Comment on Suri (2011) "Selection and Comparative Advantage in Technology Adoption"*

Emilia Tjernström[1], Dalia Ghanem[2], Aleksandr Michuda[3], Oscar Barriga-Cabanillas[4], Travis J. Lybbert[2], and Jeffrey D. Michler[5]

[1] *Monash University*
[2] *University of California, Davis*
[3] *Swarthmore College*
[4] *World Bank*
[5] *University of Arizona*

This version: December 2024

Abstract

This paper illustrates and addresses weak identification concerns in the correlated random coefficient (CRC) model that Suri (2011) uses to study agricultural technology adoption. Using the publicly available version of the dataset used in Suri (2011), cleaned per the author's instructions, we are unable to replicate the paper's main CRC model results. To understand why, we recast the CRC model as a more general random coefficient model in which the returns to hybrid adoption are restricted to be linear in comparative advantage. This reveals that the key structural parameter in the CRC model (ϕ) is prone to a weak identification problem. We then propose a procedure to conduct weak-identification robust inference on ϕ using test inversion. Only with this robust procedure accounting for weak identification are we able to reconcile our results with the original Suri (2011) results.

JEL Classification: C12, C14, O33,Q16
Keywords: heterogeneity, correlated random coefficients, weak identification, generalized method of moments

*We thank the Data Management Team at Tegemeo Institute of Agricultural Policy and Development for providing us with access to the data and Tavneet Suri for data cleaning instructions, code, and helpful comments. This research was supported by the Cornell Center for Social Sciences Data Discovery & Replication.

1 Introduction

In an influential paper, Suri (2011) addresses why many Sub-Saharan African farmers still use traditional farming techniques despite higher returns from modern agricultural technologies. Existing literature identifies market frictions like credit constraints, uninsured risk, and incomplete information as likely causes.[1] Suri (2011) instead attributes this to heterogeneous returns to adoption stemming from time-invariant unobservables.

Suri (2011) introduces a specific restriction on heterogeneous returns to hybrid maize adoption, which we call the Linearity in Comparative Advantage (LCA) restriction.[2] The key structural parameter, ϕ, in this correlated random coefficient (CRC) model indicates the slope of this linear relationship.

Using a Kenyan panel dataset in which farmers are observed growing either hybrid or non-hybrid maize, Suri (2011) finds negative and significant ϕ estimates, implying that farmers with the lowest non-hybrid productivity gain the most from hybrid adoption. She then extrapolates these estimated returns, using the LCA restriction, to non-adopters. Our reanalysis fails to replicate these negative and significant ϕ estimates using Suri (2011)'s CRC approach. Reliable parameter estimates from a model like Suri (2011) can potentially adjudicate among competing prescriptions and interventions that aim to encourage non-adopters to adopt productivity-enhancing technologies. Given these policy implications, valid inference about key parameters in this model can be consequential.

We investigate these discrepancies and the potential for weak identification of ϕ.[3] We show that the CRC model in Suri (2011) is a restricted version of a group random coefficient (GrRC) model that imposes the LCA restriction. The unrestricted GrRC model helps detect and diagnose the weak-identification problem. We then propose a weak-identification robust procedure for inference on the LCA parameter using test inversion. A simulation study demonstrates our procedure's finite-sample performance.

Finally, we reanalyze Suri (2011) using our GrRC approach. Our results suggest the presence of weak identification. Our weak-identification robust confidence intervals for ϕ are

[1]Surveys include Feder *et al.* (1985); Foster and Rosenzweig (2010); Magruder (2018); Jack (2013).

[2]Verdier (2020) shows that the approach of Suri (2011) relies on a linear extrapolation that is valid if the factors that determine selection other than the treatment effect are uncorrelated with outcomes, and proposes a more robust extrapolation approach that allows for selection into treatment to be correlated with treatment effects as well as (a subset of) covariates.

[3]It is worth noting that Suri (2011) discusses ways of estimating ϕ as a combination of production function reduced-form parameters and recognizes that identification of ϕ relies on a specific restriction on these parameters. See the discussion in Section 4.5.1 in Suri (2011), where the author also informally assesses whether the denominator in these reduced-form expressions is bounded away from zero.

negative and overlap with Suri (2011)'s original point estimates. This suggests that, after accounting for weak identification, our results align with Suri (2011)'s main finding.

2 Reanalysis of Suri (2011)

2.1 Data

We use the same panel dataset of rural Kenyan households as Suri (2011), collected by the Tegemeo Institute of Agricultural Policy and Development.[4] Our dataset closely resembles the original dataset, with minor differences: we have 1,197 households instead of 1,202; further, small differences suggest slight variations in the composition of the sample or variable construction procedures.[5] Appendix A documents the steps we take to match the original paper's dataset and compares summary statistics.

We follow Suri (2011) in constructing variables. Maize yield is the ratio of (self-reported) maize harvest to plot size.[6] Technology adoption trajectories and fertilizer use are based on whether households report using hybrid seed or fertilizer in a given year. We control for the same demographic and production variables as Suri (2011).

2.2 Reanalysis Results

Table 1 reports the OLS and fixed effects (FE) specifications that Suri (2011) uses to estimate average yield advantages of hybrid maize under the assumption of homogeneous returns. Panel A shows the original paper's point estimates, while Panel B shows our reanalysis.[7] Our results are statistically indistinguishable from Suri (2011).

The crux of our reanalysis is estimating the CRC model and the LCA parameter, ϕ. We briefly present this model before discussing our results (for more details, see Suri (2011)). The CRC model is given by:

$$y_{it} = \tau_i + \theta_i + (\beta + \phi\theta_i)h_{it} + x_{it}'\gamma + h_{it}x_{it}'\delta + \varepsilon_{it}, \quad (1)$$

where y_{it} is log of maize yield for household i at time t; τ_i is farmer i's absolute advantage,

[4] Tegemeo Institute makes the data available to researchers subject to a brief application form.

