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

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

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

This version: April 2023

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), which we clean following 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 replicate 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 access, code, and helpful comments.

1 Introduction

In an influential paper, Suri (2011) addresses a long-standing development puzzle: why do many sub-Saharan African farmers continue to use traditional farming techniques when modern agricultural technologies portend higher returns? A large literature has proposed explanations that center on market frictions preventing farmers from adopting profitable technologies, such as credit constraints, uninsured risk, and incomplete information.[1] Suri (2011) instead provides an explanation that focuses on the role of heterogeneity in the returns to adoption, which she models as stemming from time-invariant unobservables.

Central to the Suri (2011) approach is a specific restriction on the form of heterogeneous returns. The Suri (2011) model posits that farmers make technology adoption decisions based on their unobservable comparative advantage. It further imposes a key assumption, which we call the Linearity in Comparative Advantage (LCA) restriction: the returns to hybrid adoption must be linear in comparative advantage. The key structural parameter in this correlated random coefficient (CRC) model is the slope parameter in this linear relationship (ϕ), which we will refer to as the LCA parameter.

Using this assumption, Suri (2011) estimates a CRC model on a panel dataset of Kenyan farm households that grow either hybrid or non-hybrid maize. She uses a linear projection of an individual's returns to adoption of the hybrid seed technology onto their observed adoption history, building on the correlated random effects (CRE) approach in Chamberlain (1984). This allows her to recover the LCA parameter (ϕ), the sign of which influences how we interpret the results. She finds negative and statistically significant estimates of the LCA parameter across her specifications, indicating that farmers who have the lowest non-hybrid productivity stand to reap the highest productivity gains from switching to hybrids. Using the LCA restriction, Suri (2011) then extrapolates the estimated returns to non-adopters in the sample.

In our reanalysis, we are unable to replicate the negative and statistically significant LCA parameter estimates in Suri (2011) using her CRC estimation approach. An array of prescriptions and interventions aim to encourage precisely these non-adopters to adopt productivity-enhancing technologies. Reliable parameter estimates from a model like Suri (2011) can potentially adjudicate among these competing options. Correct inference about these parameters therefore has direct practical and policy relevance.

We therefore investigate why our results differ from the original paper and uncover a

[1] For surveys of this vast literature, see Feder *et al.* (1985); Foster and Rosenzweig (2010); Magruder (2018); Jack (2013).

potential for weak identification of the LCA parameter in the CRC model. We show that the CRC model in Suri (2011) is a restricted version of a group random coefficient (GrRC) model, specifically imposing the LCA restriction. The unrestricted version of the GrRC model highlights the nature of the weak-identification problem. As a result, researchers can use the unrestricted GrRC to detect potential weak identification of the LCA parameter in practice.

Next, we propose a weak-identification robust procedure for inference on the LCA parameter. Our procedure is based on test inversion of the LCA restriction on the parameters of the (unrestricted) GrRC model. We illustrate the weak identification problem and the finite-sample performance of our weak-identification robust inference procedure using a small-scale simulation study.

Finally, we revisit the reanalysis of Suri (2011) using our GrRC approach. Our unrestricted GrRC estimates clearly raise concerns regarding weak identification of the LCA parameter. The weak-identification robust confidence interval we construct for this parameter yields negative intervals across multiple specifications; further, the intervals include values for the LCA parameter (ϕ) that are similar in magnitude to the point estimates reported in Suri (2011). In sum, once when we account for the weak-identification issue, we are able to replicate the main result in Suri (2011).

In the next section, we briefly describe our reanalysis attempt. Section 3 presents the GrRC model for the two-period case and, after imposing the LCA restriction of the CRC model, introduces the weak-identification robust inference procedure for the LCA parameter. Simulations illustrate that the proposed weak-identification robust inference provides good coverage probabilities in finite samples. Section 4 discusses the relevance of our findings for the broader questions surrounding technology adoption in low-income agriculture.

2 Reanalysis of Suri (2011)

2.1 Data

We use the same panel dataset on rural Kenyan households as Suri (2011), which was collected and provided by the Tegemeo Institute of Agricultural Policy and Development Institute.[2] Appendix C documents the steps that we take to ensure our data are as close as possible to the dataset used in the original paper. The appendix also compares summary

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

statistics across the two datasets. Our data and the original dataset are very similar, with a few minor exceptions: despite obtaining the data directly from the same source as Suri (2011) and following the original author's data cleaning documentation, our reconstructed sample consists of 1,197 households instead of the 1,202 in Suri (2011). Furthermore, minor discrepancies in the summary statistics suggest that there exist other differences, either in the composition of these roughly 1,200 households or in the way that variables were constructed, imputed, or cleaned.

Throughout, we construct variables following Suri (2011). We define maize yield as the ratio of (self-reported) maize harvest to the plot size.[3] We similarly construct technology adoption trajectories and fertilizer use based on whether households report using a hybrid maize seed or fertilizer, respectively, in the relevant season. In addition, we report specifications that control for the same set of demographic and agricultural production variables as Suri (2011).

2.2 Reanalysis Results

Table 1 reports the descriptive OLS and fixed effects (FE) specifications that Suri (2011) uses to introduce her analysis. These regressions provide an estimate of the average yield advantages of hybrid maize compared to non-hybrid maize varieties under the assumption that returns are homogeneous. Panel A shows the point estimates from the original paper, while Panel B shows our reanalysis.[4] Our results are statistically indistinguishable from Suri (2011).

