 40: Empirical Distribution of the Adjustment Parameter

Note: Dot plot shows the average parameter estimate from 1000 simulations—along with 95 percent confidence intervals calculated using the percentile method—across the number of observations ($T = 35, 50, 80$) and level of AR(1) ($\rho = 0.0, 0.2, 0.5$).

In this section I use the same data from the Monte Carlo simulation in the main paper, but this time I treat an inconclusive result (i.e., if the test statistic falls *between* the I(1) and I(0) critical values) as another form of Type I error. How might an indeterminate test result be considered a form of Type I error? Consider the following example; four unrelated regressors are included in an ARDL model, and the user gets an indeterminate F-statistic result using the bounds test. The user then restricts one of the series from appearing in levels (i.e., imposes the restriction that this regressor cannot cointegrate with the dependent variable), and the resulting F-statistic now falls above the I(1) critical value. Thus, by treating indeterminate results as a potential form of Type I error, we can establish a very high standard for the bounds test to pass.

The results using the new critical values are shown in Figure 41. Recall from the main paper that this data-generating process involved creating an I(1) dependent variable, y_t, and four independent variables, x_{kt} (where $k = 1,2,3,4$), for series of length $T = 35, 50, 80$. The autoregressive process for x_{1t} was varied by $\phi_1 = (0.0(.20)1.0)$, while all of the other x_{kt} regressors (if present in the model), were I(1). I then used the bounds testing procedure, and compared this with the results of the Engle-Granger and Johansen (BIC and rank) cointegration tests. In Figure 41, the proportion of bounds test statistics falling above the I(1) critical value—or falling between the I(1) and I(0) thresholds—are shown as a black line. For reference, the original bounds test results (that is, only treating a test statistic above the I(1) critical value as Type I error) are shown as blue dots. Not surprisingly, the proportion of simulations falsely detecting cointegration using the I(1) and indeterminate cut-off

points increase. Yet this increase is not uniform across the length of the series or the number of k regressors. For instance, when there are only one or two regressors, the rates of Type I error for the bounds test are nearly identical to the original bounds test results in the main paper. Yet for three, and especially under four regressors, the rates of Type I error for the bounds test increase substantially when using the expanded definition of Type I error. However, it appears that by increasing the length of the series, this difference appears to shrink. Note also how if we count critical values that are indeterminate or exceeding the I(1) value, the rate of Type I error stays constant across the level of autoregression in the single series, x_{1t}. This is similar to the earlier findings in the main paper.

In addition to comparing the rates of Type I error under the different cut-off points for the bounds test, it also helps to compare how the new results compare to the three other cointegration tests. For series of length $T = 35$, the bounds test remains the best choice for minimizing the Type I error rate; although the Engle-Granger test performs similarly for a single regressor, the Johansen test performs at a similar rate as the bounds test when there are four regressors in the model. When $T = 50$, the bounds test remains the best choice for minimizing Type I error, except in the case of four regressors (the Johansen BIC test outperforms the bounds test when autocorrelation is high). When $T = 80$, the bounds test has the lowest rate of Type I error across all scenarios, except when there are many regressors (3 or 4) and when the level of autoregression in the x_{1t} regressor is high; in this case, the Johansen BIC test outperforms all other tests in terms of Type I error. Overall, the bounds test still appears to be the best choice for minimizing false positives.

