Improved Projection GMM-LM Tests for Linear Restrictions*

Saraswata Chaudhuri

This version: October 25, 2017

Abstract

We extend the two-step GMM-LM projection test for subvectors in Chaudhuri and Zivot (2011) to testing null hypotheses on linear restrictions on a parameter vector. This extension retains all the asymptotic properties of the original two-step projection test under more general identification conditions. A key aspect of our paper is that the identification of the linear restriction being tested may be rate-entangled since we allow for multiple rates (strength) of identification for the elements of the parameter vector. This leads to novel issues when studying the two-step projection test’s asymptotic efficiency by establishing its asymptotic equivalence with a locally efficient infeasible plug-in test. This happens because, when the null hypothesis is false, the infeasible plug-in test is not invariant to the characterization of the restrictions not being tested, while the two-step projection test is generally invariant. This asymmetry has manifestly non-ignorable consequences under rate-entangled identification. We address these issues under a unified framework. We present a simple example and a simulation study for illustrating our results.

More broadly, these same issues would also arise under our framework for any invariant test, e.g., the standard plug-in tests or variations of the two-step projection test, seeking asymptotic equivalence with an infeasible but non-invariant test. We work with the two-step projection test since this test has desirable rejection probabilities under the null and, importantly, since this test is based on our earlier work that we wish to extend in this paper.

---

*I thank M. Carrasco, J-M Dufour, A. Mikusheva, E. Renault, I. Andrews, J. Robins, T. Richardson, V. Zinde-Walsh and E. Zivot for helpful discussions. I am also grateful to D. Andrews for helpful discussions about this paper and for pointing out and discussing with me his related but more extensive work Andrews (2017a,b). Informed by this and to keep the paper short, in this updated version, I focus on issues related to local efficiency of a test for parameters with rate-entangled identification that, to my knowledge, have not been in addressed in Andrews (2017a,b) or elsewhere.

†Department of Economics, McGill University; and Cireq, Montreal, Canada. Email: saraswata.chaudhuri@mcgill.ca.
1 Introduction

Consider a parameter vector $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ whose unknown true value $\theta^0$ satisfies the moment restrictions:

$$E[g(Z_t; \theta^0)] = 0$$

where $\{Z_t\}_{t=1}^T$ are $\mathbb{R}^{d_z}$-valued random vectors, $g(\cdot; \theta) : \mathbb{R}^{d_z} \times \Theta \mapsto \mathbb{R}^{d_\theta}$ is a known (up to $\theta$) function, and $d_\theta \geq d_\theta$. Suppose that we are interested in testing:

the null hypothesis $H_0 : R\theta^0 = r_0$ against the alternative hypothesis $H : R\theta^0 \neq r_0$

where $R$ is a fixed, full row-rank, $d_R \times d_\theta$ known matrix, and $r_0$ is a $d_R \times 1$ known vector, and $d_R < d_\theta$.

This paper extends the improved two-step GMM-LM projection subvector (of $\theta$) test presented in Chaudhuri and Zivot (2011) and built on the original work of Robins (2004), to testing $H_0$ in (2). (Also see Chaudhuri (2008), Zivot and Chaudhuri (2009), Chaudhuri et al. (2010), Chaudhuri and Renault (2011).) For brevity, in the sequel we generally refer to this test and its extension as the two-step test.

Chaudhuri and Zivot (2011) established the following results in the context of testing for subvectors. First, allowing for identification failure of $\theta^0$ in (1) as in Stock and Wright (2000), this test’s asymptotic rejection probability of the true null can be non-trivially bounded from above, a property that the subvector-plug-in tests cannot possess in general [see Guggenberger et al. (2012a)]. Second, this test is generally more powerful than the standard projection tests as in Dufour (1997), Dufour and Jasiak (2001), Dufour et al. (2006), Dufour and Taamouti (2005, 2007), etc. Indeed, under the classical conditions as in Newey and McFadden (1994)’s Theorem 9.2 (henceforth NM-9.2) and a global identification condition for the nuisance subvector, this test is locally efficient. This efficiency, which is also the focus of our paper, is what fundamentally distinguishes the two-step test from the standard projection tests noted above, or the Bonferroni tests in Dufour (1990), Berger and Boos (1994), Silvapulle (1996), etc.

In this paper, we establish that all these results of Chaudhuri and Zivot (2011) remain valid for testing the hypothesis on general linear restrictions in (2) under more general characterization of the identification failure of $\theta^0$. Importantly, the efficiency result is also strengthened to cover non-classical scenarios that do not satisfy the NM-9.2 conditions, and this is the main contribution of our paper.

As noted in the abstract and further elaborated in the sequel (Sections 2 (F2) and 4.2), the discussion of efficiency under non-classical scenarios involving rate-entangled identification of $R\theta^0$ leads to novel issues that, to our knowledge, have not been addressed elsewhere in this literature. We point out and address them under a unified framework, and then further illustrate them visually with the help of a simple example and a small-scale simulation study. More broadly, this discussion is related to the
fundamentally important question – What is the benchmark for feasible local efficiency? – the answer

The rest of the paper is organized as follows. Section 2 summarizes at the outset the key features of our paper, their novelty, consequences, and relation with the literature. Section 3 describes the two-step test and heuristically discusses the efficiency result under the classical NM-9.2 setup. The key to efficiency is the test statistic used in the second step, and for this we use the (GMM) LM version of Neyman (1959)’s C-alpha statistic. We show its numerical equivalence with the efficient score statistic used in Chaudhuri and Zivot (2011). While an equivalence result is not surprising thanks to the relation between the efficient score and influence functions, this numerical equivalence is new to our knowledge.

Section 4 allows for identification failure of $\theta^0$ and treats the NM-9.2 setup as a special case. In Section 4.1, the two-step test’s asymptotic rejection probability of the true null is discussed under the setup of Andrews and Guggenberger (2014). In Section 4.2, efficiency for suitable local deviations of the null from the truth is discussed by ruling out weak or worse identification of $\theta^0$ but still allowing for multiple rates (strength) of identification in between weak and strong. For this, we impose more structure that still covers the setups of Stock and Wright (2000), Antoine and Renault (2012), etc. This efficiency consideration in Section 4.2 under rate-entangled identification of $R\theta^0$ is our focus of discussion.

Certain rather long but important definitions for Section 4.2 are collected in Appendix A. Technical proofs and expository materials are collected as Supplementary Materials in Appendices B, C and D.

2 The key features of our paper, and the relation with the literature

To put the contributions of this paper into perspective, before proceeding further, we briefly summarize the key features (F1)-(F4) of our paper, their consequences, and their relation with the literature.

(F1) Re-parameterize (1) to facilitate imposing $H_0$ in the two-step test: Guided by (2), consider a $(d_\theta-d_R)\times d_\theta$ matrix $S$ such that the $d_\theta\times d_\theta$ matrix $A_S = [R', S']'$, indexed by $S$, is nonsingular. $S$ exists since $R$ is full row-rank. Now, for this $S$, consider the invertible linear transformation of $\theta$:

$$(\beta', \gamma_S') := A_S \theta.$$  

(1) and (3) imply that $\beta^0 := R\theta^0$ and $\gamma^0_S := S\theta^0$ are the true values for $\beta$ and $\gamma_S$. The parameter space for $(\beta', \gamma_S')$ is $B \times \Gamma_S$ where $B := \{R\theta : \theta \in \Theta\} \subset \mathbb{R}^{d_R}$ and $\Gamma_S := \{S\theta : \theta \in \Theta\} \subset \mathbb{R}^{d_\theta-d_R}$. The two-step test: (i) constructs a possibly restricted-by-$H_0$ confidence set for $\gamma^0_S$ in the first step, and (ii) rejects $H_0$ if either this confidence set is empty or if the second-step statistic, which is the LM C-alpha statistic
fixed at $\beta = r_0$ and minimized over $\gamma_S$ in the confidence set, exceeds a pre-chosen critical value. Given our numerical equivalence result in Section 3, this is exactly same as the two-step projection test for $H_0 : \beta = r_0$ in Chaudhuri and Zivot (2011) designed specifically under the re-parameterized system (3).

(F2) Describe the framework in terms of the original parameter $\theta$: Note that, in practice, moment restrictions such as (1) are typically conceived in terms of $\theta$, and not $\beta := R\theta$ or an ad-hoc nuisance parameter $\gamma_S := S\theta$. Hence, we deviate from Chaudhuri and Zivot (2011) in that we use (3) only for computational ease of imposing $H_0$, while we describe the framework entirely in terms of $\theta$, more precisely, $\theta^0$, by strictly adhering to (1).\(^1\) Since we allow for multiple rates of identification for the elements of $\theta^0$, the novel issues in the efficiency consideration arise precisely because such allowance can make, concurrently, the identification of $\beta^0$ rate-entangled and that of $\gamma_S^0$ dependent on the choice of $S$.\(^2\)

To elaborate, let $d_\theta = 2$, $\theta = (\theta_1, \theta_2)'$ and $R = [1, 1]$, i.e., $\beta = \theta_1 + \theta_2$. Consider a $\gamma_S$ (e.g., $\gamma_S = \theta_1$ if $S = [1, 0]$, $\gamma_S = \theta_2$ if $S = [0, 1]$) for (3). The local efficiency of the two-step test appeals to its asymptotic equivalence with an infeasible LM test that skips the first step and instead plugs-in the unknown true $\gamma_S^0$ (hence, infeasible), regardless of $H_0$, in the LM C-alpha statistic for the second step. Naturally, the infeasible plug-in test is not invariant to $S$ unless $H_0$ is true. Indeed, it can be very different even locally for the different choices of $S$ if the strengths of identification for the corresponding $\gamma_S^0$’s are not the same.

As a result, the interpretation of the local deviations of $H_0$ from the truth, for which the said asymptotic equivalences hold, needs particular care. For example, let $\theta_1^0$ be nearly strongly and $\theta_2^0$ strongly identified. Then, fixing $\gamma_S = \theta_1^0$ at $\theta_1^0$ leads the local deviation in $\beta = \theta_1 + \theta_2$ to be along the strong direction $\theta_2$, while fixing $\gamma_S = \theta_2$ at $\theta_2^0$ leads the deviation to be along the nearly strong directions $\theta_1$. We consider both (and all others, e.g., $\gamma_S = \theta_1 - \theta_2$) under a unified framework maintained in terms of $\theta$. We present our results such that they automatically reflect that the asymptotic equivalence with the latter, which is the less powerful infeasible test ($\gamma_S = \theta_2$), holds in a larger region, i.e., less locally in terms of the local deviation, than that with the former ($\gamma_S = \theta_1$). It is important to remember that the two-step test itself is generally invariant to the choice of $S$ (and $\gamma_S$); it is the non-invariance of the infeasible plug-in test coupled with the rate-entangled identification of $R\theta^0$ that leads to this non-standard efficiency result.

\(^1\)Given this importance that we assign to $\theta^0$, we let the representation in (1) and (2) suffer from two drawbacks that hinder a satisfactory treatment of issues related to similarity and, hence, size of the two-step test. First, (1) and (2) do not adequately distinguish between the truth for $\theta$ (i.e., $\theta = \theta^0$), and what would be the truth for $\beta$ (i.e., $\beta = \beta^0$, equivalently, $R\theta = R\theta^0$) in a standard subvector test representation. The former is a point in $\Theta$, while the latter is a $(d_\theta - d_R)$ dimensional linear subspace of $\Theta$. Second, we consider the true $\theta$ (i.e., $\theta^0$) and hence, given the first drawback, the true $\beta$ (i.e., $\beta^0$) as fixed but let the hypothesized value $r_0$ vary (possibly with $T$), which is what determines if $H_0$ is true or false. Accordingly, our assumptions focus on the fixed true $\theta^0$. While both drawbacks could be bypassed by maintaining the setup in terms of $\beta$ and $\gamma_S$ (and for only the first one, by maintaining a global version of joint weak convergence assumptions of Kleibergen (2005)), given our focus on highlighting the issues related to local efficiency, we do not do so for brevity and instead refer the reader to Andrews (2017a) for a comprehensive treatment along this route for the asymptotic size of two-step tests.

\(^2\)While related, these issues are not the same as in Antoine and Renault (2009). Also, they do not arise in the study of efficiency with a single rate of identification as in Chaudhuri and Zivot (2011), (I.)Andrews (2016b), Andrews (2017a), etc. They are also standard if assumptions on identification are instead maintained directly on $\beta$ and the specific $\gamma_S$ being used.
(F3) **Impose $H_0$ in the first-step confidence set:** One might infer from (F2) and the non-invariance of the infeasible test that the two-step test could have better power for certain choices of $S$ if $H_0$ is not imposed in step one. However, this has no justification since the two-step test may still be invariant to $S$ (for which, equivariance of the first-step confidence set is sufficient but not necessary). In fact, this leads to very poor power in small samples except in the NM-9.2 setup, in which case it is actually immaterial. We do not recommend this strategy. On the other hand, this appears to better suit the LM plug-in test which then loses invariance to $S$ and resembles the corresponding $S$-dependent infeasible test [see Conniffe (2001) for a somewhat related proposal in a different context]. Nevertheless, as shown in Appendix D, this leads to over-sized LM plug-in tests in cases where a standard, i.e., restricted-by-$H_0$, LM plug-in would have worked (e.g., cases covered by Theorem 6 of Guggenberger and Smith (2005)).