The crux of our reanalysis is the estimation of the CRC model and of the LCA parameter in particular. We offer a concise presentation of this model here before discussing our results (for a more detailed exposition, see Suri (2011)). This CRC model is given by:

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

where y_{it} is log of maize yield for household i at time t; τ_i is farmer i's absolute advantage, assumed to be (mean) independent of technology adoption (i.e., $E[\tau_i|h_i] = E[\tau_i]$); $h_i = (h_{i1}, \dots, h_{iT})$ denotes individual i's adoption history with each h_{it} indicating whether farmer i adopted hybrid seeds at time t; $\epsilon_{it} \sim \mathcal{N}(0, \sigma_u^2)$ is an idiosyncratic error term, and θ_i can

[3] Note that this is not a measure of economic returns, as the variable is not net of input costs; however, we maintain the assumption in Suri (2011) that input costs, other than hybrid seeds, are constant across hybrid and non-hybrid sectors.

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

be correlated with the adoption history. To estimate the different means across different adoption history in the CRC model following Chamberlain (1984), a linear projection of θ_i onto a fully saturated model of adoption histories is used:

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

We call the restriction on response heterogeneity in this model—i.e., the assumption that the returns to hybrid adoption are linear in comparative advantage—the LCA restriction and therefore refer to ϕ as the LCA parameter.

We show the results for this CRC model in Table 2, which we estimate using the `Stata` package from Barriga-Cabanillas *et al.* (2018). Even though we use data that are nearly identical to Suri's (2011), we are only able statistically to replicate the estimates for λ_1 and λ_2. Our estimates of ϕ 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, especially since it can only be attributed to seemingly trivial differences in the working data we construct from the Tegemeo panel dataset. In order to explore why minor data differences could have such disproportionate estimation impacts, we introduce a more general model that includes the CRC model in Suri (2011) as a special case.

3 A Group Random Coefficient (GrRC) Alternative

In this section, 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 group random coefficient (GrRC) model. This model allows us to show the potential source of weak identification of this parameter and 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}, which is given by the following, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

$$y_{it} = f(h_{it}, a_i) + \varepsilon_{it}. \tag{3}$$

In this paper, as in Suri (2011), we maintain the strict exogeneity assumption, i.e. $E[\varepsilon_{it}|h_i, a_i] = 0$, where $h_i = (h_{i1}, \ldots, h_{iT})$ denotes farmer i's adoption history and each h_{it} is a binary indicator of adoption. We impose no restriction on the distribution of farmer ability (a_i) conditional on adoption history (h_i), thereby treating a_i as a "fixed effect."[5] For simplicity, we consider a model without covariates, but our results extend to the inclusion of additively separable covariates as in (1).

Since h_{it} is a binary variable, we can express the above equation equivalently as a random coefficient model by letting $\mu_i \equiv f(0, a_i)$ and $\Delta_i \equiv f(1, a_i) - f(0, a_i)$,

$$y_{it} = \mu_i + \Delta_i h_{it} + \varepsilon_{it}. \tag{4}$$

In our empirical context, μ_i denotes farmer i's expected (log) yield without adoption and Δ_i the returns to hybrid maize adoption. This model nests the CRC model of Suri (2011) as a special case. To see this, let $\mu_i = \tau_i + \theta_i$ and $\Delta_i = \beta + \phi\theta_i$,

$$y_{it} = \tau_i + \theta_i + (\beta + \phi\theta_i)h_{it} + \varepsilon_{it}, \tag{5}$$

which is the CRC model without covariates. Selection into technology adoption is determined by θ_i, farmer i's comparative advantage, which admits the normalization $E[\theta_i] = 0$.

3.2 Group Random Coefficient (GrRC) Model

Applying the GrRC approach to this context relies on the insight that with a binary variable and fixed T there is a finite number of adoption histories. We denote the realization of an adoption history h_i by $\underline{h} = (h_1, \ldots, h_T) \in \mathcal{H} = \{0,1\}^T$, the set of switcher trajectories $\mathcal{H}_S = \{\underline{h} \in \mathcal{H}_S : 0 < \sum_{t=1}^{T} h_{it} < T\}$, and the set of stayer trajectories $\mathcal{H}_S^c = \mathcal{H} \backslash \mathcal{H}_S$. Since Suri (2011) considers the case of $T = 2$ and uses a two-period panel data set, we illustrate the GrRC alternative for $T = 2$, but the results extend to any $T < \infty$.

With $T = 2$, h_i can take four possible values; formally, its support is given by $\mathcal{H} = \{0,1\}^2 = \{(0,0), (0,1), (1,0), (1,1)\}$. Since the adoption histories may entail different distributions of ability, it is natural to define subpopulations in terms of adoption histories. Suri (2011) refers to the four subpopulations in the two-period case as never-adopters, joiners,

[5] As per arguments in Chernozhukov *et al.* (2013), $f(h_{it}, a_i)$ may be viewed as the conditional mean function of a fully nonseparable model, $\phi(h_{it}, a_i, u_{it})$, where we assume time homogeneity, i.e. $u_{it}|h_i, a_i \overset{d}{=} u_{i1}|h_i, a_i$.

leavers, and always-adopters. As a result, $\mathcal{H}_S = \{(0,1),(1,0)\}$, the set of switcher subpopulations, consists of joiners and leavers, respectively. Its complement set, $\mathcal{H}_S^c = \{(0,0),(1,1)\}$, is composed of the two stayer subpopulations, never-adopters and always-adopters, respectively.