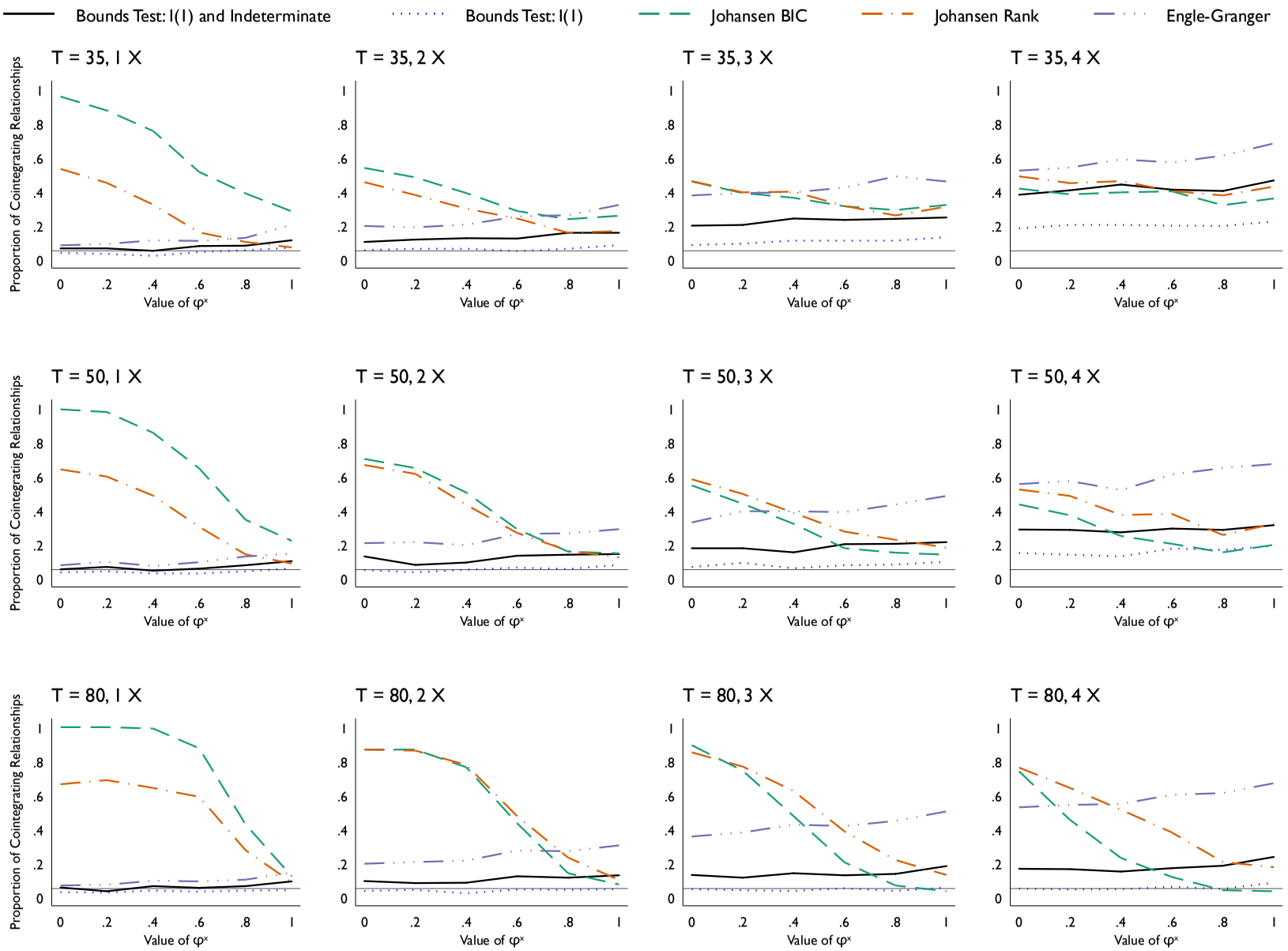

Figure 41: Proportion of Monte Carlo Simulations (falsely) Detecting Cointegration Across Various Methods: I(1) Cut-off Point Vs. I(1) and Indeterminate

Note: Each plot shows the proportion of simulations finding (at $p < 0.05$) evidence of one cointegrating relationship with up to k regressors and T observations across varying amounts of autoregression in x_{1t}, using each of the four cointegration testing procedures. In the main paper, the I(1) critical value is used as the cut-point for idenfitying spurious regression using the bounds test; the I(0) critical value is used here.

4 Proof of the Equivalence of the Triangular Error-Correction Representation to the Standard Representation

In the main paper the data-generation process (DGP) for the second Monte Carlo experiment was given by

$$x_{kt} = x_{kt-1} + \nu_{kt} \tag{27}$$

$$u_t = 0.75u_{t-1} + \eta_t \tag{28}$$

$$y_t = 0.25x_{1t} + \cdot + 0.25x_{kt} + u_t \tag{29}$$

which is also known as the triangular system error-correction representation (Phillips 1991). The DGP was specified in this way for convenience. Since this is less-commonly seen than the vector error correction model representation of Johansen (1991)—which, in turn, is derived from the Granger representation theorem (Engle and Granger 1987), although one can be derived from the other (Cappuccio and Lubian 1996)—I show that this DGP is equivalent to standard representations that can be estimated using a one-step error correction model, such as the one shown in De Boef and Keele (2008).[33]

First, consider the three data-generating processes below. Without loss of gener-

[33]Readers interested in the full analytical derivation of the VECM/MA-AR representation from the triangular representation should consult Cappuccio and Lubian (1996).

ality, assume a single regressor, x_t, that is weakly exogenous:

$$x_t = x_{t-1} + \nu_t \tag{30}$$

$$u_t = 0.75u_{t-1} + \eta_t \tag{31}$$

$$y_t = 0.25x_t + u_t \tag{32}$$

Let ν_t and η_t be i.i.d. and independent from one another. Next, subtract y_{t-1} from either side:

$$\Delta y_t = y_t - y_{t-1} = 0.25x_t + u_t - y_{t-1} \tag{33}$$

Since $y_{t-1} = 0.25x_{t-1} + u_{t-1}$, and since $u_t = 0.75u_{t-1} + \eta_t$, Equation 33 can be rewritten as:

$$\Delta y_t = 0.25\Delta x_t + (0.75 - 1)u_{t-1} + \eta_t \tag{34}$$

and since $u_{t-1} = y_t - 0.25x_t$, Equation 34 becomes:

$$\Delta y_t = 0.25\Delta x_t - 0.25(y_{t-1} - 0.25x_{t-1}) + \eta_t \tag{35}$$

which if estimated using a one-step error-correction model like the one given in De Boef and Keele (2008), would look like:

$$\Delta y_t = \alpha_0 + \alpha_1^* y_{t-1} + \beta_0 \Delta x_t + \beta_1^* x_{t-1} + \eta_t \tag{36}$$

where $\alpha_0 = 0$ (since no constant was included in the DGP), the adjustment parameter is given as $\alpha_1^* = -0.25$, the contemporaneous parameter on x_t is $\beta_0 = 0.25$, and the parameter on the lagged independent variable is $\beta_1^* = 0.0625$.[34] In the Monte Carlo experiment, up to $k = 4$ independent variables were generated, so each has a long-run multiplier of 0.25 by construction.

5 Three Replications

5.1 Replication I: Kelly and Enns (2010)

Unit root tests of the first-difference of the regressors (needed to ensure that no series is greater than I(1)), indicated that first differencing rendered each series stationary. For the first-difference of Policy Liberalism, an augmented Dickey-Fuller test with one lag on 34 observations yielded a test statistic of $Z(t) = -3.22$, which was able to reject the null hypothesis at the 0.05 level. For the first-difference of Income Inequality, the same test yielded a test statistic of $Z(t) = -6.73$, which was statistically significant at the 0.001 level or better.

Since the initial results from the ARDL model and associated bounds test in the main paper were inconclusive, the next step was to conduct unit root testing on the regressors. The results are shown in Table 3. As stated in the main paper, it appears as though all regressors are non-stationary.