(F4) **Our efficiency consideration is strictly local:** The notion of optimality/efficiency in relation to the infeasible test is less ambitious than that considered in the literature on identification failure; see, e.g., Moreira (2003), Andrews et al. (2006), Moreira and Moreira (2013), (I.)Andrews (2016a), (I.)Andrews and Mikusheva (2016), Montiel-Olea (2016), etc. By contrast, our use of the term is similar to that in Section 9 of Andrews and Guggenberger (2015), or Comment (iii) following Theorem 4.1 of Andrews and Guggenberger (2014), or the oracle equivalence considered in Andrews (2017a). Indeed, the LM-principle generally does not lead to optimality other than in a local sense since it is only based on the slope of the moment vector. Furthermore, as originally noted by Kleibergen (2005), allowing for identification failure necessitates the use of an estimator for the Jacobian matrix that is not simply the sample mean of the derivative of the moment vector, but the sample mean of the residual of the regression of this derivative on the moment vector itself. In certain cases of identification failure, this affects the intended direction along which the LM-principle maximizes local power; see, e.g., Antoine and Renault (2009), (I.)Andrews (2016b), etc. Even otherwise, this may lead to a spurious decline in power away from the truth [see Kleibergen (2005)], which, however, is partially addressed by the two-step test by virtue of a specific choice of the first-step confidence set [see Chaudhuri and Zivot (2011)].

**Related literature:** While we generalize the use of the LM and C-alpha principle in Chaudhuri (2008), Zivot and Chaudhuri (2009), Chaudhuri et al. (2010) and Chaudhuri and Zivot (2011); the LM and/or C-alpha tests were originally used in the context of identification failure by Wang and Zivot (1998), Dufour and Jasiak (2001), Kleibergen (2002), Moreira (2003), Kleibergen (2005), Guggenberger and Smith (2005), Antoine and Renault (2009), etc. It has also been considered recently in Magnusson and Mavroeidis (2010), Guggenberger et al. (2012b), Qu (2014), Dufour et al. (2015, 2016), Andrews and Mikusheva (2015), Andrews and Guggenberger (2014), etc. Even more recently, McCloskey (2015) and Andrews (2016b) propose sophisticated related methods to improve the performance of such tests.
There are three papers that are closest to ours. Of them, we have already noted above the relation and distinction with Chaudhuri and Zivot (2011). The two other papers are Andrews (2017a,b), where this basic two-step test is substantially generalized, extended, refined and demonstrably improved. However, there, the efficiency result (called as oracle equivalence) is not considered under the type of rate-entangled identification that would lead to our key feature (F2). More precisely, although testing for even nonlinear restrictions is considered in Andrews (2017a,b), efficiency is actually discussed for tests of genuine subvectors under some form of strong identification and, more importantly, under variational independence of the parameter spaces for the subvectors being tested and not tested.\footnote{Consideration of nonlinear restrictions is possible in our paper because if the linear restrictions $R\theta^0 = r_0$ are replaced by nonlinear restrictions, say, $R(\theta^0) = 0$ then the corresponding function $S(\theta^0)$ (replacing $S\theta^0$ where $S$ is as in (3)) typically exists in a tubular neighborhood of $\theta^0$. This stronger requirement for existence with a general $\theta$ is necessary for an intermediate step in the proof of our Lemma 2, but not otherwise. Nevertheless, for simplicity of exposition without obfuscating our main message by issues such as a $\theta$-dependent $S$, we choose to work with linear restrictions in this paper.}

In contrast to Andrews (2017a,b) (and all other papers noted above), we stress on the rate-entangled identification since the parameter spaces for the linear combinations being tested and not tested are not necessarily variationally independent. And, accordingly, we adopt a just-sufficiently general framework that enables us to highlight the ensuing local-efficiency-related key feature (F2) that is the novelty of our paper.

### 3 Definition and an overview of the improved two-step projection test

This section maintains the classical NM-9.2 setup as default. Define $\bar{g}_T(\theta) := \frac{1}{T} \sum_{t=1}^{T} g(Z_t; \theta)$, $G(\theta) := \frac{\partial}{\partial \theta} E\left[g(Z_t; \theta)\right]$ and $V(\theta) := Var\left(g(Z_t; \theta)\right)$. Then, the efficient GMM estimator of $R\theta^0$ has the asymptotically linear representation: $\sqrt{T}(R\theta^0 - \theta^0) = -\sqrt{T}l_T(\theta^0) + o_p(1)$ [see Appendix B.1] where

$l_T(\theta) := R\left(G'(\theta)V^{-1}(\theta)G(\theta)\right)^{-1} G'(\theta)V^{-1}(\theta)\bar{g}_T(\theta)$, if it exists. Therefore, for local optimality/efficiency, a test for $H_0$ in (2) can be based on a consistent estimator of $l_T(\theta)$:

$\hat{l}_T(\theta) := R\left(\hat{G}_T'(\theta)\hat{V}_T^{-1}(\theta)\hat{G}_T(\theta)\right)^{-1} \hat{G}_T'(\theta)\hat{V}_T^{-1}(\theta)\hat{g}_T(\theta)$

where $\hat{G}_T(\theta) \overset{p}{\to} G(\theta)$ and $\hat{V}_T(\theta) \overset{p}{\to} V(\theta)$. While the inverse in the above expression may not exist for a given $T$, we will maintain conditions such that its probability limit exists asymptotically either without (in Section 3) or with (in Section 4) an appropriate scaling. This is facilitated by our asymmetric treatment of $G(\theta)$ and $V(\theta)$, in the sense that we will never allow for any rank-failure for $V(\theta)$ [see Andrews and Guggenberger (2015) for such cases] but will allow such in Section 4 for $G(\theta)$.

Guided by the consideration of local efficiency and the above representation of $\hat{l}_T(\theta)$, we use the
following quadratic form of $\tilde{I}_T(\theta)$ for the second step of the two-step test:

$$LM_T(\theta) := T \times \tilde{V}_T(\theta) \left( R \left( \hat{G}_T(\theta) \tilde{V}_T^{-1}(\theta) \hat{G}_T(\theta) \right)^{-1} R' \right)^{-1} \tilde{I}_T(\theta)$$

$$= T \times \left( \tilde{V}_T^{-1/2}(\theta) \tilde{g}_T(\theta) \right)' \left( \hat{G}_T(\theta) \tilde{V}_T^{-1/2}(\theta) \hat{G}_T(\theta) \right)^{-1} \left( \tilde{V}_T^{-1/2}(\theta) \tilde{g}_T(\theta) \right)$$

(4)

where we use the notation $P(D) := D(D'D)^{-1}D'$ to define the projection matrix for any matrix $D$ with the same caveat about the existence of the inverse as noted above (4); and if $D$ is positive semidefinite then we define $D^{1/2}$ to be the upper triangular matrix such that $D = D^{1/2}D^{1/2}$.

$LM_T(\theta)$ is Smith (1987)'s $LLM_T$, Dagenais and Dufour (1991)'s $PC$ or Newey and McFadden (1994)'s $LM_{2n}$ statistic for testing linear restrictions. It falls under the class of Neyman (1959)'s C-alpha statistic.

Note that, (3) and (4) give:

$$LM_T(\theta) = LM_T(A_S^{-1}(\beta', \gamma_S')) .$$

(5)

**Definition:** For $\epsilon, \alpha > 0$ such that $\epsilon + \alpha < 1$, the improved two-step GMM-LM projection test, or simply the two-step test, for $H_0$ in (2) is defined as:

- **Step 1:** obtain a nominal $(1 - \epsilon)$-level confidence set $CI_T(\gamma_S; \epsilon)$ for $\gamma_S^0$;
- **Step 2:** reject $H_0$ if $CI_T(\gamma_S; \epsilon)$ is empty or if $\inf_{\gamma_0 \in CI_T(\gamma_S; \epsilon)} LM_T(A_S^{-1}(r_0', \gamma_0')) > \chi^2_{0.1}(1 - \alpha)$

(6)

where $\chi^2_{0.1}(1 - \alpha)$ is the $(1 - \alpha)$-th quantile of a central $\chi^2$ distribution with $d_R$ degrees of freedom.

$CI_T(\gamma_S; \epsilon)$ is what we refer to as the first-step confidence set, i.e., the preliminary non-point (set) estimator of the nuisance parameter $\gamma_S^0 := S\theta^0$ that is not specified by $H_0$. As noted below, it plays an important role in influencing the asymptotic properties and the computational ease of the two-step test.

**Remark 1:** Invariance of the two-step test to the choice of $S$ in (3) is preserved by the conventional confidence sets $CI_T(\gamma_S; \epsilon)$ regardless of the non-uniqueness of the infimum in the second step. If possible, however, choosing an $S$ with a better identified $\gamma_S^0$ might help with the computation in the second step.

**Remark 2:** To accommodate for identification failures of $\theta^0$, $CI_T(\gamma_S; \epsilon)$ can be obtained by inverting, e.g., the S-test, the K-test, modifications of Moreira (2003)'s CLR test (see Kleibergen (2005), Andrews and Guggenberger (2014, 2015)) for $\gamma_S$, while treating $\beta = r_0$ as known. In practice, the operations required in steps one and two can be simultaneously conducted since, to fail to reject $H_0$, it is sufficient to find a single point $\gamma_0$ that would belong in $CI_T(\gamma_S; \epsilon)$ and also satisfy the condition for step two.

**Remark 3:** It is sufficient to focus on the second-step test statistic to see the connection between (6) and the two-step projection subvector test in Chaudhuri and Zivot (2011), i.e., to see that the test in
(6) is the natural extension of Chaudhuri and Zivot (2011). To this end, for a given $S$, let $R_S^1$ and $S_S^1$ be the $d_θ \times d_R$ and $d_θ \times (d_θ - d_R)$ matrices respectively such that $A_S^{-1} = [R_S^1, S_S^1]$. Then, in this context, the test statistic in Chaudhuri and Zivot (2011)’s second step would be [see Appendix B.2]:

$$LM_{T,S}^{ES}(θ) := T \times \left( \hat{V}_T^{-1/2}(θ)\hat{g}_T(θ) \right)' P \left( \left( I_{d_θ} - P \left( \hat{V}_T^{-1/2}(θ)\hat{G}_T(θ)S_S^1 \right) \right) \hat{V}_T^{-1/2}(θ)\hat{G}_T(θ)R_S^1 \left( \hat{V}_T^{-1/2}(θ)\hat{g}_T(θ) \right) \right),$$

which they refer to as the efficient score (ES) statistic for $β = Rθ$. We show in Appendix B.3 that:

**Lemma 1** Let $\hat{G}_T'(θ)\hat{V}_T^{-1}(θ)\hat{G}_T(θ)$ be positive definite for a given $T$ and $θ ∈ Θ$. Then $LM_{T,S}^{ES}(θ) = LM_{T,S\dagger}(θ)$ for any $S = S^∗, S^1$ for which $A_S := [R', S']'$ is nonsingular.

**Lemma 2** Let $\hat{G}_T'(θ)\hat{V}_T^{-1}(θ)\hat{G}_T(θ)$ be positive definite for a given $T$ and $θ ∈ Θ$. Then $LM_T(θ) = LM_{T,S}^{ES}(θ)$ for any choice of $S$ for which $A_S := [R', S']'$ is nonsingular.

Lemma 1 builds on Dagenais and Dufour (1991) and establishes invariance for $LM_{T,S}^{ES}(θ)$ similar to that for $LM_T(θ)$ in (5). Lemma 2 reconciles between $LM_T(θ)$ and $LM_{T,S}^{ES}(θ)$ and shows that they are the same in the context of testing (2) with a general $R$. While expected from the relationship between the efficient influence function and the efficient score function, on which $LM_T(θ)$ and $LM_{T,S}^{ES}(θ)$ are based respectively, this is, to our knowledge, a new result in the C-alpha literature [see, e.g., Smith (1987), Dagenais and Dufour (1991), Bera and Bilias (2001), Dufour et al. (2015, 2016), Andrews (2017a), and also the references therein for the related papers in the statistics literature].

**Remark 4:** It is useful to note here that the conventional projection test rejects $H_0$ at the level $α$ if:

$$\inf_{θ_0 ∈ Θ: Rθ_0 = r_0} \overline{LM}_T(θ_0) > χ_{d_θ}^2 (1 - α) \quad \text{or equivalently,} \quad \inf_{γ_0 ∈ Γ_s} \overline{LM}_T \left( A_S^{-1}(r_0', γ_0') \right) > χ_{d_θ}^2 (1 - α)$$

(7)

where [see the $LM_{3n}$ statistic in Newey and McFadden (1994) and the K statistic in Kleibergen (2005)]:

$$\overline{LM}_T(θ) := T \times \left( \hat{V}_T^{-1/2}(θ)\hat{g}_T(θ) \right)' P \left( \hat{V}_T^{-1/2}(θ)\hat{G}_T(θ) \right) \left( \hat{V}_T^{-1/2}(θ)\hat{g}_T(θ) \right).$$