[5] We obtained our data from Tegemeo Institute and followed the author's data cleaning documentation.

[6] This is not a measure of economic returns, as it does not account for input costs. We assume, as in Suri (2011), that input costs other than hybrid seeds are constant across sectors.

[7] The results in Panel A correspond to Table IIIA in Suri (2011).

assumed to be (mean) independent of technology adoption and covariates (i.e., $E[\tau_i|h_i, x_i] = E[\tau_i]$), ε_{it} is an idiosyncratic error term with $E[\varepsilon_{it}|h_i, x_i, \tau_i, \theta_i] = 0$, and θ_i can be correlated with the individual's adoption history. The adoption history for individual i is denoted by $h_i = (h_{i1}, \ldots, h_{iT})$, and $x_i = (x_{i1}, \ldots, x_{iT})$ denotes the time series of covariates. To estimate different means across adoption histories in the CRC model following Chamberlain (1984), a linear projection of θ_i onto a fully saturated model of adoption histories is used:

$$\theta_i = \lambda_0 + \lambda_1 h_{i1} + \lambda_2 h_{i2} + \lambda_3 h_{i1} h_{i2} + \eta_i. \tag{2}$$

The CRC model results are in Table 2, estimated using the `Stata` package from Barriga-Cabanillas *et al.* (2018). Despite using nearly identical data to Suri (2011), we only replicate the estimates for λ_1 and λ_2. Our ϕ estimates are mostly insignificant, suggesting no detectable patterns of comparative advantage to hybrid maize production.

Our inability to reproduce this key finding in Suri (2011) is unexpected and apparently due to trivial differences in the data. To assess the sensitivity of the CRC estimates to sample composition, we re-estimate the model when randomly dropping 10 households from the sample across 10,000 simulations. Figure 1 presents the resulting distribution of CRC estimates of ϕ, which highlights significant sensitivity. This motivates us to introduce a more general model that includes the Suri (2011) CRC model as a special case and demonstrates the potential for weak identification of the key LCA parameter.

3 A Group Random Coefficient (GrRC) Alternative

We start with an unrestricted random coefficient model that nests the LCA restricted model as a special case. We then show how the LCA parameter, ϕ, can be identified from a restriction on a GrRC model. This model highlights the potential for weak identification of ϕ and allows us to propose a weak-identification robust inference procedure for it.

3.1 Unrestricted Random Coefficient Model

Suppose that (log) yield is a function of hybrid adoption, h_{it}, farmer ability, a_i, and idiosyncratic shocks, ε_{it}, given by $y_{it} = f(h_{it}, a_i) + \varepsilon_{it}$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. For simplicity, we consider a model without covariates, but our results extend to the inclusion of exogenous, additively separable covariates as in (1).[8] As in Suri (2011), we maintain the strict exogeneity assumption, $E[\varepsilon_{it}|h_i, a_i] = 0$, where $h_i = (h_{i1}, \ldots, h_{iT})$. Since h_{it} is a binary variable, we can write this as a random coefficient model

$$y_{it} = \mu_i + \Delta_i h_{it} + \varepsilon_{it}, \tag{3}$$

where $\mu_i \equiv f(0, a_i)$ denotes farmer i's expected (log) yield without adoption and $\Delta_i \equiv f(1, a_i) - f(0, a_i)$ farmer i's expected returns to hybrid maize adoption. This model nests the CRC model of Suri (2011) as a special case if we impose two restrictions: (1) $\mu_i = \tau_i + \theta_i$ and (2) $\Delta_i = \beta + \phi\theta_i$. In addition, we normalize $E[\theta_i] = 0$ such that $E[\mu_i] = E[\tau_i]$.

3.2 GrRC Model

Implementing the GrRC approach relies on the fact that with a binary variable and finite number of time periods, there is a finite number of adoption histories, $\underline{h} \equiv (\underline{h}_1, \ldots, \underline{h}_T) \in \mathcal{H} = \{0,1\}^T$. For Suri (2011)'s two-period case, we can define $\mathcal{H} = \{0,1\}^2$. This includes the set of switcher trajectories ($\mathcal{H}_S = \{\{0,1\},\{1,0\}\}$), respectively called joiners and leavers, and the set of stayer trajectories ($\mathcal{H}_S^c = \{(0,0),(1,1)\}$, respectively called never-adopters and always-adopters.

Integrating the unrestricted random coefficient model (3) with respect to $a_i|h_i$ yields the following conditional mean model under strict exogeneity, $E[\varepsilon_{it}|h_i, a_i] = 0$,

$$E[y_{it}|h_i = \underline{h}] = \mu_{\underline{h}} + \Delta_{\underline{h}}\underline{h}_t, \tag{4}$$

where $\mu_{\underline{h}} \equiv E[\mu_i|h_i = \underline{h}] = E[f(0, a_i)|h_i = \underline{h}]$, $\Delta_{\underline{h}} \equiv E[\Delta_i|h_i = \underline{h}] = E[f(1, a_i) - f(0, a_i)|h_i = \underline{h}]$, and $\underline{h}_t$ is the t^{th} element of $\underline{h}$ for $t = 1, 2$. By the time homogeneity of $\mu_{\underline{h}}$ and $\Delta_{\underline{h}}$, we can identify the average returns to adoption for subpopulations that we observe adopting and not adopting hybrids in our data, i.e., the joiners and leavers. We can only identify $\mu_{(0,0)}$ for the never-adopters and $\kappa_{(1,1)} = \mu_{(1,1)} + \Delta_{(1,1)}$ for the always-adopters.