[34] This is obtained since the coefficient on the lagged independent variable is the long-run multiplier times the adjustment parameter, $\beta_1^* = \kappa_1 \alpha_1^* = 0.0625$.

Table 3: Unit Root Test Statistics for the Regressors in Kelly and Enns (2010)

Unit-Root Test	Policy Liberalism	Income Inequality
Augmented Dickey-Fuller (with drift)	0.84	-0.44
Phillips-Perron	1.10	-0.56
Dickey-Fuller GLS (with trend)	-1.15	-2.72
Elliott-Rothenberg-Stock	-1.15	-2.72
Kwiatkowski-Phillips-Schmidt-Shin (H_0 = stationary)	1.46^*	1.75^*
Conclusion	I(1)	I(1)

Note: $* = p < 0.05$. 33 observations with 1-year lag included for all tests. H_0 = series contains a unit root.

In addition to modeling support for welfare, Kelly and Enns (2010) also examine the determinants of public mood liberalism. After first ensuring that public mood liberalism was I(1) and that none of the regressors were of an order of integration higher than I(1), I replicated their second model on how inequality affects liberal mood (Table 1, Model 2, p. 864), as shown in Table 4. I was able to replicate their results exactly. I found the GECM of Kelly and Enns (2010) to have good model fit using SBIC. Critically, the residuals were also white-noise. Therefore, I proceeded to perform the bounds F-test on all variables appearing in levels: public mood liberalism, policy liberalism, and inequality. The resulting F-statistic is 7.62. According to Narayan (2005, p. 1990), the critical values (at the 5 percent level) for two regressors, no trend and an unrestricted intercept for 55 observations is 3.987 and 5.090 for the lower- and upper-bounds, respectively. Given that the F-statistic falls well above the upper I(1) bound, we are able to reject the null hypothesis of no cointegration.

Table 4: Results of the ARDL-Bounds Model for Public Mood Liberalism (Kelly and Enns 2010)

	Original GECM & ARDL-Bounds
Liberal Mood$_{t-1}$	-0.25*
	(0.07)
ΔPolicy Liberalism$_t$	0.10
	(0.10)
Policy Liberalism$_{t-1}$	-0.09*
	(0.02)
ΔInequality$_t$	-27.07
	(34.61)
Inequality$_{t-1}$	-16.22
	(8.92)
Constant	21.70*
	(5.57)
Observations	54
Adjusted R^2	0.28
Breusch-Godfrey χ^2 of: AR(1)	0.14
AR(2)	0.15
AR(3)	5.46
Durbin's Alternative χ^2 of: AR(1)	0.12
AR(2)	0.13
AR(3)	5.06
Cumby-Huizinga χ^2 of AR(1)-AR(3)	8.74*
Shapiro-Wilk z	-1.89

Note: Dependent variable is public mood liberalism. The model shows the results from Kelly and Enns (2010) (Table 1, Model 1 in their article), which had the best fit as determined by SBIC. Standard errors in parentheses. $* = p < 0.05$.

As a robustness check of this finding, I used the bounds t-test to test for the significance of the lagged dependent variable. Recall that only asymptotic critical values are available and given by Pesaran, Shin and Smith (2001, p. 303); the I(0) and I(1) bounds are -2.86 and -3.53, respectively. Even so, the t-statistic on the lag of liberal mood is -3.85, which falls below the I(1) critical bound. Therefore, the results from the bounds t-test lends further support to the finding that public mood liberalism is cointegrating with policy liberalism and inequality.

Taken together, the results in Table 4 suggest that there is cointegration among public mood liberalism. This lends support to the authors' conclusions that are consistent with the Benabou (2000) model; inequality tends to lessen public mood liberalism in the US.

5.1.1 Different Conclusions About the Time Series Properties of Welfare Policy Mood

An interesting counterfactual to consider from the main results is whether the main results would have changed, given alternative conclusions about the time series properties of the dependent variable: welfare policy mood. As discussed in the main paper, deciding whether the dependent variable is I(0) or I(1) is crucial, since it determines whether we need to first difference the series before running the model. Obviously, unit-root testing on short series such as welfare policy mood ($\boldsymbol{T} = 33$ in the models) is very difficult. However, the Phillips-Perron (PP), Dickey-Fuller GLS (DF-GLS), and Elliott-Rothenberg-Stock (ERS) tests all indicate that the series is

I(1). Only the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests point towards a possibly stationary series, and the latter test shows very borderline results.

Although the DF-GLS and ERS tests are thought to be superior to the ADF and Phillips-Perron tests (Maddala and Kim 1998; Choi 2015; Enders 2010), the ADF test is still very common amongst time series practitioners; a Google Scholar search for "Dickey-Fuller" turns up nearly 60,000 results, and their paper has been cited almost 20,000 times (Dickey and Fuller 1979).

Assuming we only used the ADF test, we might have concluded that welfare policy mood is stationary. The next step would be to evaluate whether the independent variables are I(1) (step (c) in the schematic diagram in the main paper). As shown in Table 3, *all* unit root tests find evidence that policy liberalism and income inequality are I(1). We therefore need to first-difference the independent variables (step (f)) before including them in a model where the dependent variable is stationary.

The results of the ARDL model in lagged-dependent variable form are shown in Table 5. In order to ensure white-noise residuals, the first difference of welfare policy mood lagged three periods back was included (step (i)). As a result, we end up at step (j) in the schematic (an ARDL model with the dependent variable estimated in levels). It is clear from the results that welfare policy mood is relatively strongly related to its previous levels and past changes. In the short-run, neither change in policy liberalism nor changes in inequality appear to be related to welfare. These results echo the findings in the main paper (Table 1, Model 5) that short-run changes

in these variables are not associated with movements in welfare.