The second version in (7) explicitly imposes $H_0$ by using (3) and thus streamlines the computation for the test. Note that, $\overline{LM}_T(\tilde{θ}_T) = LM_T(\tilde{θ}_T)$ where $\tilde{θ}_T$ is the restricted-by-$H_0$ GMM estimator of $θ_T$ [see Appendix B.4]. That is, $\overline{LM}_T(\tilde{θ}_T) = LM_T(\tilde{θ}_T)$ is the standard plug-in LM statistic [see Section 2(F3)] for testing $H_0$ in (2). Then, NM-9.2 gives: $LM_T(\tilde{θ}_T) \overset{d}{\to} χ_{d_R}^2$ distribution, which is central if $H_0$ is true, and non-central under local deviations of $H_0$. Hence, the conservativeness (and thereby, the inefficiency) of the conventional projection test, that the two-step test would address, is due to the use of a $χ_{d_R}^2$ critical value in (7), while the test statistic is actually $\inf_{θ_0 ∈ Θ: Rθ_0 = r_0} \overline{LM}_T(θ_0) \leq \overline{LM}_T(\tilde{θ}_T) = LM_T(\tilde{θ}_T) \overset{d}{\to} χ_{d_R}^2$.
Remark 5: By contrast, the local efficiency of the two-step test, that one would expect from (4), results as follows. For the given S, let \( \theta_0 := A_S^{-1}(r_0', \gamma_0)' \) where \( r_0 := \beta^0 + \mu_\beta / \sqrt{T} \), \( \gamma_0 := \gamma^0_S + \mu_\gamma_S / \sqrt{T} \), \( \mu_\beta \) is a constant and \( \mu_\gamma_S = O_p(1) \). Then, NM-9.2 gives: \( \hat{G}_T(\theta_0) \overset{P}{\rightarrow} G(\theta^0) \), \( \hat{V}_T(\theta_0) \overset{P}{\rightarrow} V(\theta^0) \) and, crucially,
\[
\sqrt{T} \hat{I}_T(\theta_0) \overset{P}{\rightarrow} \sqrt{T} I_T(\theta^0) + \mu_\beta
\]
by using \( RR_S^0 = I_{d_0} \) and \( RS_S^1 = 0 \) that follow from \( A_S A_S^{-1} = I_{d_0} \). Therefore, NM-9.2 gives: \( LM_T(\theta_0) \overset{d}{\rightarrow} \chi^2_{d_{Dr}} \) with non-centrality parameter \( \mu'_\beta \left( R \left( G'(\theta^0) V^{-1}(\theta^0) G(\theta^0) \right)^{-1} R' \right)^{-1} \mu_\beta \), which does not depend on the \( \sqrt{T} \)-deviation of \( \gamma_0 \) from \( \gamma^0_S \). On the other hand, under the same conditions and a global strong identification condition for \( \gamma^0_S \) given \( \beta^0 \), it can be shown that \( \text{[see Lemma 7 in Section 4 for a more general result]} \):
\[
\sup_{\gamma_0 \in CI_T(\gamma^0_S)} \sqrt{T} \| \gamma_0 - \gamma^0_S \| = O_p(1)
\]
for conventional confidence sets, provided that they are non-empty (to fix ideas for now). Hence, by construction, \( \gamma^0_{S,T} := \arg \inf_{\gamma_0 \in CI_T(\gamma^0_S)} LM_T \left( A_S^{-1}(r_0', \gamma_0)' \right) = \gamma^0_S + \mu_{\gamma_S,T} / \sqrt{T} \) for some \( \mu_{\gamma_S,T} = O_p(1) \). Therefore, the two-step test in (6), with a non-empty first-step confidence set, is locally efficient in the sense of (F4) [see Section 2] since it is asymptotically equivalent to the locally optimal/efficient infeasible test — infeasible, since it is based on the unknown true \( \gamma^0_S \) — that rejects \( H_0 \) at the level \( \alpha \) if for \( \theta_0^{inf,S} := A_S^{-1}(r_0', \gamma^0_S)' \):
\[
LM_T(\theta_0^{inf,S}) > \chi^2_{d_{Dr}}(1 - \alpha).
\]
This optimality discussion is only under the classical NM-9.2 setup. Section 4 presents a general treatment allowing for identification failure of \( \theta^0 \) and, hence, considering the NM-9.2 setup as a special case.

4 Asymptotic Properties: When identification failure of \( \theta^0 \) is allowed

Now, we allow for possible identification failure of \( \theta^0 \) due to possible rank-failure of \( G(\theta^0) \). Then, it is imperative that the choice of \( \hat{G}_T(\theta) \) in the definition of \( LM_T(\theta) \) in (4) follows Kleibergen (2005). That is:
\[
\hat{G}_T(\theta) := \left[ \hat{G}_{1,T}(\theta), \ldots, \hat{G}_{d_0,T}(\theta) \right] \quad \text{where} \quad \hat{G}_{T,j}(\theta) := \frac{\partial}{\partial \theta_j} \hat{g}_T(\theta) - \hat{V}_{j,g,T}(\theta) \hat{V}_T^{-1}(\theta) \hat{g}_T(\theta),
\]
\( \hat{V}_{j,g,T}(\theta) \) and \( \hat{V}_T(\theta) \) are respectively \( d_0 \times d_g \) and \( d_g \times d_g \) matrices, and \( \theta_j \) is the \( j \)-th element of \( \theta \) for \( j = 1, \ldots, d_0 \). Unless otherwise noted [see Remark 9], we require \( \hat{V}_{j,g,T}(\theta) \) and \( \hat{V}_T(\theta) \) to be estimators of
\[
V_{j,g}(\theta) := \lim_{T \to \infty} T \cdot E \left[ \left( \frac{\partial}{\partial \theta_j} \hat{g}_T(\theta) - E \left[ \frac{\partial}{\partial \theta_j} \hat{g}_T(\theta) \right] \right) \hat{g}_T(\theta)' \right] \quad \text{for} \ j = 1, \ldots, d_0
\]
and \( V(\theta) := \lim_{T \to \infty} T \cdot E \left[ \left( \hat{g}_T(\theta) - E[\hat{g}_T(\theta)] \right) \hat{g}_T(\theta)' \right]. \]
provided that they exist. Also applicable are the choices of $\hat{G}_T(\theta)$ considered in Guggenberger and Smith (2005, 2008) that only deviate from $\hat{G}_T(\theta)$ defined above by an order of magnitude of $o_p(1/\sqrt{T})$.

We maintain high-level but standard assumptions on the joint distribution $F_T$ of the data $\{Z_t\}_{t=1}^T$. Allowing for a drifting data generating process (DGP) in what follows is important, and to emphasize it we index by $T$ the key parameters defined in terms of $F_T$; see, e.g., Stock and Wright (2000), Andrews and Guggenberger (2014). However, irrespective of the drifting DGP $\{F_T : T \geq 1\}$, we take the truth $\theta^0$ satisfying the moment restrictions in (1) as fixed. $H_0$ in (2) is true if the hypothesized value $r_0$ is equal to $R\theta^0$, it is false otherwise. Apart from characterizing the false $H_0$ by locally deviating (made precise later in (14) and (15)) $r_0$ from $R\theta^0$, no other assumptions involve $r_0$. For convenience, we maintain that:

**Assumption O:**

$$\theta^0 \in \text{interior}(\Theta)$$ where $\Theta$ is compact in $\mathbb{R}^{d_\theta}$.

**Notation:** We suppress the triangular array $\{Z_{t,T} : t = 1, \ldots, T; T \geq 1\}$ notation, and instead denote $Z_{t,T}$ by $Z_t$. Let $c > 0$ and $\bar{c} > 0$ denote generic constants. For any matrix $D$, define $\|D\| := \sqrt{\text{trace}(D'D)}$.

For any $a \times b$ matrix $D = [D_1, \ldots, D_b]$ define $D_{(j,k)} := [D_j, \ldots, D_k]$ as the $a \times (k - j + 1)$ matrix for $1 \leq j \leq k \leq b$. $D_{(k:j)}$ is an empty matrix for $0 \leq j < k \leq b+1$. For an $(ab) \times 1$ vector $D = (d_1, \ldots, d_{ab})'$, define $\text{devec}_b(D) := [(d_1, \ldots, d_b)', (d_{b+1}, \ldots, d_{2b})', \ldots, (d_{(a-1)b+1}, \ldots, d_{ab})']$ as a $b \times a$ matrix.

For reasons stated in footnote 1, the discussion in Section 4.1 is not a proper characterization of the two-step test’s asymptotic size, a term that we generally avoid using in the sequel. While an important topic in its own right that necessitated the use of the two-step test in the first place, we have nothing original to say about it, and Section 4.1 is only presented for completeness. Readers might want to move directly to Section 4.2, the main focus of our paper, and instead consult Andrews (2017a) for a comprehensive treatment of asymptotic size of the various two-step tests proposed by Andrews (2017a).

### 4.1 Rejection of the null hypothesis in (2) when it is true

**Assumption M:**

M1. $\frac{\partial}{\partial \theta}g(z; \theta^0)$ exists for each $z \in \mathbb{R}^{d_\theta}$. Let $\hat{G}_T(\theta) := \frac{1}{T} \sum_{t=1}^T \frac{\partial}{\partial \theta}g(Z_t; \theta)$ and $G_T := E_T[\hat{G}_T(\theta^0)]$. Then,

$$\left[\sqrt{T}\hat{g}_T(\theta^0), \sqrt{T}\text{vec}(\hat{G}_T(\theta^0) - G_T)\right] \xrightarrow{d} [\psi', \psi'_G] \sim N(0, \Sigma)$$

and $\lim_{T \to \infty} \text{Var}_T \left( \begin{bmatrix} \sqrt{T}\hat{g}_T(\theta^0) \\ \sqrt{T}\text{vec}(\hat{G}_T(\theta^0)) \end{bmatrix} \right) \equiv \lim_{T \to \infty} \begin{bmatrix} V_T & V_{G_T} \\ V_{G_T} & V_{GG} \end{bmatrix} = \Sigma := \begin{bmatrix} V & V_G \\ V_G & V_{GG} \end{bmatrix}$.

M2. $\|G_T\|, \|V_T\|, \|V_{G_T}\|, \|V_{G_{G,T}}\| \leq \bar{c}$ for all $T$. $\|\hat{V}_T(\theta^0) - V_T\| = o_p(1)$ and $\|\hat{V}_{G_T}(\theta^0) - V_{G_T}\| = o_p(1)$.
To characterize any identification failure of $\theta^0$, consider the singular value decomposition of $V_T^{-1/2}G_T$:

$$V_T^{-1/2}G_T = C_T\tilde{\Delta}_T B_T'$$

where $C_T$ and $B_T$ are $d_y \times d_y$ and $d_\theta \times d_\theta$ orthogonal matrices whose columns are respectively the eigenvectors of the matrices $V_T^{-1/2}G_TG_T V^{-1/2}$ and $G_T V^{-1}G_T$. $\tilde{\Delta}_T := [\Delta_T, 0]'$ is the $d_y \times d_\theta$ matrix where $\Delta_T := \text{diag}(\delta_{T,1}, \ldots, \delta_{T,d_\theta})$ is the $d_\theta \times d_\theta$ diagonal matrix with its diagonal elements $\delta_{T,1} \geq \delta_{T,2} \geq \ldots \geq \delta_{T,d_\theta}$ ($\geq 0$, without loss of generality) as the singular values of $V_T^{-1/2}G_T$.

**Assumption M (continued):** (identification failure of $\theta^0$ following Andrews and Guggenberger (2014))

**M3.** For the singular value decomposition in (9), there exists a $p \in \{0,1,\ldots,d_\theta\}$ such that:

(a) $\delta_{T,j} \to \delta_j$, a constant, and $\sqrt{T}\delta_{T,j} \to \infty$ for $j = 1,\ldots,p$ as $T \to \infty$ (M3(a) is void if $p = 0$);

(b) $\sqrt{T}\delta_{T,j} \to l_j$, a constant, for $j = p+1,\ldots,d_R$ as $T \to \infty$ (M3(b) is void if $p = d_\theta$);

(c) $C_T \to C$ and $B_T \to B$ as $T \to \infty$ where $B$ is a nonsingular matrix;

(d) If $p < d_\theta$, then $G^* := [C_{(1:p), C_{(p+1:d_\theta)}} L + V^{-1/2}(\theta^0)\text{devec}_{d_\theta}(\psi_G - V_G V^{-1}\psi)B_{(p+1:d_\theta)}]$ is a $d_y \times d_\theta$ matrix with full column-rank $d_\theta$ almost surely, where $L := \text{diag}(l_{p+1},\ldots,l_{d_\theta})$ is a $(d_\theta-p) \times (d_\theta-p)$ diagonal matrix with $l_{p+1},\ldots,l_{d_\theta}$ as its diagonal elements. Else $G^* := C_{(1:p)}$.

**Remark 6:** $p$ is the number of directions in $\theta$ that are better than weakly identified. The remaining $d_\theta - p$ directions in $\theta$ are at best weakly identified and necessitate the particular choice of $\hat{G}_T(\theta)$.

Assumption M3 and the representation involved in it are entirely based on the original work of Andrews and Guggenberger (2014). M3 is actually slightly stronger than what Andrews and Guggenberger (2014) require. This helps to avoid certain peripheral complications arising from the fact that $d_R < d_\theta$.

**Assumptions O, M1 and M2 are standard; see, e.g., Kleibergen (2005), Guggenberger and Smith (2005).**

**Lemma 3** Let assumptions O and M1-M3 hold. Then, for $LM_T(\theta^0)$ defined in (4), $LM_T(\theta^0) \overset{d}{\to} \chi^2_{d_R}$.

**Proposition 4** Let the null hypothesis $H_0$ in (2) be true, i.e., $r_0 = R\theta^0$ for $\theta^0$ defined in (1). Let the joint distribution $\{F_T : T \geq 1\}$ of $\{Z_t\}_{t=1}^T$ be constrained by the assumptions O and M1-M3. Let $\epsilon, \alpha > 0$ and $\epsilon + \alpha < 1$. Let $CI_T(\gamma_S; \epsilon)$ be a confidence set for $\gamma_S$ defined in (3) with asymptotic coverage $(1-\epsilon)$ for $\gamma^0_S := S\theta^0$. Then, the probability with which the improved two-step projection test in (6) rejects $H_0$ cannot exceed $(\epsilon + \alpha)$ asymptotically.

**Remark 7:** The result follows by Bonferroni’s inequality applied to Lemma 3 and the asymptotic coverage of $CI_T(\gamma_S; \epsilon)$. Importantly, the upper bound $(\epsilon + \alpha)$ is entirely under the control of the user.