Using the following GrRC model, we can consistently estimate all the above objects using

$$y_{it} = \sum_{\underline{h} \in \mathcal{H} \setminus (1,1)} \mu_{\underline{h}} 1\{h_i = \underline{h}\} + \sum_{\underline{h} \in \mathcal{H}_S} \Delta_{\underline{h}} h_{it} 1\{h_i = \underline{h}\} + \kappa_{(1,1)} h_{it} 1\{\underline{h} = (1,1)\} + \varepsilon_{it}. \tag{5}$$

[8] In supplementary analysis available upon request, we augment our approach to allow for endogenous covariates following Suri (2011).

An advantage of the GrRC model relative to the reduced form of the CRC model in Suri (2011) is that all of the GrRC coefficients have economic meaning: $\mu_{\underline{h}}$ is the average yield without hybrid adoption for subpopulation $\underline{h}$, $\Delta_{\underline{h}}$ is the average return to adoption for switcher subpopulation $\underline{h}$, and $\kappa_{(1,1)}$ is the average yield with hybrid for the always-adopters.

3.3 Unrestricted GrRC Model and LCA Parameter

Next, we impose the LCA restriction on the GrRC model, specifically $\Delta_i = \beta + \phi\theta_i$, and illustrate how the unrestricted model can indicate potential identification concerns for ϕ, the LCA parameter. We first establish the relationship between parameters in the unrestricted GrRC model and those in the Suri (2011) model.

Proposition 1. *Let* $y_{it} = \mu_i + \Delta_i h_{it} + \varepsilon_{it}$*. Assume* $\mu_i = \tau_i + \theta_i$*,* $\Delta_i = \beta + \phi\theta_i$*,* $E[\theta_i] = 0$*,* $E[\tau_i|h_i] = E[\tau_i]$*,* $E[\varepsilon_{it}|h_i, \tau_i, \theta_i] = 0$*, the following equalities hold for* $\underline{h}, \underline{h}' \in \mathcal{H} = \{0,1\}^T$*,*

(i) $\Delta_{\underline{h}} = \beta + \phi\theta_{\underline{h}}$*,*

(ii) $\mu_{\underline{h}} - \mu_{\underline{h}'} = \theta_{\underline{h}} - \theta_{\underline{h}'}$*,*

(iii) $\Delta_{\underline{h}} - \Delta_{\underline{h}'} = \phi\left(\mu_{\underline{h}} - \mu_{\underline{h}'}\right)$*, for* $\underline{h} \neq \underline{h}'$*.*

where $\theta_{\underline{h}} = E[\theta_i|h_i = \underline{h}]$.

The conditions required for the above proposition are imposed in Suri (2011). The proof of (i) follows from the definition of Δ_i and $\Delta_{\underline{h}}$ as its conditional expectation:

$$\Delta_{\underline{h}} = E[\Delta_i|h_i = \underline{h}] = \beta + \phi E[\theta_i|h_i = \underline{h}] = \beta + \phi\theta_{\underline{h}}. \tag{6}$$

The proof of (ii) follows from the mean independence restriction, $E[\tau_i|h_i] = E[\tau_i]$. Proposition 1 (iii) follows from (i) and (ii).

If $\mu_{\underline{h}} \neq \mu_{\underline{h}'}$, we can re-write ϕ as the ratio of difference in returns to adoption for different subpopulations to the difference in their comparative advantage. Since we can identify both $\mu_{\underline{h}}$ and $\Delta_{\underline{h}}$ for switcher subpopulations, ϕ is identified as follows in the two-period case when $\mu_{(1,0)} \neq \mu_{(0,1)}$,

$$\phi = \frac{\Delta_{(1,0)} - \Delta_{(0,1)}}{\mu_{(1,0)} - \mu_{(0,1)}}. \tag{7}$$

This ratio points to the source of potential weak-identification for ϕ. As we illustrate numerically in Section 3.5, this issue emerges when the difference in average yield without adoption for joiners and leavers is relatively small.[9] The unrestricted GrRC model lets us estimate both parameters without imposing the LCA restriction, similar to the first stage of an instrumental variables (IV) regression in terms of detecting potential identification concerns (see Section 3.6 for an illustration). Unlike the first stage of an IV, however, the unrestricted GrRC has an economic interpretation and indicates the degree of response heterogeneity to technology adoption, albeit only for the switcher subpopulations.

3.4 Weak-identification Robust Inference on ϕ

In practice, ϕ may be weakly identified, as in our empirical application (see Section 3.6). We use the restrictions on the LCA parameter from Proposition 1 to conduct weak-identification robust inference on this key structural parameter. We propose a weak-identification robust confidence interval for ϕ based on inverting $W_n(\phi_0)$, the Wald statistic, of

$$H_0 : \Delta_{\underline{h}} - \Delta_{\underline{h}'} = \phi_0 \left(\mu_{\underline{h}} - \mu_{\underline{h}'}\right) \text{ for } \underline{h} \in \mathcal{H}^S, \underline{h}' \in \mathcal{H}^S, \underline{h} \neq \underline{h}', \tag{8}$$

where $\mathcal{H}^S = \{\underline{h} \in \mathcal{H} : 0 < \sum_{t=1}^{T} \underline{h}_t < T\}$. Assuming sufficient regularity conditions such that $W_n(\phi_0) \xrightarrow[H_0:\phi=\phi_0]{d} \chi^2_{|\mathcal{H}^S|-1}$, the $(1-\alpha)\%$ confidence interval is defined as

$$C_\alpha = \{\phi_0 \in \Phi : W_n(\phi_0) < c_{\alpha,|\mathcal{H}^S|-1}\} \tag{9}$$

where Φ is a parameter space, $c_{\alpha,|\mathcal{H}^S|-1}$ is the $(1-\alpha)$-quantile of the $\chi^2_{|\mathcal{H}^S|-1}$ distribution. Since ϕ is a scalar parameter, we can compute the confidence interval with a fine grid search.[10]