Table 5: Results of Concluding that Welfare Policy Mood is I(0)

	ARDL Model
Welfare_{t-1}	0.82*
	(0.06)
$\Delta\text{Welfare}_{t-3}$	0.46*
	(0.11)
$\Delta\text{Policy Liberalism}_t$	0.09
	(0.09)
$\Delta\text{Inequality}_t$	-23.32
	(33.04)
Constant	10.87*
	(3.75)
Observations	51
Adjusted R^2	0.87
Breusch-Godfrey χ^2 of: AR(1)	0.50
AR(2)	0.77
AR(3)	2.56
Durbin's Alternative χ^2 of: AR(1)	0.44
AR(2)	0.67
AR(3)	2.28
Cumby-Huizinga χ^2 of AR(1)-AR(3)	2.57
Shapiro-Wilk z	-0.44

Note: Dependent variable is welfare policy mood, with lag structures determined by SBIC. Standard errors in parentheses. $* = p < 0.05$.

However, unlike Table 1, Model 5 in the main paper, the dependent variable in Table 5 appears in levels and not first differences. Therefore, changes in inequality and policy liberalism still might have a longer-run effect on welfare policy mood (although of course this effect will die out over time in a lagged dependent variable model). To examine this, I used the Stata program `dynpss` to create expected values

of welfare policy mood over time in response to a one standard deviation shock to each of the independent variables at time $t = 10$ (Philips 2016*a*).[35] The results are shown in Figure 42. As shown in the plot on the left, positive changes in policy liberalism have a small increase on the level of welfare policy mood, but this effect is temporary and not statistically significant. Overall, by deciding that welfare policy mood is I(0) rather than I(1), we reach a similar substantive conclusion about the effects of policy liberalism and inequality.

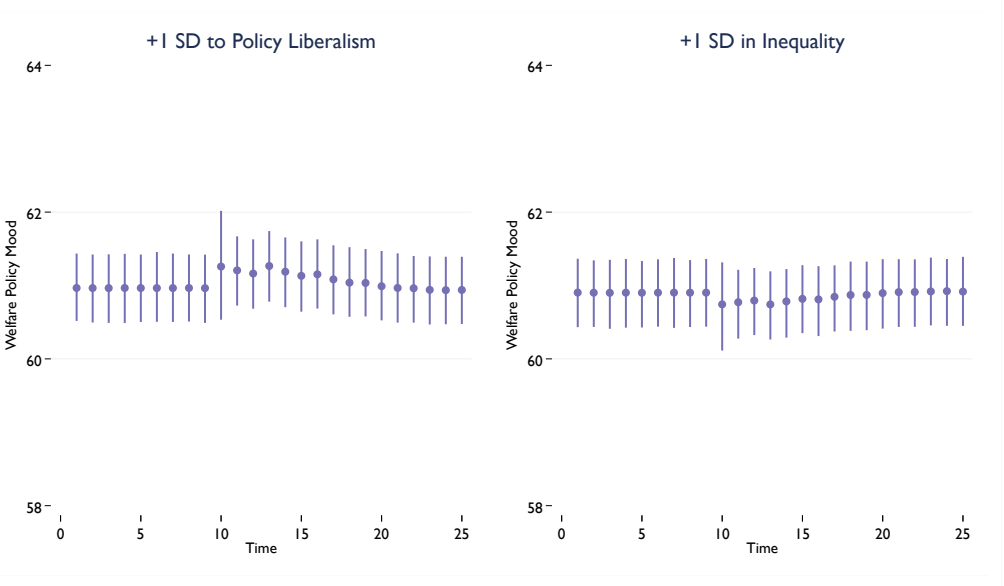

Figure 42: Policy Liberalism and Inequality's Effect on Welfare Policy Mood

Note: Plots show dynamic simulation using the ARDL model in Table 42. A 1 standard deviation increase in each variable occurs at time $t = 10$. 95% confidence intervals shown.

[35]Note that these are one standard deviations of the first difference of each variable, not the undifferenced series. This provides for a more plausible scenario of realistic changes in these variables. In fact, changes more extreme (i.e., larger than one standard deviation) actually occurred for both variables over the time period in the sample.

5.2 Replication II: Volscho and Kelly (2012)

Unit root tests of the first-difference of the regressors indicated that first differencing rendered each series stationary. For the first-difference of Democratic President, an augmented Dickey-Fuller test with one lag yielded a test statistic of -5.31, which was statistically significant at the 0.001 level. For Congressional Democrat, Divided Government, and Union Membership, the test statistics (and p-values) were -7.60 (0.001), -6.54 (0.001), and -3.37 (0.01) respectively.

As discussed in the main paper, the largest difference between Volscho and Kelly's original model and the ARDL-bounds model is the significance of the short-run effect of a Democratic president. To see if this changes the substantive conclusions of the authors, I turn to the Stata program `dynpss` for dynamic interpretations of the results (Philips 2016*a*). It uses stochastic simulation to produce 1,000 parameter estimates that are multivariate normal, with a mean equal to the estimated parameters in the results table in the main paper, and variance equal to the estimated variance-covariance matrix.[36] Next, all lagged variables are set to their sample means. All differenced variables are set to zero, which creates a stable dynamic relationship. Then, expected values are generated, which now become the new lagged dependent variable value, and the simulation is repeated again for time $t = 2, 3, \cdots, 25$. To gain dynamic inferences, at time $t = 10$ a one-period change to one of the independent variables occurs. This appears in the model first through the differenced independent

[36]Since the model contains i.i.d. residuals, the asymptotic sampling distribution is multivariate normal for the coefficients. Variance draws are from a scaled inverse χ^2 with $(n-k)$ degrees of freedom, where n is the number of observations and k is the number of parameters.

variable, then the lags, as well as any lagged differences. Expected values are then plotted over time. Measures of certainty are given by 95 percent confidence intervals calculated using the percentile method, although such confidence intervals tend to be more conservative than analytical hypothesis testing of coefficients (Philips, Rutherford and Whitten 2016*a*). Therefore, users should still conduct analytical hypothesis tests about the significance of individual coefficients.