We are interested in how the average returns to adoption varies across the different subpopulations, and therefore integrate the unrestricted random coefficient model (4) with respect to $a_i|h_i$, which yields the following conditional mean model under strict exogeneity, $E[\varepsilon_{it}|h_i, a_i] = 0$,

$$E[y_{it}|h_i = \underline{h}] = \mu_{\underline{h}} + \Delta_{\underline{h}}\underline{h}_t, \tag{6}$$

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 $t = 1,\ldots,T$.

By the time homogeneity of $\mu_{\underline{h}}$ and $\Delta_{\underline{h}}$, we can identify the average return to adoption for subpopulations that we observe with and without hybrid adoption in our data. Hence, $\Delta_{\underline{h}}$ is only nonparametrically identified for the switcher subpopulations, $\underline{h} \in \mathcal{H}_S$. For stayer subpopulations, we can either identify their average yield *with* or *without* adoption.[6] Specifically, for the never-adopters, we can identify the average yield without adoption, $\mu_{(0,0)}$, while we can only identify the average yield with adoption for the always-adopters. We denote the latter average by $\kappa_{(1,1)} = \mu_{(1,1)} + \Delta_{(1,1)}$. Without further restrictions, we cannot separately identify $\mu_{(1,1)}$ and $\Delta_{(1,1)}$. Hence, the returns to adoption are not nonparametrically identified for the stayer subpopulations.

All of the aforementioned identifiable objects can be estimated consistently using the following GrRC model,

$$y_{it} = \sum_{\underline{h} \in \mathcal{H}\setminus(1,1)} \mu_{\underline{h}} 1\{h_i = \underline{h}\} + \sum_{\underline{h} \in \mathcal{H}_S} \Delta_{\underline{h}} h_{it} 1\{h_i = \underline{h}\} + \kappa_{(1,1)} h_{it} 1\{\underline{h} = (1,1)\} + \varepsilon_{it}. \tag{7}$$

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.

[6] Since we are referring to the average yield of a subpopulation, it is by definition the expected yield of this subpopulation. We prefer the more intuitive term "average yield" to "expected yield" and use it hereinafter to refer to the latter.

3.3 Identifying the LCA Parameter

Next, we impose the LCA restriction on the GrRC model and illustrate how the unrestricted model can indicate potential identification concerns for ϕ, the LCA parameter. To this end, we first establish the relationship between parameters in the unrestricted GrRC model and those in the Suri (2011) model in the following proposition.

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). We provide the proof of Proposition 1 in Appendix A. From *(iii)* in Proposition 1, we can re-write ϕ as the ratio of difference in returns to adoption for different subpopulations to the difference in their comparative advantage assuming $\mu_{\underline{h}} \neq \mu_{\underline{h}'}$. Since we can identify both $\mu_{\underline{h}}$ and $\Delta_{\underline{h}}$ for switcher subpopulations, ϕ is identified as the following in the two-period case as long as $\mu_{(1,0)} \neq \mu_{(0,1)}$,

$$\phi = \frac{\Delta_{(1,0)} - \Delta_{(0,1)}}{\mu_{(1,0)} - \mu_{(0,1)}}. \tag{8}$$

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.[7] Since the unrestricted GrRC model enables us to estimate both parameters without imposing the LCA restriction, the unrestricted GrRC model plays a similar role to the first stage of an instrumental variables (IV) regression in terms of detecting potential identification concerns, which we illustrate in Section 3.6. Unlike the first-stage of an IV, however, the unrestricted GrRC has an economic interpretation

[7] 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 $\beta = \Delta_{(0,1)} - \phi\theta_{(0,1)} = \Delta_{(1,0)} - \phi\theta_{(1,0)}$.

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 is the case in our empirical application (see Section 3.6). Therefore, we use the restrictions on the LCA parameter obtained from Proposition 1 to conduct weak-identification robust inference on this key structural parameter. We specifically propose a weak-identification robust confidence interval for ϕ based on inverting $W_n(\phi_0)$, the Wald statistic, of

$$H_0 : \Delta_{\underline{h}} - \Delta_{\underline{h}'} = \phi_0 \left(\mu_{\underline{h}} - \mu_{\underline{h}'}\right), \text{for } \underline{h} \in \mathcal{H}^S, \underline{h}' \in \mathcal{H}^S, \underline{h} \neq \underline{h}'. \tag{9}$$

Assuming sufficient regularity conditions such that $W_n(\phi_0) \underset{H_0:\phi=\phi_0}{\overset{d}{\to}} \chi^2_{|\mathcal{H}^S|-1}$, the $(1-\alpha)\%$ confidence interval is defined as,

$$C_\alpha = \{\phi_0 \in \Phi : W_n(\phi_0) < c_{\alpha,|\mathcal{H}^S|-1}\} \tag{10}$$

where Φ is a compact 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, computing the confidence interval is straightforward using a fine grid search.[8]

3.5 Simulations

A small-scale simulation study illustrates the weak-identification concern and the weak-identification-robust inference procedure that we propose. We consider a simple two-period model that satisfies the LCA restriction described in Table 3. In addition to the CRC estimator and the weak-identification robust inference procedure, we also consider the GrRC

[8] With more than two switcher subpopulations, as in the $T > 2$ case, it is possible for the confidence interval to be empty if the over-identifying restrictions are violated. For a discussion of similar issues in the context of weak-IV robust inference see Andrews *et al.* (2019).

model with the LCA restriction (i.e., the restricted GrRC),[9]

$$y_{it} = \sum_{\underline{h} \in \mathcal{H} \setminus (1,1)} \mu_{\underline{h}} + \Delta_{(0,1)} h_{it} + \phi(\mu_{(1,0)} - \mu_{(0,1)}) h_{it} 1\{h_i = (1,0)\}$$
$$+ \left(\mu_{(1,1)} + \phi\left(\mu_{(1,1)} - \mu_{(0,1)}\right)\right) h_{it} 1\{h_i = (1,1)\} + \varepsilon_{it}. \quad (11)$$

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

Table 4 presents simulation summary statistics that illustrate how the difference between $\mu_{(0,1)}$ and $\mu_{(1,0)}$, when small, can lead to a weak-identification problem for ϕ. Both the CRC and restricted GrRC estimators suffer from severe bias that is exacerbated when the difference between $\mu_{(0,1)}$ and $\mu_{(1,0)}$ is small. In addition, for both estimators, weak identification leads the standard error to overestimate the simulation standard deviation.