3.5 Simulations

A small-scale simulation study illustrates the weak-identification concern and our proposed weak-identification-robust inference procedure. We consider a simple two-period model that

[9] If ϕ is identified, we can also identify $\mu_{(1,1)}$, which allows us to identify β and $\theta_{\underline{h}}$ for all $\underline{h} \in \mathcal{H}$. Let $\pi_{\underline{h}} = P(h_i = \underline{h})$ for $\underline{h} \in \mathcal{H}$. Note that $E[\mu_i] = \sum_{\underline{h}\in\mathcal{H}} \pi_{\underline{h}}\mu_{\underline{h}}$. Since $E[\theta_i] = 0$, $E[\mu_i] = E[\tau_i]$ and $\theta_{\underline{h}} = \mu_{\underline{h}} - \sum_{\underline{h}\in\mathcal{H}} \mu_{\underline{h}}$. Since $\Delta_{\underline{h}} = \beta + \phi\theta_{\underline{h}}$, we can therefore also identify $\beta = \Delta_{(0,1)} - \phi\theta_{(0,1)} = \Delta_{(1,0)} - \phi\theta_{(1,0)}$.

[10] With more than two switcher subpopulations, as in the $T > 2$ case, the confidence interval may be empty if the over-identifying restrictions are violated. This is reminiscent of the behavior of the Anderson-Rubin confidence sets in the weak-instrument setting with over-identifying restrictions (Andrews *et al.*, 2019). In our context, the over-identifying restrictions arise from assuming that ϕ is time-invariant. One empirically compelling approach to address this issue is to allow this parameter to vary across time.

satisfies the LCA restriction described in Table 3. In addition to the CRC estimator and the weak-identification robust inference procedure, we also include the GrRC model with the LCA restriction (restricted GrRC):[11]

$$y_{it} = \sum_{\underline{h} \in \mathcal{H} \setminus (1,1)} \mu_{\underline{h}} + \Delta_{(0,1)} h_{it} + \phi(\mu_{(1,0)} - \mu_{(0,1)}) h_{it} 1\{h_i = (1,0)\} \\ + \left(\mu_{(1,1)} + \phi\left(\mu_{(1,1)} - \mu_{(0,1)}\right)\right) h_{it} 1\{h_i = (1,1)\} + \varepsilon_{it}. \quad (10)$$

We include this estimator for completeness, but acknowledge that it will suffer from the same weak-identification problem as the CRC estimator.

Table 3 presents simulation summary statistics that illustrate how a small difference between $\mu_{(0,1)}$ and $\mu_{(1,0)}$, can lead to a weak-identification problem for ϕ. Both the CRC and restricted GrRC estimators suffer from severe bias, which worsens when the difference between $\mu_{(0,1)}$ and $\mu_{(1,0)}$ is small. For both estimators, weak identification leads the standard error to overestimate the simulation standard deviation.

We then compare the coverage probabilities for the weak-identification robust 95% CI with those obtained from the CRC and restricted GrRC estimators in Table 3. The simulation results show that the weak-identification robust inference CI has coverage close to 95%, regardless of the magnitude of $\mu_{(0,1)} - \mu_{(1,0)}$. In contrast, both the CRC and restricted GrRC tend to over-cover. The simulation results suggest that this over-coverage issue is likely related to the over-estimation of the sampling variance shown in Table 3.

3.6 Weak-identification Robust Inference: Revisiting Suri (2011)

Building on our formal analysis, we revisit our reanalysis of Suri (2011) using the GrRC approach. The unrestricted GrRC estimates help explain the inconsistent CRC results in Table 2. For $T = 2$, the returns to hybrid adoption are similar in magnitude but have opposite signs for joiners ($\Delta_{(0,1)}$) and leavers ($\Delta_{(1,0)}$). This helps explain why the estimated hybrid coefficient for the FE regression is insignificant, as it pools these two switcher subpopulations together. However, the average yield without adoption for these subpopulations ($\mu_{(0,1)}$ and $\mu_{(1,0)}$) is statistically indistinguishable, especially with control variables as in column (2).

As noted in (7) and Proposition 1 *(iii)*, ϕ will suffer from weak-identification issues when this yield difference without adoption is small. This sensitivity to the yield difference

[11] The restriction on the coefficient on $h_{it}1\{h_i = (1,1)\}$ in (10) follows from noting that $\kappa_{(1,1)} - \Delta_{(0,1)} = \mu_{(1,1)} + \Delta_{(1,1)} - \Delta_{(0,1)}$ and using Proposition 1 *(iii)*.

might explain why minor discrepancies in our working data result in disproportionately big differences in our estimates of ϕ, on which so much of the narrative in Suri (2011) rests.

Next, we construct the 95% confidence interval for ϕ for the different specifications we consider using our weak-identification robust inference procedure reported in Table 2. For the majority of the specifications, the upper and lower bounds of the confidence intervals are negative and include values for ϕ similar in magnitude to the point estimates reported in Suri (2011). Our weak-identification robust inference approach thus allows us to reconcile the inference results on the key LCA parameter in Suri (2011).

4 Concluding Remarks

Despite being widely-cited in development economics, most references to Suri (2011) do not discuss her methodological contribution or the selection patterns implied by a negative ϕ. Instead of prompting a nuanced discussion of different forms of heterogeneity, the article is often cited as generic evidence of the simple presence of heterogeneity in the returns to technology adoption. This superficial reading misses an opportunity to inform policies aimed at stimulating technology adoption since optimal program design often hinges more on the specific form of heterogeneous returns than on the mere existence of heterogeneity. We hope that, by highlighting how to detect potential weak-identification concerns and conduct weak-identification robust inference on the key structural parameter of the model, our approach will be useful for applied researchers who wish to study models like the one in Suri (2011).