Results from the dynamic simulations are shown in Figure 43. The plot on the left shows the effect of moving from a Republican to Democratic president at time $t = 10$.[37] I show simulations for both the ARDL-bounds approach as well as Volscho and Kelly's GECM.[38] In the short run, moving from a Republican to a Democratic president increases the income concentration of the top one percent. However, this effect loses statistical significance after four years, is not statistically significantly different from the predictions using Volscho and Kelly's GECM, and the long run effect is nearly zero. This is confirmed analytically by calculating the long-run multiplier, which is 0.36 and is not statistically significantly different from zero.

The plot on the right in Figure 43 shows the effect of a one standard deviation decrease in the percent of union membership at time $t = 10$. While this results in a small, instantaneous drop at time $t = 10$ using the ARDL-bounds predictions, this effect is not statistically significant from the average level of income concentration of the super-rich. Nor is it statistically significantly different from the Volscho and

[37]After $t = 10$, the first-difference of the presidential variable is set back to zero, and the lag of Democratic President moves from zero to one for the rest of the simulation.

[38]Pre-shock predicted values differ slightly due to sample size, and are jittered forward for clarity.

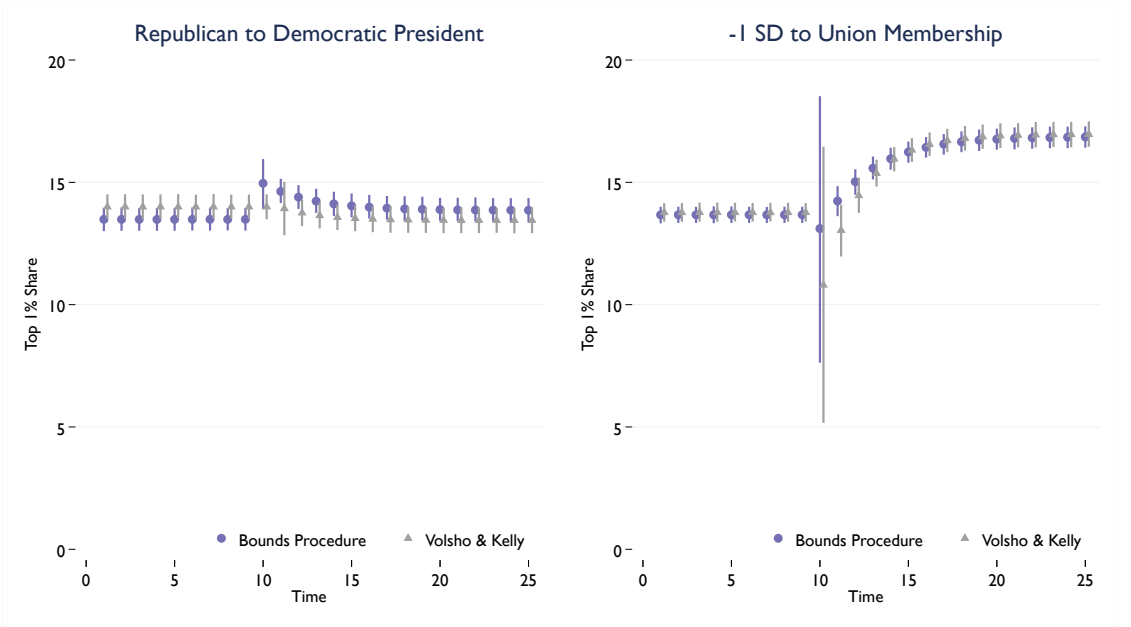

Figure 43: The Substantive Conclusions of Volscho and Kelly (2012) Remain Unchanged

Note: Plots show dynamic simulations from the Volscho and Kelly results table in the main paper. All changes occur at $t = 10$, holding all other variables at their sample means, and (for the plot on the right) president at its modal value. 95% confidence intervals shown. Volscho and Kelly simulation jittered forward in time for clarity. Pre-shock predicted values differ slightly due to sample size.

Kelly GECM prediction in the short-run. While the two predictions are almost statistically significantly different from one another when $t = 12$ and $t = 13$, they have similar long-run trajectories in response to a decrease in union membership; in the long-run the concentration of income by the top one percent increases from about 14 percent to close to 17 percent.[39] While statistically significant, the overall effects are substantively small, given that a one standard deviation decrease in union membership would be a very large shift in the structure of the labor market.

5.3 Replication III: Ura (2014)

As a third example of the utility of the approach outlined in the main paper, I replicate an article by Ura (2014), who uses a GECM to analyze how the ideology of the US Supreme Court has shaped aggregate public mood from 1956 to 2009. Like many other examples pointed out by Grant and Lebo (2016), although the choice to use an error correction model in this article is theoretically justified, there is no mention of testing for equation balance or the presence of unit roots. To see if this is the case, I performed the steps for conducting the ARDL procedure.

The first step is to test whether or not the dependent variable, Stimson's annual mood index, is I(1). These results are shown in Table 6. The relatively weak power of the Augmented Dickey-Fuller test in short samples stands out. All other tests indicate that a unit root is present in the public mood index. Therefore, based on the conclusions of the majority of the tests, we can conclude that there is an I(1)

[39] Analytically calculating out the long-run multiplier in Model 2 yields a statistically significant long-run multiplier of 3.18.

process, and proceed to model estimation in error-correction form, after ensuring that all independent variables are I(1) or less. Augmented Dickey-Fuller tests on the first difference of Policy, Unemployment, Inflation, and the Caselaw Index yielded test statistics (and significance levels) of -2.61 (0.09), -5.98 (0.001), -6.79 (0.001), and -3.23 (0.05). Although Policy was borderline I(2) when using the ADF test, the Phillips-Perron test could reject the null at the 0.01 level.