Finally, we compare the coverage probabilities for the weak-identification robust 95% CI with those obtained from the CRC and restricted GrRC estimators in Table 5. The simulation results illustrate that the coverage of the weak-identification robust inference CI is generally close to 95%, regardless of the magnitude of $\mu_{(0,1)} - \mu_{(1,0)}$. In contrast, both the CRC and restricted GrRC tend to overcover. The simulation results suggest that this over-coverage issue is likely related to the over-estimation of the variance shown in Table 4.

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 $T = 2$ (Table 7), 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 $T = 2$ FE regression is insignificant, as it pools together these two switcher subpopulations. However, the average yield without adoption for these two subpopulations ($\mu_{(0,1)}$ and $\mu_{(1,0)}$) is statistically indistinguishable, especially when we add control variables in column (2). As noted in (8) and Proposition 1 *(iii)*, ϕ will suffer from a weak-identification issue when this yield difference without adoption is small. Furthermore, the fact that identification is so sensitive to this yield difference may explain why very small discrepancies in our working data result in disproportionately big differences in our estimates

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

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 7. With the exception of one specification, which does not include control variables, the upper and lower bounds of the resulting confidence intervals are negative and include values for ϕ that are similar in magnitude to the point estimates reported in Suri (2011). As a result, our weak-identification robust inference approach allows us to replicate the inference results on this key LCA parameter in Suri (2011).

4 Concluding Remarks

Despite being well-known and widely-cited in development economics, Suri (2011) has had surprisingly limited methodological impact: rather than leading to widespread use of the CRC method and to nuanced discussion in subsequent empirical work of the specific form of heterogeneity it implies, the article is largely cited as generic evidence of heterogeneity in the returns to new technologies. Such a superficial reading of the article misses an important opportunity to inform policies and interventions aimed at stimulating technology adoption since optimal program design often hinges on the specific form of heterogeneous returns rather than the mere existence of such heterogeneity. We aim to revive the methodological contribution of Suri (2011) by proposing an approach that allows empirical researchers to detect potential weak-identification concerns as well as conduct weak-identification robust inference on the key structural parameter of the model.

In addition to allowing us to address the weak-identification issues, the GrRC approach provides several appealing features, especially when $T > 2$. First, the Suri (2011) approach to estimating the CRC model is cumbersome to adapt to the multiple-period case due to multicollinearities that arise in the reduced form whenever some adoption histories are unobserved in a given dataset.[10] Since the regressors in our GrRC approach consist of dummy variables for the adoption histories, this issue is circumvented by the inclusion of dummy variables for the observed trajectories. Second, the unrestricted GrRC model, unlike the reduced form of the CRC model, has an economic interpretation and provides the practitioner with insights on potential identification concerns pertaining to the parameter ϕ. Finally, relating the Suri (2011) model with the panel identification literature provides alternative

[10] To obtain the reduced form of the CRC model in Suri (2011), θ_i is projected onto a fully saturated model of h_{it} for all $t = 1, \ldots, T$. As soon as any adoption history is unobserved, then at least two of the independent variables in this projection become collinear.

identification strategies, such as exchangeability and other nonparametric correlated random effects restrictions (Altonji and Matzkin, 2005; Bester and Hansen, 2009; Ghanem, 2017).

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.

Chernozhukov, V., Fernandez-Val, I., Hahn, J. and Newey, W. (2013). Average and quantile effects in nonseparable panel data models. *Econometrica*, **81** (2), pp.535–580.

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.

— (2018). Data documentation for Suri 2011.

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***	0.530***	0.0139	0.0319
	(0.0457)	(0.0443)	(0.0425)	(0.0696)	(0.0661)
Acres			-0.00906		-0.0638***
			(0.00964)		(0.0219)
Seed rate (kg/acre) x 10			0.203***		0.177***
			(0.0275)		(0.0310)
Land prep (Ksh/acre) x 1000			0.0169***		0.0191***
			(0.00313)		(0.00520)
Fertilizer (Ksh/acre) x 1000			0.0223***		0.0113***
			(0.00274)		(0.00389)
Hired labor (Ksh/acre) x 1000			0.0314***		0.0239***
			(0.00844)		(0.00925)
Family labor (hours/acre) x 1000			0.193**		0.243**
			(0.0788)		(0.107)
2004	0.538***	0.518***	0.408***	0.483***	0.447***
	(0.0348)	(0.0333)	(0.0369)	(0.0318)	(0.0443)
Observations	1197	1197	1197	1197	1197
$N \times T$	2394	2394	2394	2394	2394
District FE	No	Yes	Yes		
Controls	No	No	Yes	No	Yes
Adj. R^2	0.27	0.40	0.49	0.49	0.56