**Remark 8:** An example of the first-step confidence set $CI_T(\gamma_S; \epsilon)$ that possesses the desired property is
\[
CI_T^{SW}(\gamma_S; r_0, \epsilon) := \left\{ \gamma_0 \in \Gamma_S : T \times Q_T(A_S^{-1}(r_0', \gamma_0')) \leq \chi^2_{d_\theta}(1 - \epsilon) \right\}.
\] (10)

It is obtained by inverting the \(S\)-test of Stock and Wright (2000) (SW). \(r_0\) is the hypothesized value from \(H_0\) in (2). \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) imposes \(H_0\) and utilizes the re-parameterization in (3) for the computation of

\[
Q_T(\theta) := \hat{g}_T(\theta)\hat{V}_T^{-1}(\theta)\hat{g}_T(\theta),
\] (11)

the continuously updated (CU) GMM criterion function. By Theorem 2 of Stock and Wright (2000), the asymptotic coverage of \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) for \(\gamma_0^0 := S\theta^0\) is \((1 - \epsilon)\) when \(H_0\) in (2) is true and when:

(a) \(\sqrt{T}\hat{g}_T(\theta^0) \xrightarrow{d} \psi\) and (b) \(\hat{V}_T(\theta^0) \xrightarrow{P} V\). Since (a) and (b) are included in M1 and M2, the asymptotic coverage for \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) holds under weaker conditions than what we maintain here.

What also works when \(H_0\) is true, but not recommended otherwise since it generally leads to very poor power except under NM-9.2 setup [see (F3) in Section 2, and also Appendix D], is:

an unrestricted-by-\(H_0\) version: \(CI_T^{ ur-SW}(\gamma_S; \epsilon) := \left\{ \gamma = S\theta : \theta \in \Theta, T \times Q_T(\theta) \leq \chi^2_{d_\theta}(1 - \epsilon) \right\}\). (12)

We have always found \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) in (10) to be much more useful in practice for all purposes.\(^4\)\(^5\)

4.2 Rejection of the null hypothesis in (2) when it locally deviates from the truth

We focus on constructing local deviations of the null from the truth that make the null locally false, and generate results resembling that in the classical NM-9.2 setup. While the construction is generic, as noted in Section 2 (F2), its interpretation for specific choices of \(S\) in (3) helps to clarify the local efficiency properties of the two-step test. This is a key aspect of this subsection. To appeal to contiguity arguments, we rule out weak or worse identification of \(\theta^0\). In terms of assumption M3, it means \(p = d_\theta\).

Given this, a key issue is the rate at which \(CI_T(\gamma_S; \epsilon)\) shrinks to \(\gamma_0^0\) [see Remark 5]. This necessitates characterizing \(E_T[\hat{g}_T(\theta)]\) globally for \(\theta \in \Theta\) (more specifically, \(E_T[\hat{g}_T(A_S^{-1}(r_0', \gamma_0'))]\) for \(\gamma_S \in \Gamma_S\)).\(^6\) Two

\(^4\)Amongst the well-known identification-robust confidence sets [see Remark 2], Chaudhuri and Zivot (2011) recommend the use of \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) because of its: (i) validity under weak and general conditions, (ii) computational simplicity, and (iii) effectiveness in eliminating certain spurious declines in power of the GMM-LM test from the second step of the improved projection test. The \(\epsilon\) in the upper bound in Proposition 4 is, in practice, primarily due to the fact that \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) can be empty with nonzero probability (that increases with \(\epsilon\)). While possibly unsatisfying in theory [see Andrews (2017a) for ways to avoid it], this feature is actually useful for (iii) and also for (ii), and hence is accommodated in the definition of the improved two-step projection test in (6). Thus, the recommendation is in spite of the concern raised in Davidson and MacKinnon (2014) and Muller and Norets (2016) (page 2184) that \(CI_T^{SW}(\gamma_S; r_0, \epsilon)\) does not properly reflect the parameter uncertainty. (This concern is at least partly addressed by the second step of the improved two-step projection test.)

\(^5\)This is a continuation of footnote 4. While it is clear that \(CI_T(\gamma_S; \epsilon)\) based on Kleibergen (2005)’s GMM-LM principle cannot be helpful for (iii) in general, it should be noted that \(CI_T(\gamma_S; \epsilon)\) based on Moreira (2003)’s conditional likelihood ratio principle may not also be helpful for (iii). Simulation evidence and discussion on this can be found in Section 7.2.1 of Andrews (2016b). Both such \(CI_T(\gamma_S; \epsilon)\)’s can also be less appealing in terms of (i) and (ii) [see Mikusheva (2010) for (ii)].

\(^6\)See Remark 5. Andrews (2017a) ensures this in a similar setup by maintaining a global strong-identification condition SI2(i)/(13.2) for \(\gamma_0^S\) given \(\beta = r_0\) and a local strong-identification condition SI2(ii)/SS2LM(i) for \(\theta^0\). By contrast, our local and global conditions are going to be inter-related, and they allow for worse than strong but better than weak identification.
well-known setups for this are Stock and Wright (2000) and Antoine and Renault (2012). They differ in terms of their consequences in that the former results in rate-disentangled local identification of $\theta^0$, while the latter generally leads to rate-entanglement. Both will lead to rate-entangled $R\theta^0$ in general.

Both setups (and also that in Section 4.1) share a common trait that the local behavior of $E_T[\hat{g}_T(\theta)]$ under them can be characterized by the existence of a nonsingular matrix (depends on $T$ and the setup) which, when pre-multiplied by the (scaled) expected Jacobian, converges to a finite $d_g \times d_\theta$ matrix of full column-rank. Hence, the local analysis under both setups are similar except that Antoine and Renault (2012) demand a little extra work since it makes the identification of $\theta^0$ itself rate-entangled.

Therefore, we take the following strategy for the sake of brevity: We explicitly follow Antoine and Renault (2012) for the discussion in this subsection, and remark on the simplifications that are possible under Stock and Wright (2000). Since modeling $E_T[\hat{g}_T(\theta)]$ explicitly has implications on its derivative and since local efficiency considerations rule out weak/joint-weak identification anyway, there is no reason to continue with the general setup from Section 4.1; we abandon it. (See Andrews (2017a) for a rigorous treatment of efficiency with that setup under some form of strong identification [see footnote 6].)

Our strategy leads to a framework that is (a little more than) just-sufficiently general for us to highlight the key feature (F2) related to local efficiency under rate-entangled identification and the variational dependence of the parameter spaces for the linear combinations being tested and not tested.

Accordingly, following Antoine and Renault (2012), for some $\rho : \Theta \mapsto \mathbb{R}^{d_\theta}$ and a sequence of diagonal matrices $\{\Lambda_T : T \geq 1\}$, let

$$E_T[\hat{g}_T(\theta)] = \frac{\Lambda_T}{\sqrt{T}} \rho(\theta).$$

(13)

**Notation:** Let $1_c$ denote the $1 \times c$ vector with all elements equal to 1. For a set of $d_g$-dimensional vectors $\{a_j\}_{j=1}^q$, let $\text{diag}(a_1, \ldots, a_q)$ denote the $\sum_{j=1}^q d_g$-dimensional diagonal matrix with diagonal elements as the elements of $a_1, \ldots, a_q$ respectively. Let $\mathcal{N}(\theta^0) \subset \Theta$ denote a generic open neighborhood of $\theta^0$.

**Assumption N:** (following Antoine and Renault (2012))

N1. $\rho(\theta^0) = 0$ and $\inf_{\|\theta - \theta^0\| \geq c} \|\rho(\theta)\| > 0$ for any $c > 0$.

N2. $\psi_T(\theta) := \sqrt{T} (\hat{g}_T(\theta) - E_T[\hat{g}_T(\theta)]) \Rightarrow \psi(\theta)$ where $\psi(\theta)$ is a Gaussian process on $\Theta$ with mean zero and covariance function $V(\theta_1, \theta_2)$. $V(\theta^0) = V$ (as in M1) where $V(\theta) := V(\theta, \theta)$.

N3. $\{\Lambda_T : T \geq 1\}$ is a deterministic sequence of $d_g \times d_g$ diagonal matrices with positive diagonal elements.

$I^*$ is a $d_g \times d_g$ matrix whose rows are a suitable permutation of the rows of $I_{d_g}$ such that $I^* \Lambda_T I^* = \text{diag}(\lambda_{T,1} k_1, \ldots, \lambda_{T,1} k_l)$ where $k_j > 0$ for $j = 1, \ldots, l$ and $\sum_{j=1}^l k_j = d_g$. $\lambda_{T,j} = o(\lambda_{T,j+1})$ for $j = 1, \ldots, l-1$; $\lim_T \lambda_{T,1} = \infty$ but $\lim_T \lambda_{T,l}/\sqrt{T} < \infty$.\footnote{$I^{-1} = I^*$. $I^*$ is not unique unless $k_1 = \ldots = k_l = 1$ and thus $l = d_g$. The multiplicity of the elements can be made}
N4. The $d_y \times d_y$ matrix $\rho_0(\theta) := \frac{\partial}{\partial \theta} \rho(\theta)$ exists, has full column-rank $d_y$, and is continuous in $\theta \in \mathcal{N}(\theta^0)$.  

N5. $g(z; \theta)$ is differentiable in $\theta \in \mathcal{N}(\theta^0)$ for each $z \in \mathbb{R}^{d_z}$.

N6. \[ \frac{\partial}{\partial \theta} \psi_T(\theta^0) = \sqrt{T} \left[ \tilde{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}} \rho_0(\theta^0) \right] = O_p(1). \]

N7. (N7 (a) and (b) are grouped together to allow us to briefly discuss a tradeoff in Remark 9 below.)

(a) $\rho(\theta)$ is twice continuously differentiable in $\theta \in \mathcal{N}(\theta^0)$. $g(z; \theta)$ is twice differentiable in $\theta \in \mathcal{N}(\theta^0)$ for each $z \in \mathbb{R}^{d_z}$.\[ \sup_{\theta \in \mathcal{N}(\theta^0)} \left\| \frac{\partial}{\partial \theta} \left[ \tilde{G}_T(\theta) - \frac{\Lambda_T}{\sqrt{T}} \rho_0(\theta) \right] \right\| = O_p(\lambda_{T,1}/\sqrt{T}) \] for $i = 1, \ldots, d_y$. 

(b) $\lambda_{T,1}^2/\lambda_{T,1} \to \infty$ as $T \to \infty$.

N8. \[ \sup_{\theta \in \Theta} \left\| \tilde{V}_T(\theta) - V(\theta) \right\| = O_p(1). \sup_{\theta \in \mathcal{N}(\theta^0)} \left\| \tilde{V}_{Gg,T}(\theta) - V_{Gg}(\theta) \right\| = o_p(1). \]

Remark 9: The assumptions in N describes the setup following Antoine and Renault (2012), who also provide discussion on each of them. Some assumptions are maintained locally since that suffices for us. Given the rest, N7(b) can be slightly weakened at the cost of messy notation; in particular, it suffices if $\lambda_{T,1}^2/\lambda_{T,1} \to \infty$ as $T \to \infty$ where $\lambda_{T,1}$ has a rather long definition stated in Appendix A.2.2. It should, however, be noted that there is a tradeoff between the smoothness assumption N7(a) and the rate assumption in N7(b). This explains why, e.g., in a linear instrumental variables regression where the higher order derivatives are necessarily zero, one does not, unlike N7(b), need to impose a lower bound to the rate at which $\lambda_{T,1} \to \infty$.\[ \sup_{\theta \in \Theta} \max[\text{eigen values}(V(\theta))] \leq \bar{c} < \infty \text{ and } \inf_{\theta \in \Theta} \min[\text{eigen values}(V(\theta))] \geq \underline{c} > 0. \]

Remark 10: In this setup, Antoine and Renault (2012) characterize an orthogonal rotation of $\theta^0$ with rate-disentangled identification as $\Pi'_{\rho_0} \theta^0$ and with the rates given by $\sqrt{T}D_{T,\rho_0}^{-1}$. $\Pi_{\rho_0}$ and $D_{T,\rho_0}$ are, respectively, $d_y \times d_y$ orthogonal and diagonal matrices dependent on $\frac{\Lambda_T}{\sqrt{T}} \rho_0(\theta^0)$ and with rather involved definitions that, for the sake of readability, are presented in detail in Appendices A.2.1 and A.2.2.

Definition: For a fixed $d_R \times 1$ vector $\mu_\beta$, define the local deviation of the null $H_0$ from the truth as:

\[ \sqrt{T}D_{T,R}\Pi'_{R}(r_0 - \beta^0) = \mu_\beta, \] (14)

where $D_{T,R}$ and $\Pi_{R}$ are, respectively, $d_R \times d_R$ diagonal and orthogonal matrices dependent on $R$, $\Pi_{\rho_0}$ and $D_{T,\rho_0}$. The rather involved definitions of $\Pi_{R}$ and $D_{T,R}$ are presented in Appendices A.3.1 and A.3.2.\[ ^{10} \]
However, as noted in Section 2 (F2), (14) is not sufficient for a study of local efficiency in a rate-entangled setup like ours. More is needed to complete the description of “local”. Accordingly, consider an arbitrary and possibly random sequence \( \{\gamma_{S,T} : T \geq 1\} \in \Gamma_S \) such that \( \theta_T := A_S^{-1}(\gamma_0, \gamma_{S,T})' \) satisfies:

\[
\sqrt{T} D_{T,\rho_0}^{-1} \Pi_{\rho_0} (\theta_T - \theta^0) \equiv \sqrt{T} D_{T,\rho_0}^{-1} \Pi_{\rho_0} \left( R_S^1(\gamma_0 - \theta^0) + S_S^1(\gamma_{S,T} - \gamma_S^1) \right) = \mu_T,\theta \quad \text{for some } \mu_T,\theta = O_p(1).
\]  

(15)

**Definition:** Define the full row-rank matrix \( R^* \) [Appendix A.3.3 contains the details of its construction in (24)] as:

\[
R^* := \lim_{T \to \infty} D_{T,R} \Pi_R R \Pi_{\rho_0} D_{T,\rho_0} \quad \text{and note that, } \quad R^* \mu_T,\theta \xrightarrow{P} \mu_\beta.
\]

**Definition:** Define the full column-rank matrix \( G^* \) [Appendix A.2.3 contains the details of its construction in (21)] as:

\[
G^* := \lim_{T \to \infty} \frac{\Lambda_T}{\sqrt{T}} \rho_\theta(\theta^0) \Pi_{\rho_0} D_{T,\rho_0}.
\]

**Remark 11:** \( R^* \) and \( G^* \) correspond to \( R \) and \( G(\theta^0) \) respectively in the classical NM-9.2 setup [see Remark 5]. However, \( R \) and \( R^* \) can be very different otherwise [see Remark 16 for an example].