Table 6: Unit Root Test Statistics (Ura 2014)

Unit-Root Test	Public Mood Liberalism
Augmented Dickey-Fuller (with drift)	-1.73*
Phillips-Perron	-1.75
Dickey-Fuller GLS (with trend)	-1.79
Elliott-Rothenberg-Stock	-1.79
Kwiatkowski-Phillips-Schmidt-Shin (H_0 = stationary)	0.89*
Conclusion:	I(1)

Note: $* = p < 0.05$. 54 observations with 1-year lag included for all tests. H_0 = series contains a unit root for all tests except KPSS.

The second step is to estimate the ARDL model in error-correction form. The results are shown in Table 7, Model 2. For reference, the original GECM results from Ura (2014) are also shown in Model 1. Successive lags of the first difference of all variables were chosen via SBIC—only the lag of the first difference of *Unemployment* was needed in order to minimize the information criterion. While theory may have been used to guide lag specification in Ura (2014), I find that the Cumby-Huizinga test for autocorrelation finds evidence of up to an AR(3) process at the .05 level, and both the Breusch-Godfrey and Durbin's Alternative LM test suggest that there is an

AR(3) process at the .10 level of significance. This autocorrelation was eliminated by adding the lag of the first difference of unemployment, as shown in the second model in Table 7. Across both models, the Cook-Weisberg and Shapiro-Wilk tests find no evidence of heteroskedasticity or violation of normality, respectively.

After ensuring that we have a stable model, we can then use the bounds procedure of Pesaran, Shin and Smith (2001) to test if there is a cointegrating relationship between public mood, Supreme Court and Congressional liberalism, inflation, and unemployment. Running a joint F-test that all of the lagged coefficients are jointly equal to zero (i.e. $Mood_{t-1}$, $Policy_{t-1}$, $Unemployment_{t-1}$, $Inflation_{t-1}$, and $Caselaw\ Index_{t-1}$) yields a F-statistic of 5.15. Using the critical values provided in Narayan (2005) for $k = 4$ independent variables, and assuming an unrestricted constant and no trend, it is clear that the F-statistic exceeds the upper I(1) bound of 4.334 at the 95 percent level of significance.[40] We can also use the one-sided test on the significance of the lagged dependent variable to confirm the F-test. Although the the t-statistic on $Mood_{t-1}$, -3.19, lies within the stationary and I(1) bounds, which suggests that individual cointegration testing is necessary (-2.86 and -3.99, respectively, as given in Pesaran, Shin and Smith (2001)), these are asymptotic critical values and the result of the t-statistic still suggests that it is near the I(1) bound. Therefore, we can conclude that a cointegrating relationship is present in this example.

Looking back at Table 7, most of the results appear to hold across both specifica-

[40]By construction, if the I(1) bound is exceeded for the F-test, the stationary bound (3.068 in this case) is exceeded.

Table 7: Results of the ARDL Model (Ura 2014)

	(1) Original GECM	(2) ARDL Model
Mood_{t-1}	-0.28*** (0.08)	-0.24*** (0.08)
ΔPolicy_t	0.07 (0.07)	0.05 (0.07)
Policy_{t-1}	-0.07*** (0.02)	-0.07*** (0.02)
$\Delta\text{Unemployment}_t$	-0.32 (0.27)	-0.11 (0.27)
$\Delta\text{Unemployment}_{t-1}$		-0.54** (0.24)
$\text{Unemployment}_{t-1}$	-0.24 (0.19)	-0.02 (0.20)
$\Delta\text{Inflation}_t$	-0.30** (0.13)	-0.31** (0.12)
Inflation_{t-1}	-0.29** (0.13)	-0.30** (0.12)
$\Delta\text{Caselaw Index}_t$	-0.09** (0.04)	-0.09** (0.04)
$\text{Caselaw Index}_{t-1}$	0.02** (0.01)	0.03** (0.01)
Constant	19.49*** (5.14)	15.65*** (5.05)
Observations	54	53
Adjusted R^2	0.30	0.34
Breusch-Godfrey's χ^2 of: AR(1)	1.84	0.20
AR(2)	1.96	0.65
AR(3)	7.29*	3.06
Durbins's Alternative's χ^2 of: AR(1)	1.51	0.16
AR(2)	1.58	0.50
AR(3)	6.40*	2.39
Cumby-Huizinga χ^2 of AR(1)-AR(3)	7.95**	4.23
Shapiro-Wilk's z	-0.35	0.62

Note: Model 1 shows results from Ura (2014) and Model 2 shows results using ARDL procedure, with lag structure determined by SBIC. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

tions. The adjustment parameter, $Mood_{t-1}$, is only slightly smaller, which indicates a slower rate of return to equilibrium. The variables for the Supreme Court's liberalism (*Caselaw Index*), the *Inflation* level of the US, and the index of Congressional liberalism in policy passage (*Policy*) are virtually the same across model specifications. The one large difference is the variable for unemployment. While Ura (2014) found that unemployment had no effect on policy mood (in both the short and long-run via calculation of the long-run multiplier), the ARDL model suggests that unemployment may actually make public mood more conservative.

To see the substantive implications of this, I plot the result of a one standard deviation increase in unemployment (about 1.38 percentage points) using the program `dynpss` (Philips 2016*a*). The results are shown in Figure 44. The left-side plot shows the one standard deviation increase in unemployment, holding all else equal, while the right-side plot shows a one standard deviation increase in the *Caselaw Index*—which is similar to the figure shown in Ura (2014, pg. 120).