Notes: Dependent variable is ln(yield). Covariates follow Suri (2011). All regressions include acreage, land preparation costs, fertilizer, hired labor, family labor, main season rainfall, household size and age-sex composition of the household (includes indicator variables for the number of boys (aged<16 years), the number of girls, the number of men (aged 17–39), the number of women, and the number of older men (aged>40 years). OLS specification: $y_{it} = \delta + \beta h_{it} + X_{it}^{'}\gamma + \varepsilon_{it}$ FE specification: $y_{it} = \delta + \alpha_i + \beta h_{it} + X_{it}^{'}\gamma + \epsilon_{it}$

Table 2: CRC results replication (Table VIIIA in Suri 2011)

	Full sample			No HIV districts		
λ_1	0.716***	0.564***	0.760***	0.630***	0.393***	0.352*
	(0.0761)	(0.0701)	(0.194)	(0.0782)	(0.0703)	(0.187)
λ_2	0.923***	0.527***	0.930***	0.937***	0.435***	0.908***
	(0.108)	(0.0958)	(0.240)	(0.116)	(0.102)	(0.247)
λ_3	-0.334	26.79	0.0318	-0.297	-0.178	0.218
	(0.405)	(.)	(0.993)	(0.305)	(0.988)	(0.379)
β	0.0195	-13.98***	-0.388	-0.0151	0.0937	-0.246
	(0.0842)	(0.0719)	(0.442)	(0.133)	(0.318)	(0.358)
ϕ	0.104	-0.990***	-0.0855	-0.0443	-0.150	-0.236
	(0.935)	(0.00209)	(1.034)	(0.605)	(3.591)	(0.319)
Observations	1197	1197	1197	1057	1057	1057
Controls	No	Yes	Yes	No	Yes	Yes
Interactions	No	No	Yes	No	No	Yes

Notes: We follow the specifications of Suri (2011). Structural coefficients reported are the average return to hybrid (β), the comparative advantage coefficient (ϕ), and the coefficients (λ_1 and λ_2). Following the original paper, we use OMD as the optimal weighting matrix for the minimum-distance procedure. Results for the no-HIV districts omit two districts where HIV was prevalent, following Suri (2011). All specifications with covariates assume that all covariates are exogenous. Covariates follow Suri (2011). All regressions include acreage, land preparation costs, fertilizer, hired labor, family labor, main season rainfall, household size and age-sex composition of the household (includes indicator variables for the number of boys (aged<16 years), the number of girls, the number of men (aged 17–39), the number of women, and the number of older men (aged>40 years).

Table 3: Simulation Design

Outcome Equation	$y_{it} = \mu_i + (\beta + \phi\theta_i)h_{it} + u_{it}$ for $i = 1, \ldots, n$, $t = 1, 2$.
Unobservables	$\mu_i \vert (h_{i1}, h_{i2}) \overset{i.i.d.}{\sim} N(\mu_{(h_{i1},h_{i2})}, \sigma_\mu^2)$, $\theta_i = \mu_i - E[\mu_i]$, $u_{it} \vert (h_{i1}, h_{i2}) \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$.
Subpopulations	$\pi_{(0,1)} + \pi_{(1,0)} = 0.2$, $\pi_{(0,1)} = \pi_{(1,0)}$, $\pi_{(0,0)} = \pi_{(1,1)}$.

Notes: Before generating the unobservables and the outcome model, we randomly assign each unit i the trajectory using a uniform random variable with the proportions given above, where we define $\pi_{(h_1,h_2)} = P((h_{i1}, h_{i2}) = (h_1, h_2))$. We fix $\beta = 0.25$, $\sigma_u^2 = 1$, $\sigma_\mu^2 = 0.25$, $\mu_{(0,0)} = 1$, $\mu_{(0,1)} = 0$, $\mu_{10} = -\eta$, and $\mu_{(1,1)} = 3$. By varying $\eta = \mu_{(0,1)} - \mu_{(1,0)}$, we vary the degree of weak identification in our design.

Table 4: CRC and Restricted GMM Point Estimation of ϕ ($\phi = -0.5$)

$\mu_{(0,1)} - \mu_{(1,0)}$	CRC						Restricted GRC					
	Mean	Median	SD	MAE	RMSE	SE/SD	Mean	Median	SD	MAE	RMSE	SE/SD
$n = 1,000$												
0.1	2.23	-0.15	6.20	2.95	6.78	10.10	-0.35	-0.48	14.60	2.96	14.60	19.00
0.25	0.79	-0.36	4.08	1.53	4.28	6.28	-0.16	-0.41	8.58	1.46	8.59	8.57
0.5	-0.31	-0.48	1.14	0.40	1.16	1.76	-0.33	-0.48	1.94	0.36	1.95	1.55
1	-0.48	-0.50	0.20	0.15	0.20	1.19	-0.49	-0.50	0.17	0.13	0.17	0.98
$n = 1,500$												
0.1	2.27	-0.14	6.18	2.98	6.77	9.25	-0.83	-0.47	16.50	2.76	16.50	14.40
0.25	0.43	-0.40	3.30	1.15	3.43	4.66	-0.18	-0.42	8.68	1.11	8.68	11.20
0.5	-0.41	-0.49	0.60	0.27	0.61	1.17	-0.45	-0.49	0.36	0.23	0.36	0.99
1	-0.50	-0.50	0.15	0.12	0.15	1.21	-0.50	-0.51	0.14	0.11	0.14	0.98
$n = 2,000$												
0.1	1.94	-0.15	5.39	2.65	5.92	8.20	-0.87	-0.44	24.90	3.21	24.90	20.20
0.25	0.17	-0.45	2.74	0.90	2.82	3.95	-0.42	-0.47	6.64	0.87	6.64	8.56
0.5	-0.45	-0.49	0.33	0.22	0.34	1.17	-0.46	-0.49	0.28	0.20	0.28	0.97
1	-0.49	-0.50	0.13	0.10	0.13	1.23	-0.49	-0.50	0.12	0.09	0.12	1.00
$n = 5,000$												
0.1	1.07	-0.30	3.77	1.79	4.08	5.84	-0.50	-0.42	11.70	1.66	11.70	9.52
0.25	-0.37	-0.48	0.73	0.33	0.74	1.49	-0.42	-0.49	1.27	0.28	1.27	1.63
0.5	-0.49	-0.50	0.17	0.13	0.17	1.19	-0.49	-0.50	0.15	0.12	0.15	1.01
1	-0.50	-0.50	0.08	0.06	0.08	1.24	-0.50	-0.50	0.07	0.06	0.07	1.00