Our GrRC approach provides additional appealing features, especially when $T > 2$. First, the Suri (2011) approach is cumbersome for multiple periods and may suffer from multicollinearities in the reduced form.[12] Our GrRC approach circumvents this issue by only requiring dummy variables for observed adoption trajectories. Second, the unrestricted GrRC model, unlike the CRC model's reduced form, has an economic interpretation and provides insights into potential identification concerns for ϕ. Finally, relating the Suri (2011) model with the panel identification literature provides alternative identification strategies, such as exchangeability and other nonparametric correlated random effects restrictions (Altonji and Matzkin, 2005; Bester and Hansen, 2009; Ghanem, 2017).

[12] To obtain the reduced form of Suri (2011)'s CRC model, θ_i is projected onto a fully saturated model of h_{it} for all $t = 1, \ldots, T$. Any unobserved adoption history will lead to at least two of the independent variables in this projection becoming collinear.

References

ALTONJI, J. and MATZKIN, R. (2005). Cross-section and panel data estimators for nonseparable models with endogenous regrssors. *Econometrica*, **73** (3), 1053–1102.

ANDREWS, I., STOCK, J. H. and SUN, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, **11** (1), 727–753.

BARRIGA-CABANILLAS, O., MICHLER, J. D., MICHUDA, A. and TJERNSTRÖM, E. (2018). Fitting and Interpreting Correlated Random Coefficient (CRC) Models Using Stata. *Stata Journal*, **18** (1), 159–173.

BESTER, C. A. and HANSEN, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business & Economic Statistics*, **27** (2), 235–250.

CHAMBERLAIN, G. (1984). Panel data. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Elsevier, pp. 1247–1318.

FEDER, G., JUST, R. E. and ZILBERMAN, D. (1985). Adoption of agricultural innovations in developing countries: A survey. *Economic Development and Cultural Change*, **33** (2), 255–98.

FOSTER, A. and ROSENZWEIG, M. (2010). Microeconomics of technology adoption. *Annu. Rev. Econ.*, **2** (1), 395–424.

GHANEM, D. (2017). Testing identifying assumptions in nonseparable panel data models. *Journal of Econometrics*, **197** (2), 202–217.

JACK, B. K. (2013). Market inefficiencies and the adoption of agricultural technologies in developing countries. *CEGA White Paper.*

MAGRUDER, J. R. (2018). An assessment of experimental evidence on agricultural technology adoption in developing countries. *Annual Review of Resource Economics*, **10**, 299–316.

SURI, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica*, **79** (1), 159–209.

VERDIER, V. (2020). Average treatment effects for stayers with correlated random coefficient models of panel data. *Journal of Applied Econometrics*, **35** (7), 917–939.

Table 1: OLS and FE models (Table IIIA in Suri 2011)

Panel A: Suri 2011	OLS			FE	
Hybrid	1.074	0.695	0.541	0.017	0.090
	(0.040)	(0.039)	(0.041)	(0.070)	(0.065)
Panel B: Re-analysis					
Hybrid	1.072	0.692	0.582	0.0139	0.0330
	(0.0457)	(0.0443)	(0.0424)	(0.0696)	(0.0656)
Acres			-0.00995		-0.0675
			(0.00964)		(0.0218)
Seed rate (kg/acre) x 10			0.202		0.180
			(0.0271)		(0.0317)
Land prep (Ksh/acre) x 1000			0.0162		0.0193
			(0.00305)		(0.00511)
Fertilizer (Ksh/acre) x 1000			0.0243		0.0111
			(0.00283)		(0.00385)
Hired labor (Ksh/acre) x 1000			0.0311		0.0238
			(0.00856)		(0.00921)
Family labor (hours/acre) x 1000			0.196		0.238
			(0.0796)		(0.107)
2004	0.538	0.518	0.374	0.483	0.517
	(0.0348)	(0.0333)	(0.0407)	(0.0318)	(0.0535)
N	1197	1197	1197	1197	1197
Province FE	No	Yes	Yes		
Controls	No	No	Yes	No	Yes
Adj. R^2	0.27	0.40	0.48	0.49	0.56

Notes: Dependent variable is ln(yield). Regressions with covariates follow Suri (2011), i.e., in addition to the covariates reported, we include controls for household size, the number of boys (males $<$ age 16), the number of girls (females $<$ age 16), the number of men (aged 17 to 39), the number of women (aged 17 to 39), the number of older men (males $>$ age 40), and main season rainfall. All specifications assume that covariates are exogenous.
OLS specification: $y_{it} = \delta + \beta h_{it} + X_{it}^{'}\gamma + \varepsilon_{it}$
FE specification: $y_{it} = \delta + \beta h_{it} + X_{it}^{'}\gamma + \alpha_p + \epsilon_{ipt}$, for province p

Table 2: CRC and GrRC models (Table VIII A in Suri 2011)