**Remark 12:** Finally, note the modifications to these definitions that would also make the results in the sequel applicable to the setup of Stock and Wright (2000). To distinguish between the setups, use an “upper bar” for the corresponding quantities now. Resembling N3, define \( \Lambda_T := \text{diag}(\lambda_T,1,k_1,\ldots,\lambda_T,1_{k_1}) \) where \( \tilde{k}_j > 0 \) for \( j = 1,\ldots,l \), \( \sum_{j=1}^l \tilde{k}_j = d_\theta \), \( \lambda_T,j/\lambda_T,j+1 = o(1) \) for \( j = 1,\ldots,l-1 \); \( \lim_T \lambda_T,j = \infty \) and \( \lim_T \lambda_T,j/\sqrt{T} < \infty \) (\( k_0 = 0, \tilde{k}_j > 0 \) for \( j = 1,\ldots,l \), \( \sum_{j=1}^l \tilde{k}_j = d_\theta \), and, for \( j = 1,\ldots,l \), the functions \( \rho(\cdot) \) are \( d_\theta \times 1 \) deterministic functions satisfying the restrictions in Section 3.3 of Stock and Wright (2000), and other conditions as needed.

since we want the results to resemble NM-9.2. For example, suppose instead that one wishes to test \( \theta^0 = \theta_0 \). So, \( d_\theta = d_\theta \). Then, under the setup of Section 4.1, the proper local deviation should be \( \lim_T \text{diag}(1/\delta_{1,1},\ldots,1/\delta_{T,\rho}) E_T^2(\theta_0 - \theta^0) = \mu_\theta \) for a fixed \( \mu_\theta \); while under the setup here in Section 4.2, this should resemble (15). Both are still nonstandard.

A direct generalization of Chaudhuri and Zivot (2011) adhering to Stock and Wright (2000)’s setup would lead to the above structure giving rate-disentangled \( \theta^0 \). As opposed to (13), here one could, for example, model \( E_T[\tilde{g}_T(\theta)] \) as [c.f. (13)]:

\[
E_T[\tilde{g}_T(\theta)] = \sum_{j=1}^l \frac{\lambda_T,j}{\sqrt{T}} \left( \tilde{\rho}(\theta_{j-1+1},\ldots,\theta_{j}) - \tilde{\rho}(\theta_{j-1+1},\ldots,\theta_{d_\theta}) \right)
\]

where \( \theta = (\theta_1,\ldots,\theta_{d_\theta})' \), \( \tilde{k}_0 = 0, \tilde{k}_j > 0 \) for \( j = 1,\ldots,l \), and, for \( j = 1,\ldots,l \), the functions \( \tilde{\rho}(\cdot) \) are \( d_\theta \times 1 \) deterministic functions satisfying the restrictions in Section 3.3 of Stock and Wright (2000), and other conditions as needed.
4.2.1 Asymptotic results:

Lemma 5 Let assumptions O and N hold. For the given S in (3), let \( r_0 \) in (14) be such that for the true value \( \gamma^0_S \) of \( \gamma_S \), the sequence \( \theta_{0,S}^{inf} := A_S^{-1}(r_0', \gamma^0_S) \equiv R^1_S r_0 + S^1_S \gamma^0_S \) satisfies (15). Consider any sequence \( \theta_T = A_S^{-1}(r_0', \gamma_{S,T}') : T \geq 1 \) where \( \{ \gamma_{S,T} : T \geq 1 \} \) is such that (15) holds. Then, the following results hold for \( LM_T(\theta_T) \) defined in (4), as \( T \to \infty \):

(a) \( LM_T(\theta_T) = LM_T(\theta_{0,S}^{inf}) + o_p(1). \)

(b) \( LM_T(\theta_T) \xrightarrow{d} \chi^2_d \) with non-centrality parameter \( \mu_\beta \left( R^*(G^*V^{-1}G^*)^{-1} R^* \right)^{-1} \mu_\beta. \)

Lemma 6 Let assumptions O and N hold. For the given S in (3), let \( r_0 \) in (14) be such that for the true value \( \gamma^0_S \) of \( \gamma_S \), the sequence \( \theta_{0,S}^{inf} := A_S^{-1}(r_0', \gamma^0_S') \equiv R^1_S r_0 + S^1_S \gamma^0_S \) satisfies (15). For \( \epsilon, \alpha > 0 \) such that \( \epsilon + \alpha < 1 \), let \( CI_T(\gamma_S; \epsilon) \) be a confidence set for \( \gamma^0_S \) such that:

\[
\sup_{\gamma_0 \in CI_T(\gamma_S; \epsilon)} \sqrt{T} \left\| D^{-1}_{T, \rho_0} \Pi_{\rho_0}' \left( (R^1_S(r_0 - \beta^0) + S^1_S(\gamma_0 - \gamma^0_S)) \right) \right\| = O_p(1) \tag{16}
\]

where \( \Pi_{\rho_0} \) and \( D_{T, \rho_0} \) are defined in (18) and (19) in Appendices A.2.1 and A.2.2 respectively. Then,

\[
\inf_{\gamma_0 \in CI_T(\gamma_S; \epsilon)} LM_T \left( A_S^{-1}(r_0', \gamma_0') \right) = LM_T(\theta_{0,S}^{inf}) + o_p(1) \text{ for } LM_T(\theta) \text{ defined in (4)}.
\]

Remark 13: Lemmas 5 and 6 formalize the discussion of asymptotic equivalence from Remark 5. As we will see in Lemma 7 below, the asymptotic behavior of the two-step test itself does not require imposing the condition (15) since the first step based on \( CI_T^W(\gamma_S; r_0, \epsilon) \) in (10) automatically incorporates (15) by virtue of (16). Condition (15) is, however, needed to completely characterize the local deviation of \( H_0 \) for which the asymptotic equivalence with the corresponding \( (S\text{-dependent}) \) infeasible test in (8) holds. This is an important observation that fundamentally distinguishes our paper from the related literature on efficiency. Sections 4.2.2 and 4.2.3, respectively, analytically and visually illustrate this observation.

Remark 14: Lemma 5(a) establishes the key property of \( LM_T(\theta) \) that its asymptotic behavior is invariant to certain local deviations of the nuisance parameter \( \gamma_S \) from \( \gamma^0_S \). This motivated Chaudhuri and Zivot (2011)’s use of LM C-alpha in a projection test. Lemma 5(b) describes the asymptotic distribution of \( LM_T(\theta) \), which closely resembles that from the NM-9.2 classical setup; the latter is a special case. Lemma 6 utilizes Lemma 5(a) to establish that the second-step test statistic is asymptotically equivalent to the infeasible statistic under the condition (16) imposed on the first-step confidence set. Then, the asymptotic rejection rate of the improved two-step projection test in (6) follows from Lemma 5(b).

Remark 15: Condition (16) is important for benefitting from the use of (LM) C-alpha, and this is what we ideally expect the first-step confidence set to satisfy. Given the local deviation in (14),
it characterizes the worst possible convergence of the first-step confidence set to $\gamma_S^0$ that still allows appealing to (15), which is necessary for discussing the local asymptotic properties of $LM_T(\theta)$. In the case where $\Lambda_T = \lambda_T I_d$, for some $\lambda_T \to \infty$ (but $\lim_T \lambda_T/\sqrt{T} < \infty$), i.e., all the rates are equal, and if our interest lies in testing subvectors of $\theta$ (e.g., $R = |I_d|, [0]$), then, by virtue of (14), the condition in (16) boils down to $\sup_{\gamma_0 \in CI_T(\gamma_S; \varepsilon)} \lambda_T \left\| \gamma_0 - \gamma_S^0 \right\| = O_p(1)$, which becomes $\sup_{\gamma_0 \in CI_T(\gamma_S; \varepsilon)} \sqrt{T} \left\| \gamma_0 - \gamma_S^0 \right\| = O_p(1)$ if, additionally, we focus only on the strong identification classical setup as done in the related literature.

It is, however, clear that our recommended confidence set $CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ from Section 4.1 cannot satisfy (16) since $CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ can be empty with positive probability. Nevertheless, as noted earlier in footnote 4, we are of the opinion that as long as it satisfies the requirement of Proposition 4, the practical benefit of an empty $CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ (with a small $\varepsilon$) in: (i) eliminating spurious declines in power and (ii) easing computation, may be worth the cost associated with it. With the caveat of emptiness, which we sidestep by redefining the supremum in (16), we now show that $CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ would satisfy (16).

**Lemma 7** Let assumptions O and N hold. Let $r_0$ satisfy (14). Define the supremum in (16) as zero if $CI_T(\gamma_S, \varepsilon)$ is empty for a given $\varepsilon > 0$. Then, $CI_T(\gamma_S, \varepsilon) = CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ satisfies (16) for $\varepsilon > 0$, i.e.,

$$\sup_{\gamma_0 \in CI_T^{SW}(\gamma_S; r_0, \varepsilon)} \sqrt{T} \left\| D^{-1}_{T, \rho_0} \Pi' \left( (R^1(r_0 - \beta^0) + S_1^1(\gamma_0 - \gamma_S^0)) \right) \right\| = O_p(1).$$

For completeness, we summarize Lemmas 6 and 7 to present our final result in Proposition 8.

**Proposition 8** Let assumptions O and N hold. Consider the $S$ given in (3). Let the sequence of hypothesized value $r_0$ in (2) locally deviate from the truth $\beta^0 = R \theta^0$ following (14), and such that the infeasible sequence $\theta^{int}_{0, S} := A^{-1}_S (r_0', \gamma_S')' \equiv R^1_S r_0 + S^1_S \gamma_S^0$ satisfies (15). Then, for $\varepsilon, \alpha > 0$ such that $\varepsilon + \alpha < 1$, the asymptotic probability of rejection of $r_0$ by the improved two-step projection test in (6), based on the choice $CI_T^{SW}(\gamma_S; r_0, \varepsilon)$ in (10), cannot be smaller than that by the infeasible test in (8).

### 4.2.2 Discussion of the asymptotic results with an example (reference: Remark 13):

Consider the example from Section 2 (F2) where $d_{\theta} = 2$, $\theta = (\theta_1, \theta_2)'$ and $R = [1, 1]$. Consider the two choices of $S$: $S^* = [1, 0]$ and $S^\dagger = [0, 1]$, giving $\gamma_{S^*} = \theta_1$ and $\gamma_{S^\dagger} = \theta_2$ respectively. We focus on what our local results say about the asymptotic equivalence with the infeasible statistic, in this case for the two choices $S^*$ and $S^\dagger$ respectively, and how, through (15), that restricts the local deviation in (14).

Complying with both Antoine and Renault (2012) and Stock and Wright (2000), let, for example,

$$E_T[\tilde{G}_T(\theta^0)] = \begin{bmatrix} \frac{\lambda_{1,1}}{\sqrt{T}} & 0 \\ 0 & \frac{\lambda_{2,2}}{\sqrt{T}} \end{bmatrix},$$

and define $\rho_1 := (\rho_1, 0)' \neq 0$ and $\rho_2 := (0', \rho_2^2)' \neq 0$. (17)
Therefore, $\lambda_{T,1} = \tilde{\lambda}_{T,1}$, $\lambda_{T,2} = \tilde{\lambda}_{T,2}$, $\Pi_{\rho g} = \tilde{\Pi}_{\rho g} = I_2$, $D_{T,\rho g} = \tilde{D}_{T,\rho g} = \sqrt{T}\text{diag}(\lambda_{T,1}^{-1}, \lambda_{T,2}^{-1})$, $\Pi_R = \tilde{\Pi}_R = 1$ and $D_{T,R} = \tilde{D}_{T,R} = \lambda_{T,1}/\sqrt{T}$. This is a setup where identification of $\theta^0$, but not $R\theta^0$, is rate-disentangled.

**Remark 16:** This setup gives $R^* = \tilde{R}^* = \lim_T[1, \lambda_{T,1}/\lambda_{T,2}]$. Unless $\lambda_{T,1} = \lambda_{T,2}$, we have $R^* = \tilde{R}^* = [1, 0] \neq R = [1, 1]$ [see Remark 11]. On the other hand, here $G^* = [\rho_1, \rho_2]$ (block-diagonal). Therefore, defining $m_{ij} := \rho'_j V^{-1}\rho_i$ for $j = 1, 2$, this implies that the non-centrality parameter in Lemma 5(b) is $\mu_3^2/(m_{11} - m_{12}^2/m_{22})$, where the denominator is what one should obtain in a subvector test for $\theta^0$, the element with less strong identification [c.f. Proposition 2.2, Antoine and Renault (2009)].