Notes: The above table presents simulation statistics for the estimators of ϕ using different values of $\mu_{(0,1)} - \mu_{(1,0)}$ and sample sizes of the design described in Table 3. The summary statistics are computed using 5,000 simulation replications. We set $\beta = 0.25$, $\sigma^2_\mu = 0.25$ and $\sigma^2_u = 1$, with the total share of switchers set as 20% of the sample. Only simulation results where the model converged are displayed. 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.

Table 5: Coverage Probability of 95% CI on ϕ ($\phi = -0.5$)

$\mu_{(0,1)} - \mu_{(1,0)}$	Weak-Id Robust			CRC	Restricted GRC
	All simulations	CRC Conv.	R-GRC conv.	CRC conv.	R-GRC conv.
$n = 1,000$					
0.1	0.950	0.966	0.959	0.939	0.999
0.25	0.940	0.950	0.961	0.945	0.998
0.5	0.945	0.950	0.964	0.969	0.991
1	0.947	0.947	0.948	0.985	0.963
$n = 1,500$					
0.1	0.943	0.957	0.956	0.964	0.999
0.25	0.947	0.957	0.966	0.967	0.998
0.5	0.950	0.956	0.961	0.986	0.985
1	0.945	0.945	0.945	0.983	0.955
$n = 2,000$					
0.1	0.940	0.951	0.956	0.979	0.999
0.25	0.950	0.959	0.969	0.981	0.997
0.5	0.944	0.949	0.953	0.986	0.977
1	0.954	0.954	0.954	0.984	0.963
$n = 5,000$					
0.1	0.944	0.955	0.961	0.998	0.999
0.25	0.955	0.961	0.969	0.995	0.990
0.5	0.950	0.950	0.950	0.986	0.963
1	0.948	0.948	0.948	0.985	0.951

Notes: The above table presents the coverage probability of 95% confidence interval on ϕ for each procedure using different values of $\mu_{(0,1)} - \mu_{(1,0)}$ and sample sizes of the design described in Table 3. The coverage probabilities are computed using 5,000 simulation replications. We set $\beta = 0.25$, $\sigma_\mu^2 = 0.25$ and $\sigma_u^2 = 1$ with the total share of switchers set as 20% of the sample. R-GRC conv. denotes the subset of simulations where the restricted GRC model converged, whereas CRC conv. denotes the subset of simulations where the CRC converged.

Table 6: Coverage Probability of 95% CI on ϕ ($\phi = -0.5$)

$\mu_{01} - \mu_{10}$	$100/\sqrt{n}$			$50/\sqrt{n}$			$10/\sqrt{n}$		
	WIR	CRC	R-GRC	WIR	CRC	R-GRC	WIR	CRC	R-GRC
$n = \quad 500$	0.946	0.985	0.948	0.949	0.983	0.953	0.954	0.949	0.998
$n = 1,000$	0.947	0.984	0.949	0.945	0.986	0.949	0.951	0.968	0.998
$n = 1,500$	0.949	0.983	0.950	0.945	0.984	0.952	0.946	0.980	0.998
$n = 2,000$	0.946	0.985	0.947	0.946	0.984	0.952	0.946	0.911	0.998

Notes: The above table presents the coverage probability of 95% confidence interval for each proceduring using different values for $\mu_{(0,1)} - \mu_{(1,0)}$ and sample sizes. The coverage probabilities are computed based on 5,000 simulation replications of the design described in Table 3. We set $\beta = 0.25$, $\sigma_\mu^2 = 0.25$ and $\sigma_u^2 = 1$ with the total share of switchers set as 20% of the sample. Only simulations results where the model converged are displayed. WIR, CRC, R-GRC abbreviate the weak-identification robust inference procedure we propose, the correlated random coefficient estimator proposed in Suri (2011) and the restricted GRC estimator, respectively.

Table 7: Unrestricted GrRC Estimates and Weak-Identification Robust 95% CI on ϕ

	Full Sample			No HIV districts		
$\mu_{(0,0)}$	5.246	4.457	4.501	5.354	4.531	4.527
	(0.0366)	(0.0715)	(0.110)	(0.0435)	(0.0728)	(0.116)
$\mu_{(0,1)}$	5.942	4.917	4.931	6.068	4.893	4.781
	(0.0949)	(0.114)	(0.150)	(0.100)	(0.118)	(0.159)
$\mu_{(1,0)}$	6.215	5.358	5.289	6.279	5.331	5.182
	(0.0737)	(0.0951)	(0.129)	(0.0753)	(0.0948)	(0.133)
$\Delta_{(0,1)}$	0.508	0.475	0.527	0.500	0.525	0.803
	(0.134)	(0.124)	(0.198)	(0.142)	(0.129)	(0.203)
$\Delta_{(1,0)}$	-0.476	-0.427	-0.312	-0.520	-0.482	-0.213
	(0.104)	(0.0963)	(0.168)	(0.107)	(0.0970)	(0.169)
$\kappa_{(1,1)}$	6.637	5.567	5.649	6.641	5.455	5.638
	(0.0259)	(0.0788)	(0.0994)	(0.0252)	(0.0783)	(0.0945)
WIR 95% CI: ϕ	(-20.1, -2.2)	(-1.04, -.3)	(-2.16, -1.29)	$(-\infty, \infty)$	(-1.25, -.5)	(-2.36, -1.39)
Observations	2394	2394	2394	2114	2114	2114
Controls	No	Yes	Yes	No	Yes	Yes
Interactions	No	No	Yes	No	No	Yes