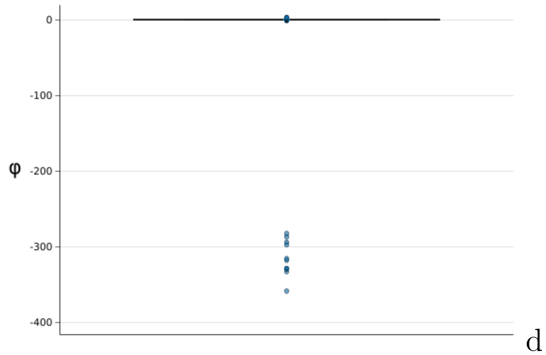
Panel A:	**CRC**					
Suri 2011	Full sample			No HIV districts		
λ_1	0.648	0.565	0.456	0.471	0.305	0.139
	(0.093)	(0.087)	(0.090)	(0.099)	(0.089)	(0.092)
λ_2	1.007	0.665	0.473	1.139	0.710	0.466
	(0.112)	(0.104)	(0.116)	(0.122)	(0.112)	(0.123)
λ_3	1.636	-1.69	-0.485	-4.800	-0.936	-0.497
	(4.854)	(4.316)	(0.199)	(9.173)	(0.308)	(0.257)
β	-0.543	1.0323	3.534	2.287	0.623	0.790
	(1.874)	(1.480)	(24.05)	(4.222)	(0.100)	(0.169)
ϕ	-0.794	-1.317	-17.82	-1.010	-1.518	-2.196
	(0.411)	(1.262)	(137.4)	(0.228)	(0.310)	(1.142)
Panel B:			**CRC**			
Re-analysis						
λ_1	0.716	0.538	0.651	0.630	0.389	0.324
	(0.0761)	(0.0702)	(0.193)	(0.0782)	(0.0704)	(0.185)
λ_2	0.923	0.522	1.110	0.937	0.440	1.068
	(0.108)	(0.0961)	(0.238)	(0.116)	(0.103)	(0.246)
λ_3	-0.334	46.67	-0.00903	-0.297	-0.156	0.193
	(0.405)	(.)	(0.369)	(0.305)	(0.928)	(0.301)
β	0.0195	-24.43	-0.492	-0.0151	0.0872	-0.359
	(0.0842)	(0.0720)	(0.307)	(0.133)	(0.339)	(0.331)
ϕ	0.104	-0.994	-0.0808	-0.0443	-0.207	-0.228
	(0.935)	(0.00121)	(0.381)	(0.605)	(2.876)	(0.238)
Panel C:			**GrRC**			
Re-analysis						
μ_{00}	5.246	4.431	4.502	5.354	4.536	4.517
	(0.0460)	(0.0765)	(0.129)	(0.0546)	(0.0806)	(0.134)
μ_{01}	5.942	4.927	4.999	6.068	4.950	4.907
	(0.0917)	(0.111)	(0.155)	(0.0913)	(0.109)	(0.162)
μ_{10}	6.215	5.274	5.226	6.279	5.240	5.062
	(0.0747)	(0.101)	(0.152)	(0.0784)	(0.102)	(0.156)
$\kappa_{(1,1)}$	6.637	5.524	5.533	6.641	5.420	5.523
	(0.0242)	(0.0835)	(0.0966)	(0.0241)	(0.0872)	(0.0977)
Δ_{01}	0.508	0.370	0.315	0.500	0.345	0.527
	(0.0980)	(0.0910)	(0.183)	(0.104)	(0.0951)	(0.193)
Δ_{10}	-0.476	-0.318	-0.304	-0.520	-0.326	-0.132
	(0.0951)	(0.0920)	(0.182)	(0.0998)	(0.0957)	(0.192)
WIR 95% CI: ϕ	(-20.42, -2.13)	(-4.23, -1.32)	(-57.41, -1.55)	($-\infty$ ∞)	(-7.03, -1.47)	($-\infty$ ∞)
Controls	No	Yes	Yes	No	Yes	Yes
Interactions	No	No	Yes	No	No	Yes

Notes: The dependent variable is ln(yield). Regressions with covariates follow Suri (2011), i.e., include controls for household size, the number of boys (males < age 16), the number of girls (females < age 16), the number of men (aged 17 to 39), the number of women (aged 17 to 39), and the number of older men (males > age 40), the number of maize acres, the seed rate (kg per acre), land preparation expenditures (KShs per acre), fertilizer expenditure (KShs per acre), hired labor (KShs per acre), family labor (hours per acre), and main season rainfall. All specifications assume that covariates are exogenous. Standard errors are clustered at the household level and reported in parentheses. The CRC results in Panels A and B use OMD optimal-weighted minimum distance as in Suri (2011). The unrestricted GrRC results in Panel C show estimates of the average yield without hybrid for the never-adopters and the switcher trajectories (μ_{00}, μ_{01}, and μ_{10}), the average yields for the always-adopters ($\kappa_{(1,1)}$), and the returns to adoption for the switcher trajectories (Δ_{01} and Δ_{10}). In Panel C, we report weak-identification robust 95% confidence intervals for the LCA parameter, ϕ. We construct these using a grid search over $[-5 * 10^4, 5 * 10^4]$ with 0.01 increments.

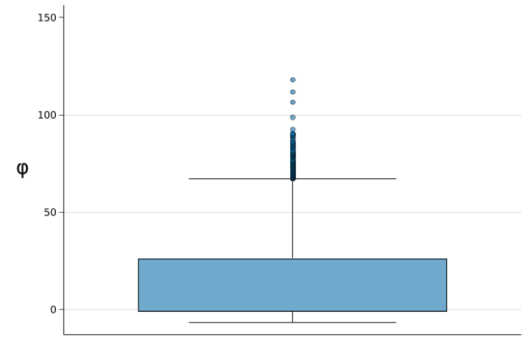
Table 3: CRC and Restricted GrRC Point Estimation and Weak-identification Robust Inference on ϕ ($\phi = -0.5$)