**Remark 17:** In this example, $\mu_\beta = \lambda_{T,1}(r_0 - \beta^0)$ by (14). However, (15) gives $\lambda_{T,1}(\theta_1 - \theta_1^0) + \lambda_{T,2}(\theta_2 - \theta_2^0) = \mu_{T,\theta}$, which implies that $\lambda_{T,1}(r_0 - \beta^0) + (\lambda_{T,2} - \lambda_{T,1})(\theta_2 - \theta_2^0) = \mu_{T,\theta}$. Hence, while (14) captures deviations of order up to $\lambda_{T,1}^{-1}$ for $\beta := \theta_1 + \theta_2$, any deviation of order bigger than $\lambda_{T,2}^{-1}$ along the $\theta_2$-axis, that causes this, is beyond the scope of our results. (On the other hand, everything is standard along the $\theta_1$-axis.) This is a consequence of allowing for multiple rates that make $R\theta^0 := \theta_1^0 + \theta_2^0$ rate-entangled.

**Remark 18:** Remark 17 is not relevant for the asymptotic behavior of the two-step test itself, but is important for its asymptotic equivalence with the infeasible tests. Note that, for the choices $S = S^*, S^T$, we have $\theta_{0,S}^{inf,s}$ as $\theta_{0,S}^{inf,s} = (\gamma_{S^*}^0 := \theta_1^0, r_0 - \gamma_{S}^0, \gamma_{S}^1) = (\mu_\beta, \gamma_{S}^1)$ respectively. Consider $r_0$ satisfying (14), i.e., $(r_0 - \beta^0) = \mu_\beta/\lambda_{T,1}$ for some constant $\mu_\beta \neq 0$, which, for $\theta_{0,S}^{inf,s}$ and $\theta_{0,S}^{inf,s}$ respectively, means that $(\theta_{T,2} - \theta_2^0) = \mu_\beta/\lambda_{T,1}$ and $(\theta_{T,1} - \theta_1^0) = \mu_\beta/\lambda_{T,1}$. Hence from Remark 17 we know that for a constant $\mu_\beta \neq 0$ in (14), $\theta_{0,S}^{inf,s}$ falls directly under the scope of our results, while $\theta_{0,S}^{inf,s}$ does not. For the latter, we need $\mu_\beta = O(\lambda_{T,1}/\lambda_{T,2})$, a deviation more local than what is implied by (14) alone.

**Remark 19:** This is the role of (15). It refines the deviation in (14) to precisely tell where the asymptotic equivalence holds. Interestingly, since the two-step test (and the plug-in tests) are generally invariant to $S$, this means that irrespective of what it uses for $S$, the asymptotic equivalence holds in a larger region with the infeasible test with $S = S^T$, i.e., the one using the better identified nuisance parameter $\gamma_S = \theta_2$. This nicely aligns with the practical implementation of the two-step projection test since it is generally computationally easier to use the same, i.e., the most strongly identified, $\gamma_S$, if possible [see Remark 1].

### 4.2.3 Simulation study with a linear instrumental variables model:

Let us now visually demonstrate the points made in Section 4.2.2 with the help of a linear instrumental variables model, the leading example of GMM. A linear (in $\theta$) $g(Z_i; \theta)$ gives the following simplifications.

**Remark 20:** The rate-restrictions $\lambda_{T,j}(\theta_{T,j} - \theta_j^0) = o_p(1)$ for $j = 1, 2$ in (15) can be weakened for the properly scaled terms inside the projection matrix $P(.)$ in $LM_T(\theta)$ in (4) to converge to the concerned limits. All we need are $(\theta_{T,1} - \theta_1^0) = o_p(1)$ and $\lambda_{T,2}(\theta_{T,2} - \theta_2^0) = o_p(\lambda_{T,1})$.\(^{12}\) Hence, now, for deviations

\(^{12}\)Let $\theta_T \overset{p}{\to} \theta^0$. Then, Lemma 11 (a), (b) and (c) hold [see Appendix C.1]. On the other hand, for $j = 1, 2$, use N2 to obtain
satisfying this condition instead of (15), our results give that only the infeasible test with $S = [1, 0]$, i.e., with $\gamma_\theta = \theta_1$, rejects $H_0$ with probability approaching one, except spuriously [c.f. Remarks 17 and 19].

**Simulation design:** Consider an i.i.d. sample $\{Z_{i,T}\}_{T=1}^T$ written as $\{Z_t := (y_t, X_{1t}, X_{2t}, W_t^j)\}_{T=1}^T$ where dependent variable: $y_t = X_{1t}\theta_1^0 + X_{2t}\theta_2^0 + u_t$,

endogenous regressors: $X_{jt} = W_t^j\pi_{jT} + v_{jt}$ for $j = 1, 2$,

while the instruments $W_t \sim N(0, I_3)$ are independent of the model errors $(u_t, v_{1t}, v_{2t})$: $u_t \sim N(0, 1)$, $v_{jt} \sim N(0, 1)$ with Cov$(u_t, v_{jt}) = .8$ and Cov$(v_{1t}, v_{2t}) = .3$ for $j = 1, 2$. We take $\theta_1^0 = \theta_2^0 = .5$.

The moment vector corresponding to (1) is $g(Z_t; \theta) = W_t(y_t - X_{1t}\theta_1 - X_{2t}\theta_2)$, which is $4 \times 1$ dimensional. Note that, this design closely follows Chaudhuri and Zivot (2011) except that we will now simplify the structure of $[\pi_{1T}, \pi_{2T}]$ now such that it conforms not only to the setup of Stock and Wright (2000) but also to that of Antoine and Renault (2012). In particular, relating to the setup and discussion in Section 4.2.2, we specify $\pi_{jT}$ for $j = 1, 2$ such that (17) holds with:

- $\rho_1 = \sqrt{4/3}(1, 1, 0, 0)'$ if $\lambda_{T,1} \neq \sqrt{T}$, and $\rho_1 = 20(1, 1, 0, 0)'$ if $\lambda_{T,1} = \sqrt{T}$,
- $\rho_2 = \sqrt{4/3}(0, 0, 1, 1)'$ if $\lambda_{T,2} \neq \sqrt{T}$, and $\rho_2 = 20(0, 0, 1, 1)'$ if $\lambda_{T,2} = \sqrt{T}$.

We take $T = 100$ and consider six specifications for the identification strength of the elements of $\theta^0$:

(i) $\lambda_{T,1} = \lambda_{T,2} = 1$, (ii) $\lambda_{T,1} = 1, \lambda_{T,2} = T^{1/6}$, (iii) $\lambda_{T,1} = 1, \lambda_{T,2} = \sqrt{T}$, (iv) $\lambda_{T,1} = T^{1/6}, \lambda_{T,2} = T^{1/6}$, (v) $\lambda_{T,1} = T^{1/6}, \lambda_{T,2} = \sqrt{T}$, and (vi) $\lambda_{T,1} = \sqrt{T}, \lambda_{T,2} = \sqrt{T}$. (i)-(iii) are out of the scope of Section 4.2. However, as noted in Remark 20, (iv)-(v) are covered, in spite of violating N7(b), especially since $g(Z_t; \theta)$ is linear in $\theta$. (i)-(vi) are under the scope of Section 4.1 if interest lies in the empirical size.)

We take $R = [1, 1]$ in (2). We consider two choices of $S$ in (3): $S^* = [1, 0]$ and $S^\dagger = [0, 1]$ giving the true $\gamma_\theta$ as $\gamma_{S^*}^0 = \theta_1^0$ and $\gamma_{S^\dagger}^0 = \theta_2^0$. We consider six different tests below, three for each choice of $S$:

- The infeasible test in (8).
- The standard plug-in test that rejects $H_0$ if $LM_T(r_0, \gamma_{\theta}(r_0)) > \chi^2_{2\alpha}(1 - \alpha)$ where $\gamma_{\theta}(r_0)$ is the restricted-by-$H_0$ CU-GMM estimator of $\gamma_\theta$, i.e., $\gamma_{\theta}(r_0) := \arg\min_{\gamma \in \Gamma_S} Q_T(A^{-1}_S(r_0', \gamma)')$ with $Q_T(.)$ as in (11). Here, $\gamma_{\theta}(r_0)$ is the restricted limited information maximum likelihood (LIML) estimator.
- The two-step projection test in (6) with $CI_T^{SW}(\gamma_{\theta}; r_0, \epsilon)$ in (10) as the first-step confidence set.

**Simulation results:** Figure 1 plots the empirical rejection probability of these tests with $S = S^*, S^\dagger$ and under specifications (i)-(vi). Since the plug-in test (in this case) and the two-step projection test that

\[
\frac{\chi^2_{2\alpha}(1 - \alpha)}{\lambda_{T,1}^2} g_T(\theta_T) = O_p \left( \frac{1}{\lambda_{T,1}^2} \right) + \frac{\chi^2_{2\alpha}(1 - \alpha)}{\lambda_{T,2}^2} p_{\theta_T} + \frac{\chi^2_{2\alpha}(1 - \alpha)}{\lambda_{T,2}^2} p_{\theta_T} = o_p(1), \quad \text{a less stringent rate restriction than (15). It is also worth noting here that since identification is better than weak, i.e., } \lambda_{T,1} \rightarrow \infty, \text{ when taken in conjunction with Lemma 7, this discussion explains the part in Remark 9 that led to footnote 9.}
Figure 1: Empirical rejection probabilities of the two-step projection test (2-step) in (6) with \( \epsilon = 0.05 \); the infeasible test (infeas) in (8) with \( \alpha = 0.045 \); and the standard plug-in test (LIML) based on the restricted by \( H_0 \) CU-GMM (LIML) estimator for \( S \), with \( \alpha = 0.045 \). Two choices of \( S \): \( S_1 = \theta_1 \) and \( S_2 = \theta_2 \) respectively. Two choices of \( \gamma_S \): \( \gamma_S = 2-step \) and \( \gamma_S = \text{infeas} \). Results are based on 10,000 Monte Carlo trials.

Horizontal axis: deviation of \( H_0 \) in (2) from the truth [see (14)]. Title: Identification strength that corresponds to specifications (i)-(vi) respectively.
are both invariant to the choice of \( S \), only one plot for each test is reported.\(^{13}\) (Note that the plug-in test is not invariant if \( H_0 \) is not imposed while estimating \( \gamma_S \). This has serious adverse consequences on size, and is studied in Appendix D.) Results under different variations of this specification are similar.

It must be acknowledged that while considerable care — e.g., at least seven different starting values, far more stringent than default conditions for optimization in Matlab, cross-checking with both and other choices of \( S \), etc. — is taken in the implementation of the two-step test, it is technically possible that the second step minimization is not accurate and hence the rejection probabilities are overestimated.\(^{14}\)

We find that the simulation results under (v) and (vi) and, to a lesser extent, under (iv) corroborate our local asymptotic results even for \( T = 100 \). They are also encouraging for the two-step test under (ii) and (iii) that are actually outside our scope. When \( \lambda_{T,1} \neq \lambda_{T,2} \), the two infeasible tests behave very differently as \( H_0 \) deviates a bit far from the truth, and we find that the two-step and the standard plug-in tests, both of which are invariant to \( S \), resemble the infeasible test using \( \gamma_S = \theta_2 \), the better identified component of \( \theta^0 \).\(^{15}\) In other words, the simulation results demonstrate that the rejection rate of the feasible test, i.e., the two-step projection test (and the standard plug-in test) can only be as good as that of the less powerful infeasible test. This observation is intuitive and this also confirms the novel aspect of the local efficiency result, i.e., the key feature (F2), that we sought to highlight in our paper.

References

\(^{13}\)The invariance to \( S \) for the two-step in (6) follows once we note (a) and (b) below. (a) For any \( S \), the optimization problem in (6) is: min_{\gamma_S} LM_T(A_S^{-1}(r_0; \gamma_S')) such that \( h_T(A_S^{-1}(r_0; \gamma_S')) \leq c \) where the inequality constraint represents the first step of the two-step test with \( h_T(\cdot) \) as the first-step test statistic and \( c \) as the critical value. (\( h_T(\cdot) = Q_T(\cdot) \) and \( c = 1 - \epsilon \) in our case; see (10).) (b) The minimization problem in (a) is the same as: min_{\gamma_S} LM_T(\theta) such that \( h_T(\theta) \leq c \) and \( R\theta = r_0 \), where the last equality relates to (2) and (3). The problem in (b) does not depend on \( S \). Hence, the invariance.

\(^{14}\)This is more likely for the results in Appendix D, but not problematic in that context since any possible overestimation actually reinforces the point about the poor power of the unrestricted-by-\( H_0 \) version of the test that we wish to make there.

\(^{15}\)More extensive simulations (1 million trials and grid size .001) not reported here suggest that when \( \lambda_{T,1} \neq \lambda_{T,2} \), the two infeasible tests behave similarly roughly in the interval \([-1,1]\) around \( r_0 - \beta \). Their difference under (i) and (iv), where \( \lambda_{T,1} = \lambda_{T,2} \), is due to small sample size; the difference vanishes if, for example, \( T = 2000 \) for (i) and \( T = 250 \) for (iv).


Appendix A: Important constructions and definitions for Section 4

A.1 UBT and LBT Constructions:

We extensively use the following constructions that are adapted from the original work of Antoine and Renault (2012), Andrews and Cheng (2014), Cheng (2015), etc. Let \( \{W_T = [W_{T,1}, \ldots, W_{T,m_T}] : T \geq 1\} \) be a sequence of \( r \times c \) (for some \( r, c \)) matrix of full row-rank \( r(\leq c) \) where \( W_{T,j} \) is \( r \times c_{T,j} \) (and empty if \( c_{T,j} = 0 \)) for \( j = 1, \ldots, m_T \) and such that \( \sum_{j=1}^{m_T} c_{T,j} = c \) for each \( T \geq 1 \).

A.1.1 UBT-Construction: An upper block-triangular (UBT) construction

We construct a sequence of \( r \times r \) matrix \( \Pi_T = [\Pi_{T,1}, \ldots, \Pi_{T,m_T}] : T \geq 1 \) such that the \( c \times r \) matrix \( W'_T \Pi_T \) has an UBT structure for each \( T \geq 1 \). For any given \( T \), the following steps give such a \( \Pi_T \).