Notes: Dependent variable is ln yield. Covariates follow Suri (2011). All regressions include acreage, land preparation costs, fertilizer, hired labor, family labor, main season rainfall, household size and age-sex composition of the household (includes indicator variables for the number of boys (aged<16 years), the number of girls, the number of men (aged 17–39), the number of women, and the number of older men (aged>40 years). Weak-Id Robust (WIR) 95% CI is conducted using a grid search over $[-5 * 10^4, 5 * 10^4]$ with 0.01 increments.

Online Appendix

A Proof of Proposition 1

(i) follows by the definition of Δ_i and $\Delta_{\underline{h}}$ as its conditional expectation as follows,

$$\Delta_{\underline{h}} = E[\Delta_i | h_i = \underline{h}] = \beta + \phi E[\theta_i | h_i = \underline{h}] = \beta + \phi \theta_{\underline{h}} \tag{12}$$

(ii)

$$\mu_{\underline{h}} - \mu_{\underline{h}'} = E[\tau_i + \theta_i | h_i = \underline{h}] - E[\tau_i + \theta_i | h_i = \underline{h}'] = \theta_{\underline{h}} - \theta_{\underline{h}'} \tag{13}$$

where the last equality follows from the mean-independence assumption, $E[\tau_i | h_i] = E[\tau_i]$. (iii) is straightforward from (i) and (ii). □

B Restricted GrRC Estimation for Multiple-Period Model

Here we provide the unrestricted GrRC model for any $T \geq 2$

$$y_{it} = \sum_{\underline{h} \in \mathcal{H} : \sum_{t=1}^{T} h_t < T} \mu_{\underline{h}} 1\{h_i = \underline{h}\} + \sum_{\underline{h} \in \mathcal{H}_S} \Delta_{\underline{h}} h_{it} 1\{h_i = \underline{h}\} + \kappa_{\underline{h}_T} h_{it} 1 \left\{ \sum_{t=1}^{T} h_{it} = T \right\} + \varepsilon_{it}. \tag{14}$$

where $\underline{h}_T$ denotes the always-adopter trajectory.

Using Proposition 1, we can obtain a restricted version of the above model,

$$\begin{aligned} y_{it} = & \sum_{\underline{h} \in \mathcal{H} : \sum_{t=1}^{T} h_t < T} \mu_{\underline{h}} + \Delta_{\underline{h}_0} h_{it} + \sum_{\underline{h} \in \mathcal{H}_S \setminus \underline{h}_0} \phi(\mu_{\underline{h}} - \mu_{\underline{h}_0}) h_{it} 1\{h_i = \underline{h}\} \\ & + \left(\mu_{\underline{h}_T} + \phi \left(\mu_{\underline{h}_T} - \mu_{\underline{h}_0}\right)\right) h_{it} 1 \left\{ \sum_{t=1}^{T} h_{it} = T \right\} + \varepsilon_{it}, \end{aligned} \tag{15}$$

for some baseline trajectory $\underline{h}_0 \in \mathcal{H}_S$.

C Data appendix

In this appendix, we provide summary statistics of key variables to enable the reader to get a sense of the differences between our dataset and that used in Suri (2011). We construct our dataset following the step-by-step instructions that the original author provides (Suri, 2018), applying them to the publicly available data from Tegemeo Institute.[11] Despite the careful documentation, some differences remain, likely due to slight modifications in Tegemeo Institute's data processing between the 2006 version of the data used in Suri (2011) and the dataset that is now publicly available.

Table C.1 presents summary statistics of key control variables across years and data sets. Overall, these control variables have similar means and distributions. The biggest differences appear in the 2004 data, for households' fertilizer application rate (`Fertilizer`), hired and family labor (`Hired labor` and `Family labor`, respectively). This is consistent with the notion that Tegemeo Institute applied further data cleaning to the 2004 data set after sharing it with the author, while the 1997 data set was already finalized.

The other difference to note is the number of observations in the two data sets. The final balanced panel in Suri (2011) has 1202 households, while our panel only has 1197 households. This difference arises because some households have missing values for control variables, while they are non-missing in the Suri (2011) data.[12] Overall, the differences between the data sets are relatively minor and should not make a substantial difference if the econometric results are reasonably robust. The fertilizer expenditures variable in 2004 differs more from Suri (2011) than the other variables in the study. We discuss one potential reason for this discrepancy in Section C.1.

C.1 Fertilizer price data

The data have a large number of missing values for district-level fertilizer prices, likely evidence of how thin the ag-input market was in the mid-1990s. Suri (2018) reports addressing the missing values by replacing them with the fertilizer median for that fertilizer type, but some districts have insufficient observations to compute the district-level median. For these districts, we use the same process as suggested for missing price data and assign them the sample median price for that fertilizer type.