It is clear that the small move towards a more conservative public mood due to an increase in unemployment is not statistically significant. This stands in contrast with a one standard deviation increase in the liberal stance of the Supreme Court (as proxied by the caselaw index)—public mood grows more conservative in the short run, becomes larger than the average public mood after just five years, and eventually moves to a new equilibrium where public mood equals 62. Therefore, in the long-run the public becomes more liberal—by about four points on the policy mood scale—in response to a more liberal Supreme Court.

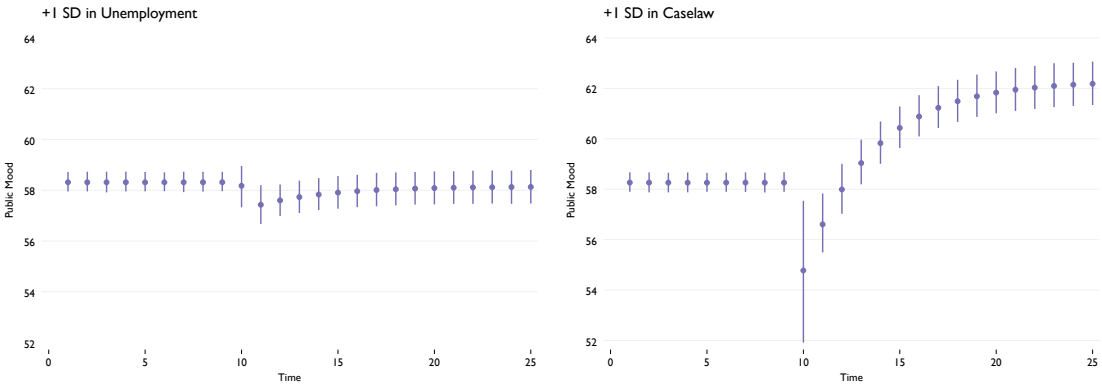

Figure 44: The Substantive Results of Ura (2014) Hold

Note: Plots show dynamic simulation using the ARDL model in Table 7. A 1 standard deviation increase in each variable occurs at time $t = 10$. 95% confidence intervals shown.

5.4 A Comparison of the Replications to the Replications of Grant and Lebo (2016)

In their Supplemental Materials, Grant and Lebo (2016) (henceforth GL) also replicate Kelly and Enns (2010) (KE) and Volscho and Kelly (2012) (VK). I briefly summarize the findings of GL and comment on their similarities and differences to my findings below.

Turning the the KE replication, GL find evidence that *Welfare Support* is I(1); this is consistent with the findings in the main paper. GL then examine the Type I error rates of the adjustment parameter by regressing each of KE's dependent variables on randomly generated I(1) series. They find that the adjustment parameter on *Welfare Support* (the dependent variable of the model I replicate in the main paper in Table 1) is spuriously significant about 85 percent of the time when using one-tailed t-tests,

but only about 1.3 percent when using the appropriate MacKinnon values.[41] They also include two independent variables (beef consumption and coal emissions) that should be unrelated to *Welfare Support*. However, they find that although these "nonsense" variables were insignificant (for their lags and first differences), there was still a significant adjustment parameter. Last, GL run a fractional error-correction model on *Welfare Support*, and fail to find evidence of fractional cointegration.

In the main paper, I took a slightly different approach to replicating KE, but largely arrived at the same conclusions as GL. In Table 1 in the main paper, I also have significant adjustment parameters (as judged by a two-tail t-test). Using the bounds procedure recommended in the main paper, I could find no evidence of cointegration, and as GL discuss, there appears to be no fractional cointegration either. Therefore, policy liberalism and income inequality may still affect public mood towards welfare policy, but only in the short-run. In both GL and my analysis, there does not appear to be a long-run relationship.

GL also replicate VK in their Supplemental Materials. They note that VK say they have a mix of I(1) and I(0) independent variables. GL then estimate a fractional error-correction model, finding that only market factors seem to affect the concentration of income of the top one percent. Note that while GL test *all* models in VK, I examine only a single model.

In the main paper, I tested the main dependent variable of VK, *Top 1% Share*,

[41] GL also regress a randomly-generated bounded series on the independent variables in KE's models; again finding extremely high rates of significant adjustment parameters (using one-tailed t-tests), and much lower (though still greater than five percent) rates when using the appropriate MacKinnon critical values.

and find evidence of a unit root. This is largely confirmed by the estimation of the d parameter that GL find, once standard errors are taken into account.[42] However, while I find evidence of cointegration using the bounds procedure for VK's model, GL find no evidence using their three-step fractional error-correction model that a Democratic president, the percentage of congressional Democrats, and union membership affect the *Top 1% Share*. They do, however, find that in the short run a Democratic president increases the income concentration of the super-rich (a finding confirmed by my analysis in the main paper).

To conclude, my findings largely confirm the replication of KE that GL perform; neither find evidence of a cointegrating relationship. In contrast, while my replication of VK found evidence of cointegration when using the bounds test, GL did not find any evidence of cointegration when using a fraction error-correction model.

[42]GL find $d = 0.93$ with an asymptotic standard error of 0.10.

References

Balke, Nathan S and Thomas B Fomby. 1997. "Threshold cointegration." *International Economic Review* 38(3):627–645.

Benabou, Roland. 2000. "Unequal societies: Income distribution and the social contract." *American Economic Review* 90(1):96–129.

Box-Steffensmeier, Janet M and Renee M Smith. 1998. "Investigating political dynamics using fractional integration methods." *American Journal of Political Science* 42(2):661–689.