- Let \( \text{rank}(W_{T,m_T}) = c^*_{T,m_T} \leq \min(r, c_{m_T}) \). Define \( \Pi_{T,m_T} \) as the \( r \times c^*_{T,m_T} \) matrix such that its columns form an orthogonal basis for the column space of \( W'_{T,m_T} \). Stop if \( m_T = 1 \).

- Let \( \text{rank}([W_{T,m_T-1}, W_{T,m_T}]) - \text{rank}(W_{T,m_T}) = c^*_{T,m_T-1} \leq \min(r, c_{m_T-1}) \). Define \( \Pi_{T,m_T-1} \) as the \( r \times c^*_{T,m_T-1} \) matrix such that the columns of \([\Pi_{T,m_T-1}, \Pi_{T,m_T}]\) form an orthogonal basis for the column space of \([W_{T,m_T-1}, W_{T,m_T}]\)'s. Stop if \( m_T = 2 \).

- Continue step-by-step, as above, for \( j = m_T - 2, \ldots, 1 \) and for each \( j \), define \( \Pi_{T,j} \) as the \( r \times c^*_{T,j} \) matrix, where \( c^*_{T,j} = \text{rank}([W_{T,j}, \ldots, W_{T,m_T}]) - \text{rank}([W_{T,j+1}, \ldots, W_{T,m_T}]) \leq \min(r, c_{j+1}) \), such that the columns of \([\Pi_{T,j}, \ldots, \Pi_{T,m_T}]\) form an orthogonal basis for the column space of \([W_{T,j}, \ldots, W_{T,m_T}]\)'s.

As a convention, \( \Pi_{T,j} \) is an empty matrix if \( c^*_{T,j} = 0 \). \( \Pi_T \) is an orthogonal matrix by construction and

(i) for some integer \( q_T \in \{1, \ldots, \min(r, m_T)\} \), the \( q_T \) blocks \( W'_{T,j+1 \cdot T} \Pi_{T,j+1 \cdot T} \) for \( k = 1, \ldots, q_T \), and where \( 1 \leq j+1 \cdot T < \ldots < j+q_T \cdot T \leq m_T \), each has full column-rank \( c^*_{T,j+1 \cdot T} > 0 \) satisfying \( \sum_{k=1}^{q_T} c^*_{T,j+1 \cdot T} = r \);

(ii) \( W'_{T,j \cdot T} \Pi_{T,k} = 0 \), a zero matrix of suitable (according to the above) dimension, for all \( 1 \leq k < j \leq m_T \).

A.1.2 LBT-Construction: A lower block-triangular (LBT) construction

We construct a sequence of \( r \times r \) matrix \( \{\Pi_T = [\Pi_{T,1}, \ldots, \Pi_{T,m_T}] : T \geq 1\} \) such that the \( c \times r \) matrix \( W'_T \Pi_T \) has a BLT structure for each \( T \geq 1 \). For any given \( T \), the following steps give such a \( \Pi_T \). (This is same as the UBT-Construction, but in reverse order. Hence to save new notation, we continue to use the same notation as in the UBT-Construction and hope that this is not confusing.)

- Let \( \text{rank}(W_{T,1}) = c^*_{T,1} \leq \min(r, c_1) \). Define \( \Pi_{T,1} \) as the \( r \times c^*_{T,1} \) matrix such that its columns form an orthogonal basis for the column space of \( W'_{T,1} \). Stop if \( m_T = 1 \).

- Let \( \text{rank}([W_{T,1}, W_{T,2}]) - \text{rank}(W_{T,1}) = c^*_{T,2} \leq \min(r, c_2) \). Define \( \Pi_{T,2} \) as the \( c \times c^*_{T,2} \) matrix such that the columns of \([\Pi_{T,1}, \Pi_{T,2}]\) form an orthogonal basis for the column space of \([W_{T,1}, W_{T,2}]\)'s. Stop if \( m_T = 2 \).
Continue step-by-step, as above, for $j = 3, \ldots, m_T$ and for each $j$, define $\Pi_{T,j}$ as the $r \times c^*_{T,j}$ matrix, where $c^*_{T,j} = \text{rank}([W_{T,1}, \ldots, W_{T,j}]) - \text{rank}([W_{T,1}, \ldots, W_{T,j-1}]) \leq \min(r, c_{T,j})$, such that the columns of $[\Pi_{T,1}, \ldots, \Pi_{T,j}]$ form an orthogonal basis for the column space of $[W_{T,1}, \ldots, W_{T,j}]'$.

As a convention, $\Pi_{T,j}$ is an empty matrix if $c^*_{T,j} = 0$. $\Pi_T$ is an orthogonal matrix by construction and

(i) for some integer $q_T \in \{1, \ldots, \min(r, m_T)\}$, the $q_T$ blocks $W'_{T,jk} \Pi_{T,jk}$ for $k = 1, \ldots, q_T$, and where $1 \leq j_1, T < \ldots < j_{q_T}, T \leq m_T$, each has full column-rank $c^*_{T,jk} > 0$ satisfying $\sum_{k=1}^{q_T} c^*_{T,jk} = r$;

(ii) $W'_{T,j} \Pi_{T,k} = 0$, a zero matrix of suitable (according to the above) dimension, for all $1 \leq j < k \leq m_T$.

### A.2 Construction of $\Pi_{\rho_0}, D_{T,\rho_0}$ and $G^*$:

The efficient rate-disentangled directions of $\theta$ that are identified from (13) under our assumptions are given by $\Pi_{\rho_0}^{-1} \theta$ where $\Pi_{\rho_0}$ is a $d_\theta \times d_\theta$ orthogonal matrix, and the appropriate rates along these directions, in the given order, are given by the $d_\theta \times d_\theta$ diagonal matrix $\sqrt{T}D_{T,\rho_0}^{-1}$ [see Antoine and Renault (2012)].

#### A.2.1 Construction of $\Pi_{\rho_0}$:

Let $\rho_0 := \rho_0(\theta^0)$, i.e., $\partial \rho(\theta^0)/\partial \theta'$. Using N3 write $I^*\Lambda_T \rho_0 \equiv I^*\Lambda_T I' I^* \rho_0 = [\lambda_{T,1}, \rho'_{0,1}, \ldots, \lambda_{T,l}, \rho'_{0,l}]'$ where $\rho_{0,j}(\theta)$ is a $k_j \times d_\theta$ matrix for $j = 1, \ldots, l$. Take $W_T = [\rho'_{0,1}, \ldots, \rho'_{0,l}] = (I^* \rho_0)'$ (not depending on $T$) in the UBT-Construction in Appendix A.1.1. To emphasize the non-dependence on $T$, write $W_T$ as $W$, and accordingly write the rest of the notation from the UBT-Construction. Thus $r = d_\theta$, $c = d_g$ and $m = l$ in terms of the notation from the UBT-Construction. $W$ is full row-rank $r$ (= $d_\theta$) by N4.

$$\Pi_{\rho_0} = [\Pi_{\rho_0,1}, \ldots, \Pi_{\rho_0,l}]$$ is the $d_\theta \times d_\theta$ matrix $\Pi$ from the UBT-Construction with $W = (I^* \rho_0(\theta^0))'$. \hspace{1cm} (18)

#### A.2.2 Construction of $D_{T,\rho_0}$:

The construction of $D_{T,\rho_0}$ depends on the matrix $I^*\Lambda_T I' I^* \rho_0(\theta^0)\Pi_{\rho_0}$. Let $c^*_{\rho_0,j} = c^*_j \geq 0$ denote the number of columns of $\Pi_{\rho_0,j}$ for $j = 1, \ldots, l$, and $q_{\rho_0} = q$ from (i) in the UBT-Construction (of $\Pi_{\rho_0}$). Let $(j_1, \ldots, j_{q_{\rho_0}})$ denote the indices such that the block $\rho_{0,j_i}\Pi_{\rho_0,j_i}$ of dimension $k_{j_i} \times c^*_{\rho_0,j_i}$ is full column-rank $c^*_{\rho_0,j_i} > 0$ for $i = 1, \ldots, q_{\rho_0}$ and $\sum_{i=1}^{q_{\rho_0}} c^*_{\rho_0,j_i} = d_\theta$. Thus, the corresponding block of $I^*\Lambda_T I' I^* \rho_0(\theta^0)\Pi_{\rho_0}$ is $\lambda_{T,j_i}\rho_{0,j_i}\Pi_{\rho_0,j_i}$. Accordingly, for $I^*\Lambda_T I' I^* \rho_0(\theta^0)\Pi_{\rho_0}$, the columns from $(d_\theta - \sum_{i'=1}^{q_{\rho_0}} c^*_{\rho_0,j_i'})$ to $(d_\theta - \sum_{i'=1}^{q_{\rho_0}} c^*_{\rho_0,j_i'})$ for $i = 1, \ldots, q_{\rho_0}$ are represented by the $d_g \times c^*_{\rho_0,j_i}$ matrix:

$$\left[ \lambda_{T,1}(\rho_{0,1}, \Pi_{\rho_0,1})', 0 \right]'$$ if $j_i = 1$,  

$$\left[ \lambda_{T,1}(\rho_{0,1}, \Pi_{\rho_0,1})', \ldots, \lambda_{T,j_i}(\rho_{0,j_i}, \Pi_{\rho_0,j_i})', 0 \right]'$$ otherwise.

In both cases: $j_i = 1$ and $j_1 > 1$, the 0’s inside the big matrices denote sub-matrices of zeros with
number of rows, which can be zero, such that the number of rows of the corresponding big matrix is $d_g$.

Now, conforming to this above structure, define the $d_g \times d_g$ matrix $D_{T,\rho_0}$ as:

$$D_{T,\rho_0} := \sqrt{T} \text{diag} \left( \lambda_{T,j_1}^{-1} 1_{c_{\rho_0,j_1}}, \ldots, \lambda_{T,j_{q_{\rho_0}}}^{-1} 1_{c_{\rho_0,j_{q_{\rho_0}}}} \right).$$  \hspace{1cm} (19)

A.2.3 Construction of $G^*$:

Define the $d_g \times d_g$ matrix $G^\dagger$ as the following limit:

$$G^\dagger := \lim_{T \to \infty} \frac{1}{\sqrt{T}} I^* \Lambda_T I^\prime I^* \rho_0(\theta^0) \Pi_{\rho_0} D_{T,\rho_0}. \hspace{1cm} (20)$$

By construction, $G^\dagger$ is finite, and its columns from $(d_g - \sum_{i=1}^{q_{\rho_0}} c_{\rho_0,j_i}^*)$ to $(d_g - \sum_{i=1}^{q_{\rho_0}} c_{\rho_0,j_i}^* + c_{\rho_0,j_i}^*)$ for $i = 1, \ldots, q_{\rho_0}$ are represented by the $d_g \times c_{\rho_0,j_i}^*$ matrix:

$$[0', (\rho_{\theta,j_i} \Pi_{\rho_0,j_i}, 0')]' \text{ otherwise.}$$

As above, 0 denotes sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding matrix is $d_g$. Naturally, under our assumptions $G^\dagger$ is full column-rank.

Now define the $d_g \times d_g$ finite matrix of full column-rank $G^*$ as:

$$G^* := I^* G^\dagger. \hspace{1cm} (21)$$

A.3 Construction of $\Pi_R$, $D_{T,R}$ and $R^*$:

$\Pi_R$ and $D_{T,R}$ are quantities used to characterize the appropriate local deviation of the null from the truth. Their construction depends on the constructions of $\Pi_{\rho_0}$ and $D_{T,\rho_0}$.

A.3.1 Construction of $\Pi_R$:

Take $W_T = R \Pi_{\rho_0} = [W_{T,1} = R \Pi_{\rho_0,j_1}, \ldots, W_{T,q_{\rho_0}} = R \Pi_{\rho_0,j_{q_{\rho_0}}}]$ (not depending on $T$) in the LBT-Construction in Appendix A.1.2. Note that the partition of $W_T$ was informed by the indices $j_1, \ldots, j_{q_{\rho_0}}$ defined immediately after constructing $\Pi_{\rho_0}$. These indices do not depend on $T$, and they also informed the construction of $D_{T,\rho_0}$. Once again, to emphasize the non-dependence on $T$, write $W_T$ as $W$, and accordingly write the rest of the notation from the LBT-Construction. Thus $r = d_R$, $c = d_\theta$ and $m = q_{\rho_0}$.

$W$ is full row-rank by the definition of $R$, $\Pi_{\rho_0}$ and Lemma 10 (in Appendix C).

$$\Pi_R = [\Pi_{R,1}, \ldots, \Pi_{R,q_{\rho_0}}]$$ is the $d_R \times d_R$ matrix $\Pi_T$ from the LBT-Construction with $W = R \Pi_{\rho_0}. \hspace{1cm} (22)$$
A.3.2 Construction of $D_{T,R}$:

The construction of $D_{T,R}$ depends on the matrix $D_{T,p_0} \Pi_{p_0} R' \Pi R$. Let $c_{R,j}^* = c_j^* \geq 0$ denote the number of columns of $\Pi_{R,j}$ for $j = 1, \ldots, q_{p_0}$, and $q_R = q$ from (i) in the LBT-Construction (of $\Pi_R$). Let $(j_{n_1}, \ldots, j_{n_{q_R}})$ denote the sub-indices of the indices $(j_1, \ldots, j_{q_{p_0}})$ such that the block $\Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}$ of dimension $c_{p_0,j_{n_1}} \times c_{R,n_1}^*$ is full column-rank $c_{R,n_1}^* > 0$ for $i = 1, \ldots, q_R$, and $\sum_{t=1}^{q_R} c_{R,n_1}^* = d_R$. Thus, the corresponding block of $D_{T,p_0} \Pi'_{p_0} R' \Pi R$ is $\frac{\sqrt{T}}{\lambda_{T,j_{n_1}}} \Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}$. Accordingly, for $D_{T,p_0} \Pi'_{p_0} R' \Pi R$, the columns from $(d_R - \sum_{t=1}^{q_R} c_{R,n_1}^* R_{n_1}^*)$ to $(d_R - \sum_{t=1}^{q_R} c_{R,n_1}^* + c_{R,n_1}^*)$ for $i = 1, \ldots, q_R$ are represented by the $d_\theta \times c_{R,n_1}^*$ matrix:

$$
\left[
\begin{array}{c}
0', \left(\Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}\right)'
\end{array}
\right]' \text{ if } n_i = q_{p_0},
$$

$$
\left[
\begin{array}{c}
0', \frac{\sqrt{T}}{\lambda_{T,j_{n_1}}} \left(\Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}\right)', \ldots, \frac{\sqrt{T}}{\lambda_{T,j_{n_1}}} 
\end{array}
\right]' \text{ otherwise.}
$$

In both cases, 0 represents the sub-matrix of zeros with number of rows that make the number of rows of the corresponding matrix equal to $d_\theta$.