[11] Interested readers can request access to the Open Access Data from Tegemeo Institute's website

[12] We lose two households due to missing labor variables (one missing in 1997 and the other in 2004). We further lose three households due to missing data for the household head's education.

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

	1997			2004		
		Mean	Min		Mean	Min
	Obs.	s.d.	Max	Obs.	s.d.	Max
Panel A: Suri (2011)						
Hybrid	1202	0.66	0	1202	0.60	0
		0.47	1		0.49	1
Acres	1202	1.90	0.020	1202	1.96	0.040
		3.22	66.1		2.69	49
Seed rate (kg/acre) x 10	1202	9.58	0.0050	1202	9.07	0.75
		7.80	125		6.86	168.8
Land prep (Ksh/acre) x 1000	1202	960.9	0	1202	541.4	0
		1237.1	8038.6		1022.8	16000
Fertilizer (Ksh/acre) x 1000	1202	1361.7	0	1202	1354.6	0
		2246.3	38585.2		1831.2	20660
Hired labor (Ksh/acre) x 1000	1202	1766.0	0	1202	1427.4	0
		3346.4	71327.2		2130.3	19200
Family labor (Ksh/acre) x 1000	1202	293.2	0	1202	354.3	0
		347.5	5306		352.7	3052.5
Panel B: TGCLMM (2022)						
Hybrid	1197	0.66	0	1197	0.60	0
		0.48	1		0.49	1
Acres	1197	1.91	0.020	1197	1.96	0.040
		3.22	66.1		2.69	49
Seed rate (kg/acre) x 10	1197	10.1	0.0050	1197	10.9	0.25
		8.36	125		8.09	75
Land prep (Ksh/acre) x 1000	1197	2169.3	0	1197	1182.2	0
		4905.1	160771.7		1649.9	16000
Fertilizer (Ksh/acre) x 1000	1197	1426.3	0	1197	3182.3	0
		2394.7	43408.4		6610.2	89600
Hired labor (Ksh/acre) x 1000	1197	1661.3	0	1197	1992.4	0
		3219.9	69742.2		2962.2	26000
Family labor (Ksh/acre) x 1000	1197	292.4	0	1197	560.9	0
		347.7	5306		767.6	17624

In the 2004 data, a few other complications arise. Suri (2018) mentions several data files that have different names in the open access version of the data.[13] Note, however, that fertilizer prices also appear in the household-level dataset `hh04`, where it is elicited in a way similar to the 1997 data: "What is the price of a `[ fertilizer unit ]` bag of `[ fertilizer type ]` in this area?"

> "Merge on fertilizer prices using the fertilizer prices file (key 5) and, as above, merge by district, year, fertilizer type and fertilizer unit. Use the variable `pfert` to value the fertilizer quantities in KShs."

Furthermore, merging on district, fertilizer type, and fertilizer unit is not entirely straightforward. There are a few differences between the coding used for input units and input types between the datasets `tfert04` and `fert04`. Table C.2 shows the mismatches between the two datasets to be merged. The first three columns are field-level reports of the amount of a specific fertilizer type the household used, while the last three were elicited at the household-fertilizer type level. Columns 3 and 6 show the number of fertilizer observations in each category. While some of the discrepancies are very minor (e.g. grams only affects 4 fields), 80 fields have fertilizer use in *gorogoros* (a local unit corresponding to roughly 2 kg) but their fertilizer prices are reported in another unit, and 151 households report fertilizer prices in 5, 10, or 25 kg bags—a unit that does not exist in the field-level data.[14]

We address this by computing median prices at the district-fertilizer type-fertilizer unit level, converting these to a per-kilo price, and merging these prices onto the field-level data by household id (rather than actually merging on district, fertilizer type, and fertilizer unit).[15] We do not know whether Suri (2011) noticed this discrepancy, as it is easy to miss. This is therefore another possible reason why our two datasets are not identical.

[13] The dataset `pricefert` does not exist. The open access data instead contain a dataset called `tfert04`, which contains household- and fertilizer-type-level fertilizer price data. We assume that this is the fertilizer price data Suri (2018) is referring to, and compute district- and fertilizer-type-level prices based on the variable `inputpr`.

[14] There are similar discrepancies in the fertilizer type categories, but they do not end up being relevant in the sample of fertilizer-using maize fields/households.

[15] For the small number of households who report more than one purchase instance of a given input type, we use the mean of the implied per-kilo prices.

Table C.2: Mismatched fertilizer units across two datasets

`fert04`			`tfert04`		
Unit	Label	Frequency	Unit	Label	Frequency
1	90 kg bag	-	-	-	-
2	kg	2,712	2	kg	795
3	litre	100	3	litre	65
4	crate	-	-	-	-
5	numbers	-	17	number	-
6	bunches	-	-	-	-
7	handfuls	-	-	-	-
9	**gorogoro**	**80**	-	-	-
10	tonnes	-	10	tonnes	-
11	50 kg bag	1,487	11	50 kg bag	889
12	debe	-	-	-	-
13	**grams**	**4**	**4**	**gram**	**2**
14	wheelbarrow	-	14	wheelbarrow	-
15	cart	-	15	cart	-
16	canter	-	-	-	-
17	pickup	-	-	-	-
-	-	-	**5**	**5 kg bag**	**1**
-	-	-	**6**	**10 kg bag**	**70**
-	-	-	**7**	**25 kg bag**	**80**
-	-	-	16	days	-

Notes: The frequencies in columns 3 and 6 correspond to observations on fertilizer input use and purchases, as we are not concerned about mismatched labels for other inputs.