$\mu_{(0,1)} - \mu_{(1,0)}$	CRC							Restricted GRC							WIR CI
	Mean	Median	SD	MAE	RMSE	SE/SD	Cov.	Mean	Median	SD	MAE	RMSE	SE/SD	Cov.	Cov.
$n = 1,000$															
0.1	2.23	-0.15	6.20	2.95	6.78	10.10	0.939	-0.35	-0.48	14.60	2.96	14.60	19.00	0.999	0.950
0.25	0.79	-0.36	4.08	1.53	4.28	6.28	0.945	-0.16	-0.41	8.58	1.46	8.59	8.57	0.998	0.940
0.5	-0.31	-0.48	1.14	0.40	1.16	1.76	0.969	-0.33	-0.48	1.94	0.36	1.95	1.55	0.991	0.945
1	-0.48	-0.50	0.20	0.15	0.20	1.19	0.985	-0.49	-0.50	0.17	0.13	0.17	0.98	0.963	0.947
$n = 2,000$															
0.1	1.94	-0.15	5.39	2.65	5.92	8.20	0.979	-0.87	-0.44	24.90	3.21	24.90	20.20	0.999	0.940
0.25	0.17	-0.45	2.74	0.90	2.82	3.95	0.981	-0.42	-0.47	6.64	0.87	6.64	8.56	0.997	0.950
0.5	-0.45	-0.49	0.33	0.22	0.34	1.17	0.986	-0.46	-0.49	0.28	0.20	0.28	0.97	0.977	0.944
1	-0.49	-0.50	0.13	0.10	0.13	1.23	0.984	-0.49	-0.50	0.12	0.09	0.12	1.00	0.963	0.954
$n = 5,000$															
0.1	1.07	-0.30	3.77	1.79	4.08	5.84	0.998	-0.50	-0.42	11.70	1.66	11.70	9.52	0.999	0.944
0.25	-0.37	-0.48	0.73	0.33	0.74	1.49	0.995	-0.42	-0.49	1.27	0.28	1.27	1.63	0.990	0.955
0.5	-0.49	-0.50	0.17	0.13	0.17	1.19	0.986	-0.49	-0.50	0.15	0.12	0.15	1.01	0.963	0.950
1	-0.50	-0.50	0.08	0.06	0.08	1.24	0.985	-0.50	-0.50	0.07	0.06	0.07	1.00	0.951	0.948

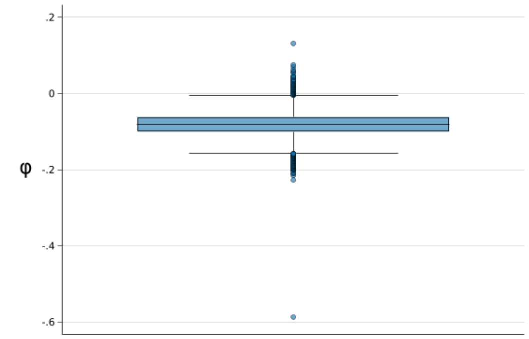
Notes: The above table presents simulation statistics for the estimators of ϕ as well as the coverage of the weak-identification robust confidence interval ($WIR\ CI$). SD, MAE, RMSE and SE/SD abbreviate the simulation standard deviation, median absolute error, root-mean squared error, and the ratio of the average standard error to the simulation standard deviation, respectively. Cov. abbreviates coverage probability for a 95% confidence interval. The summary statistics are computed using 5,000 simulation replications. The outcome in our design is given by $y_{it} = \mu_i + (\beta + \phi\theta_i)h_{it} + u_{it}$ for $i = 1, \ldots, n$, $t = 1, 2$, where $\theta_i = \mu_i - E[\mu_i]$, $\mu_i|(h_{i1}, h_{i2}) \overset{i.i.d.}{\sim} N(\mu_{(h_{i1},h_{i2})}, \sigma_\mu^2)$, $u_{it}|(h_{i1}, h_{i2}) \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$. We set $\beta = 0.25$, $\mu_{(0,0)} = 1$, $\mu_{(0,1)} = 0$, $\mu_{(1,0)} = -\eta$, $\mu_{(1,1)} = 3$, $\sigma_\mu^2 = 0.25$, $\sigma_u^2 = 1$, $\pi_{(0,1)} + \pi_{(1,0)} = 0.2$, $\pi_{(0,1)} = \pi_{(1,0)}$, $\pi_{(0,0)} = \pi_{(1,1)}$. For the CRC and Restricted GrRC, the simulation results are based on the replications where the estimator in question converges. Since $WIR\ CI$ is based on closed-form estimators, we report its simulation coverage probability across all simulations.