Cappuccio, Nunzio and Diego Lubian. 1996. "PRACTITIONERS'CORNER: Triangular Representation and Error Correction Mechanism in Cointegrated Systems." *Oxford Bulletin of Economics and Statistics* 58(2):409–415.

Cheung, Yin-Wong and Kon S Lai. 1993. "A fractional cointegration analysis of purchasing power parity." *Journal of Business & Economic Statistics* 11(1):103–112.

Choi, In. 2015. *Almost all about unit roots: Foundations, developments, and applications*. Cambridge University Press.

Clarke, Harold D and Matthew Lebo. 2003. "Fractional (co) integration and governing party support in Britain." *British Journal of Political Science* 33(02):283–301.

De Boef, Suzanna and Luke Keele. 2008. "Taking time seriously." *American Journal of Political Science* 52(1):184–200.

Dickey, David A and Wayne A Fuller. 1979. "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American statistical association* 74(366a):427–431.

Enders, Walter. 2010. *Applied econometric time series*. 3 ed. John Wiley and Sons.

Engle, Robert F and Clive WJ Granger. 1987. "Co-integration and error correction: representation, estimation, and testing." *Econometrica* 55(2):251–276.

Enns, Peter K, Nathan J Kelly, Takaaki Masaki and Patrick C Wohlfarth. 2016. "Don't jettison the general error correction model just yet: A practical guide to avoiding spurious regression with the GECM." *Research and Politics* 3(2):1–13.

Esarey, Justin. 2016. "Fractionally integrated data and the autodistributed lag model: Results from a simulation study." *Political Analysis* 24:42–49.

Fraley, C, F Leisch, M Maechler, V Reisen and A Lemonte. 2006. "fracdiff: Fractionally differenced ARIMA aka ARFIMA (p, d, q) models." *R package version* pp. 1–3.

Gandrud, Christopher, Laron K Williams and Guy D Whitten. 2016. "dynsim: Dynamic simulations of autoregressive relationships." *R package version 1.2.2* .

Gelman, Andrew, John B Carlin, Hal S Stern and Donald B Rubin. 2014. *Bayesian data analysis*. Vol. 2 Chapman and Hall/CRC Boca Raton, FL, USA.

Gonzalo, Jesus and Tae-Hwy Lee. 1998. "Pitfalls in testing for long run relationships." *Journal of Econometrics* 86(1):129–154.

Grant, Taylor and Matthew J. Lebo. 2016. "Error correction methods with political time series." *Political Analysis* 24:3–30.

Helgason, Agnar Freyr. 2016. "Fractional integration methods and short time series: Evidence from a simulation study." *Political Analysis* 24(1):59–68.

Johansen, Søren. 1991. "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models." *Econometrica: Journal of the Econometric Society* 59(6):1551–1580.

Jordan, Soren and Andrew Q Philips. 2016. "`pss`: `R` package to perform the bounds test for cointegration and create dynamic simulations." Available at: https://github.com/andyphilips/pss. R package version 1.3.9.

Keele, Luke and Nathan J Kelly. 2006. "Dynamic models for dynamic theories: The ins and outs of lagged dependent variables." *Political Analysis* 14(2):186–205.

Keele, Luke, Suzanna Linn and Clayton M Webb. 2016. "Treating time with all due seriousness." *Political Analysis* 24:31–41.

Kelly, Nathan J and Peter K Enns. 2010. "Inequality and the dynamics of public opinion: The self-reinforcing link between economic inequality and mass preferences." *American Journal of Political Science* 54(4):855–870.

Maddala, Gangadharrao S and In-Moo Kim. 1998. *Unit roots, cointegration, and structural change*. Cambridge University Press.

Narayan, Paresh Kumar. 2005. "The saving and investment nexus for China: Evidence from cointegration tests." *Applied Economics* 37(17):1979–1990.

Pesaran, M Hashem, Yongcheol Shin and Richard J Smith. 2001. "Bounds testing approaches to the analysis of level relationships." *Journal of Applied Econometrics* 16(3):289–326.

Philips, Andrew Q. 2016*a*. "`dynpss`: Stata module to dynamically simulate autoregressive distributed lag (ARDL) models." Available at: https://andyphilips.github.io/dynpss/.

Philips, Andrew Q. 2016*b*. "`pssbounds`: Stata module to conduct the Pesaran, Shin, and Smith (2001) bounds test for cointegration." Available at: http://andyphilips.github.io/pssbounds/.

Philips, Andrew Q, Amanda Rutherford and Guy D Whitten. 2016*a*. "Dynamic pie: A strategy for modeling trade-offs in compositional variables over time." *American Journal of Political Science* 60(1):268–283.

Philips, Andrew Q, Amanda Rutherford and Guy D Whitten. 2016*b*. "dynsimpie: A command to examine dynamic compositional dependent variables." *Stata Journal* 16(3):662–677.

Phillips, Peter CB. 1991. "Optimal inference in cointegrated systems." *Econometrica: Journal of the Econometric Society* pp. 283–306.

Tomz, Michael, Jason Wittenberg and Gary King. 2003. "CLARIFY: Software for interpreting and presenting statistical results." *Journal of Statistical Software* 8(1):1–30.

Ura, Joseph Daniel. 2014. "Backlash and legitimation: Macro political responses to supreme court decisions." *American Journal of Political Science* 58(1):110–126.

Volscho, Thomas W and Nathan J Kelly. 2012. "The rise of the super-rich: Power resources, taxes, financial markets, and the dynamics of the Top 1 percent, 1949 to 2008." *American Sociological Review* 77(5):679–699.

Wickham, Hadley and Winston Chang. 2015. "devtools: Tools to make developing R code easier." *R package version* 1(0).

Williams, Laron K and Guy D Whitten. 2011. "Dynamic simulations of autoregressive relationships." *Stata Journal* 11(4):577–588.