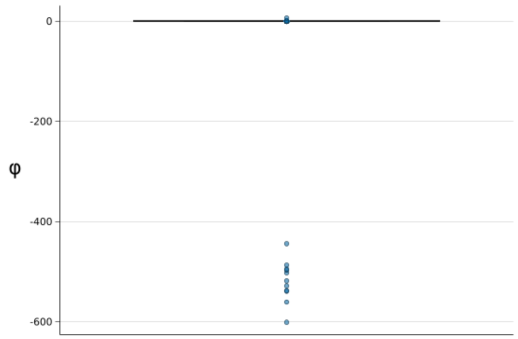
(a) Full sample; no controls



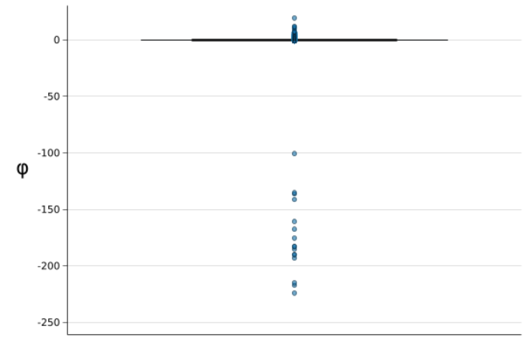
(b) Full sample; base controls



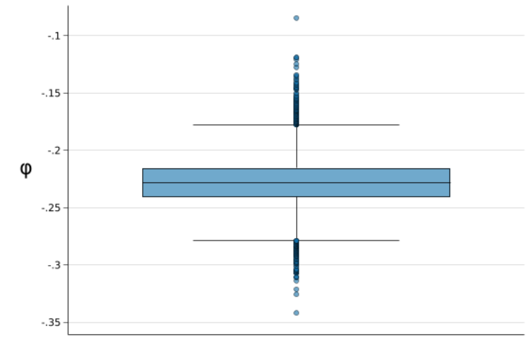
(c) Full sample; interaction controls



(d) No HIV districts; no controls



(e) No HIV districts; base controls



(f) No HIV districts; interaction controls

Figure 1: Sensitivity of CRC Estimates of ϕ

Notes: The above box plot shows the distribution of the ϕ parameter estimated from the random coefficient model as developed in Suri (2011) when ten households are removed from the sample at random across 10,000 simulation replications. We use the `randcoef` Stata command for estimation (Barriga-Cabanillas *et al.*, 2018). The dependent variable is ln(yield). Regressions with covariates follow Suri (2011), i.e., include controls for household size, the number of boys (males $<$ age 16), the number of girls (females $<$ age 16), the number of men (aged 17 to 39), the number of women (aged 17 to 39), and the number of older men (males $>$ age 40), the number of maize acres, the seed rate (kg per acre), land preparation expenditures (KShs per acre), fertilizer expenditure (KShs per acre), hired labor (KShs per acre), family labor (hours per acre), and main season rainfall. All specifications assume that covariates are exogenous.

Online Appendix

A Data appendix

Our data processing follows the original author's instructions, applying them to the publicly available data from Tegemeo Institute. The summary statistics in Table A.1 compare key variables across our dataset and Suri (2011), as well as across years. Overall, we find similar means and distributions, with some minor differences. Our balanced panel has 1,197 households compared to 1,202 in Suri (2011), due to missing values for control variables.[13]

The biggest differences appear in the 2004 data for fertilizer application rates (`Fertilizer`) and labor variables (`Hired labor` and `Family labor`). This suggests that Tegemeo Institute may have further cleaned the 2004 data after sharing it with the original author, while the 1997 data set was already finalized at the time of sharing.

However, the 2004 fertilizer expenditures variable deserves some additional discussion due to the larger discrepancies. In the data, district-level fertilizer prices contain missing values. The data documentation suggests replacing missing values with the median price for that fertilizer type. For districts with insufficient observations, we use the sample median for that fertilizer type.

In the 2004 data, a few other complications arise. The documentation mentions several data files that have different names in the open access data.[14] Fertilizer prices also appear in the household-level dataset `hh04`, where they are elicited as in 1997. Merging fertilizer prices on district, fertilizer type, and fertilizer unit, as suggested in the data documentation, is challenging due to differences in coding between `fert04` (field-level) and `tfert04` (household-level). Some discrepancies are minor, while others involve significant numbers of observations, such as 80 fields reporting fertilizer use in *gorogoros* (roughly 2 kg) but reporting prices in another unit.[15] We address this by computing median prices at the district-fertilizer type-fertilizer unit level, converting these to per-kilo prices, and merging them onto field-level data by household id.[16] This discrepancy may explain some differences between our dataset and Suri (2011).

[13] We lose two households due to missing labor variables and three to missing household head education.

[14] The dataset `pricefert` does not exist in the open access data. Instead, the dataset `tfert04` contains household- and fertilizer-type-level fertilizer price data. We assume this is the relevant dataset and compute district- and fertilizer-type-level prices based on the variable `inputpr`.

[15] Another 151 households report fertilizer prices in 5, 10, or 25 kg bags, a unit that is absent in `fert04`.

[16] For households with multiple purchases of a given input type, we use the mean of the per-kilo prices.

Table A.1: Comparison of key variables between our dataset and Suri (2011)

Panel A: Suri 2011	1997 Sample		2004 Sample	
	Mean	(s.d.)	Mean	(s.d.)
Yield (log maize harvest per acre)	5.907	(1.153)	6.350	(0.977)
Acres planted	1.903	(3.217)	1.957	(2.685)
Total seed planted (kg per acre)	9.575	(7.801)	9.072	(6.863)
Hybrid (dummy)	0.658	(0.475)	0.604	(0.489)
Total fertilizer expenditure (KShs per acre)	1361.7	(2246.3)	1354.6	(1831.2)
Land preparation costs (KShs per acre)	960.88	(1237.1)	541.43	(1022.8)
Family labor (hours per acre)	293.25	(347.49)	354.27	(352.68)
Hired labor (KShs per acre)	1766.0	(3346.4)	1427.4	(2130.3)
Main season rainfall (mm)	620.83	(256.43)	728.11	(293.29)
Household size	7.109	(2.671)	8.409	(3.521)
Panel B: TGMBLM 2024	1997 Sample		2004 Sample	
	1997 mean	sd	2004 mean	sd
Yield (log maize harvest per acre)	5.907	(1.155)	6.389	(0.979)
Acres planted	1.908	(3.223)	1.959	(2.689)
Total seed planted (kg per acre)	10.11	(8.361)	10.92	(8.092)
Hybrid (dummy)	0.657	(0.475)	0.605	(0.489)
Total fertilizer expenditure (KShs per acre)	1426.3	(2394.7)	3182.3	(6610.2)
Land preparation costs (KShs per acre)	2169.3	(4905.1)	1182.2	(1649.9)
Family labor (hours per acre)	292.4	(347.7)	560.9	(767.6)
Hired labor (KShs per acre)	1661.3	(3219.9)	1992.4	(2962.2)
Main season rainfall (mm)	620.9	(256.5)	728.5	(293.4)
Household size	7.107	(2.669)	8.349	(3.480)

Notes: This table corresponds to Table IIA in Suri (2011), reporting the subset of variables used in estimation. KShs denotes Kenyan shillings. All monetary variables are in real terms.