Now, conforming to this above structure, define the $d_R \times d_R$ matrix $D_{T,R}$ as:

$$
D_{T,R} := T^{-1/2} \text{diag} \left(\lambda_{T,j_{n_1}} 1_{c_{R,n_1}^*}, \ldots, \lambda_{T,j_{n_{q_R}}} 1_{c_{R,n_{q_R}}^*}\right). 
$$

(23)

A.3.3 Construction of $R^*$:

Define the $d_R \times d_\theta$ matrix $R^*$ as the transpose of the following limit:

$$
R^{*'} := \lim_{T \to \infty} D_{T,p_0} \Pi'_{p_0} R' \Pi R D_{T,R}.
$$

By construction, $R^{*'}$ is finite, and its columns from $(d_R - \sum_{t=1}^{q_R} c_{R,n_1}^*)$ to $(d_R - \sum_{t=1}^{q_R} c_{R,n_1}^* + c_{R,n_1}^*)$ for $i = 1, \ldots, q_R$ are represented by the $d_\theta \times c_{R,n_1}^*$ matrix:

$$
\left[
\begin{array}{c}
0', \left(\Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}\right)'
\end{array}
\right]' \text{ if } n_i = q_{p_0},
$$

$$
\left[
\begin{array}{c}
0', \left(\Pi'_{p_0,j_{n_1}} R' \Pi_{R,n_1}\right)', 0'
\end{array}
\right]' \text{ otherwise.}
$$

(As above, 0 denotes sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding matrix is $d_\theta$). Naturally, under our assumptions $R^*$ is full row-rank.
Supplemental Appendix B: For the references from Section 3

Appendix B.1: Efficient influence function for $\beta^0 := R\theta^0$ under (1)

It is well-known that under the assumptions that (1) holds, $G(\theta^0)$ is full column-rank, and $V(\theta^0)$ is positive definite: the efficient estimator of $R\theta^0$ has an asymptotically linear representation $-\sqrt{T}l_T(\theta^0) + o_p(1)$. Unfortunately, we could not find a paper to cite the proof of it. So we provide a standard proof.

**Lemma 9** Let $\{Z_t\}_{t=1}^T$ be i.i.d. copies of a random variable $Z$, and let (1) holds. If $G := \frac{\partial}{\partial \theta} E[g(Z; \theta)]|_{\theta=\theta^0}$ is a full column-rank $d_g \times d_g$ matrix and $V := E[g(Z; \theta^0)g'(Z; \theta^0)]$ is a $d_g \times d_g$ positive definite matrix, then the asymptotic variance lower bound for any regular estimator of the $d_R \times 1$ parameter vector $\beta^0 := R\theta^0$ where $d_R \leq d_g$ is $(R(G'V^{-1}G)R')^{-1}$. The regular estimator whose asymptotic variance attains this bound has the asymptotically linear representation $\sqrt{T}(\hat{\beta}^0 - \beta^0) = -\sqrt{T}l_T(\theta^0) + o_p(1)$.

**Proof:** Consider a parametric path $\xi$ of the distribution of $Z$ such that for the unique value $\xi^0$ we have the joint density $f_{\xi^0}(z) = f(z)$, the true density. Let $s_\xi(Z)$ denote the score with respect to $\xi$. Without any other restrictions, the tangent space for the model is simply $T = a(z)$ where $a(z)$ satisfies $E[a(Z)] = 0$, and $E[.]$ equivalently stands for $E_{\xi^0}[.]$. Since $d_g > d_R$, (1) equivalently requires that for any given $d_R \times d_g$ matrix $B$, the relation $B E[g(Z; \theta^0)] = 0$ holds. Take $B$ as full row-rank without loss of generality. Now, differentiating with respect to $\xi$ under the expectation we obtain $\frac{\partial \theta(\xi^0)}{\partial \xi} = -(BG)^{-1}E[Bg(Z; \theta^0)s_{\xi^0}(Z)]$ and thus $\frac{\partial \beta(\xi^0)}{\partial \xi} = -(BG)^{-1}E[Bg(Z; \theta^0)s_{\xi^0}(Z)]$. Therefore, any regular estimator for $\beta^0$ will be asymptotically linear with the influence function $\varphi(B) := -(BG)^{-1}B g(Z; \theta^0)$. Given the structure of the tangent space $T$, (1) implies that the projection of this influence function $\varphi(B)$ onto $T$ is $\varphi(B)$ itself. For this given $B$, $\text{Var}(\varphi(B)) = \Sigma(B) := (BG)^{-1}BV B'(BG)^{-1} R'$.

Thus the efficient influence function is obtained by choosing $B^* := \text{arg min}_B \Sigma(B) = G'V^{-1}$, giving $\Sigma(B^*) = (G'V^{-1}G)^{-1} R'$ and $\varphi(B^*) = -R(G'V^{-1}G)^{-1}G'V^{-1}g(Z; \theta^0)$. This completes the proof. ■

Appendix B.2: The second-step test statistic $LM_{T,S}^{ES}(\theta)$ in Chaudhuri and Zivot (2011):

Given the choice of $S$ in (3), the scores for $\beta$ and $\gamma_s$, by which we mean here the population version of the optimal rotations, in the efficient GMM sense, of $\bar{g}_T(\theta^0)$ along the directions of $\beta$ and $\gamma_s$ are:

$$l_{\beta,S,T}(\theta) := R_S^T G'(\theta)V^{-1}(\theta)\bar{g}_T(\theta)$$

and

$$l_{\gamma_s,S,T}(\theta) := S_S^T G'(\theta)V^{-1}(\theta)\bar{g}_T(\theta)$$

respectively. It is important to note that while the definition of $\beta := R\theta$ does not depend on $S$, the score for $\beta$ in the re-parameterized model generally depends on $S$ through $R_S^T$ [see Remark 21].
Following Chaudhuri and Zivot (2011), the efficient score for $\beta$ would be the residual from a regression:

$$l_{\beta,\gamma_S,T}(\theta) := l_{\beta,S,T}(\theta) - \text{Cov} \left( \sqrt{T} l_{\beta,S,T}(\theta), \sqrt{T} l_{\gamma_S,T}(\theta) \right) \text{Var}^{-1} \left( \sqrt{T} l_{\gamma_S,T}(\theta) \right) l_{\gamma_S,T}(\theta).$$

Define $\Omega(\theta) := G'(\theta) V^{-1}(\theta) G(\theta)$. Then, NM-9.2 gives $\sqrt{T} \left( l_{\beta,\gamma_S,T}(\theta) - E \left[ l_{\beta,\gamma_S,T}(\theta) \right] \right) \overset{d}{\to} N(0, \Xi(\theta))$ where

$$\Xi_S(\theta) := \left( R_S^1 \Omega(\theta) R_S^1 \right) - \left( R_S^1 \Omega(\theta) S_S^1 \right) \left( S_S^1 \Omega(\theta) S_S^1 \right)^{-1} \left( S_S^1 \Omega(\theta) R_S^1 \right)$$

$$= R_S^1 G'(\theta) V^{-1/2}(\theta) \left( I_{d_y} - P \left( V^{-1/2}(\theta) G(\theta) S_S^1 \right) \right) V^{-1/2}(\theta) G(\theta) R_S^1.$$

Using the definitions of $G(\theta)$ and $V(\theta)$, the feasible version for $l_{\beta,\gamma_S,T}(\theta)$ (as is $\hat{l}_T(\theta)$ for $l_T(\theta)$) is:

$$\hat{l}_{\beta,\gamma_S,T}(\theta) = R_S^1 \hat{G}_T(\theta) \hat{V}_T^{-1/2}(\theta) \left( I_{d_y} - P \left( \hat{V}_T^{-1/2}(\theta) \hat{G}_T(\theta) S_S^1 \right) \right) \hat{V}_T^{-1/2}(\theta) \hat{g}_T(\theta),$$

and, similarly, $\hat{\Xi}_T(\theta)$ for $\Xi_T(\theta)$. Then, the statistic in Chaudhuri and Zivot (2011) would be defined as:

$$LM_{T,S}^{ES}(\theta) := T \times \hat{l}_{\beta,\gamma_S,T}(\theta) \hat{\Xi}_S^{-1}(\theta) \hat{l}_{\beta,\gamma_S,T}(\theta)$$

$$= T \times \left( \hat{V}_T^{-1/2}(\theta) \hat{g}_T(\theta) \right)' P \left( \left( I_{d_y} - P \left( \hat{V}_T^{-1/2}(\theta) \hat{G}_T(\theta) S_S^1 \right) \right) \hat{V}_T^{-1/2}(\theta) \hat{G}_T(\theta) R_S^1 \right) \left( \hat{V}_T^{-1/2}(\theta) \hat{g}_T(\theta) \right).$$

Chaudhuri and Zivot (2011) noted that Remark 5 from Section 3 is equally applicable to $LM_{T,S}^{ES}(\theta)$.

**Appendix B.3: Proofs of Lemmas 1 and 2 from Section 3:**

We will repeatedly use the following relations that follow since $A_S = [R', S']'$ and $A_S^{-1} = [R_S^1, S_S^1]$:

$$RR_S^1 = I_{d_R}, \quad RS_S^1 = 0, \quad SR_S^1 = 0, \quad SS_S^1 = I_{d_S - d_R} \quad \text{and} \quad R_S^1 R + S_S^1 S = I_{d_S}. \quad (25)$$

We will suppress the dependence of the quantities on $\theta$ to avoid notational clutter.

**Proof of Lemma 1:** Consider any $(d_0 - d_R) \times d_0$ full row-rank matrix $S$ in (3) such that $[R', S']'$ is nonsingular. Let $\zeta$ be a $d_0 \times (d_0 - d_R)$ matrix whose columns form a basis for the null space of $R$. Therefore, since $RS_S^1 = 0$ by (25) while $S_S^1$ is full column-rank by definition, we have $S_S^1 = \zeta B_S$ for some $(d_0 - d_R) \times (d_0 - d_R)$ nonsingular matrix $B_S$. Therefore, for any two such choices of $S = S^*, S^1$, we have the corresponding $S_S^{1*} = \zeta B_{S^*}$ and $S_S^{11} = \zeta B_{S^1}$ for some $(d_0 - d_R) \times (d_0 - d_R)$ nonsingular matrices $B_{S^*}$ and $B_{S^1}$. This implies that $S_S^{1*} = S_S^{11} B$ where $B = B_{S^1}^{-1} B_{S^*}$ is $(d_0 - d_R) \times (d_0 - d_R)$ and nonsingular.
Now for any \( d_\theta \times d_\theta \) nonsingular matrix \( M = [M_1, M_2] \), where \( M_1 \) is \( d_\theta \times d_R \), define:

\[
\Phi_T(M) := T \times \left( \tilde{V}_T^{-1/2} g_T \right)^T \left( \tilde{V}_T^{-1/2} \tilde{G}_T M \right) \left( \tilde{V}_T^{-1/2} g_T \right), \\
\Phi_{1,2,T}(M) := T \times \left( \tilde{V}_T^{-1/2} g_T \right)^T \left( \left( I_{d_\theta} - P \tilde{V}_T^{-1/2} \tilde{G}_T M_1 \right) \tilde{V}_T^{-1/2} \tilde{G}_T M_1 \right) \left( \tilde{V}_T^{-1/2} g_T \right), \\
\Phi_{2,T}(M_2) := T \times \left( \tilde{V}_T^{-1/2} g_T \right)^T \left( \tilde{V}_T^{-1/2} \tilde{G}_T M_2 \right) \left( \tilde{V}_T^{-1/2} g_T \right),
\]

and note that, by construction:

(i) \( \Phi_T(M) = \Phi_T(I_{d_\theta}) \) since \( M \) is nonsingular,

(ii) \( \Phi_T(M) = \Phi_{1,2,T}(M) + \Phi_{2,T}(M_2) \) since \( M \) is partitioned as \( M = [M_1, M_2] \),

(iii) \( \Phi_{2,T}(M_2) = \Phi_{2,T}(M_2B) \) since \( B \) is a \((d_\theta - d_R) \times (d_\theta - d_R)\) nonsingular matrix.

In the above, now take \( M = M^*, M^1 \) where \( M^* = [R_{S^*}, S_{S^*}] \) and \( M^1 = [R_{S^1}, S_{S^1}] \) correspond to the two choices \( S = S^*, S^1 \) respectively. Thus we obtain:

\[
\Phi_T(M^*) = \Phi_T(M^1) \quad \text{[by (i)]} \\
\Phi_{1,2,T}(M^*) + \Phi_{2,T}(M^2) = \Phi_{1,2,T}(M^1) + \Phi_{2,T}(M^2) \quad \text{[by (ii)]} \\
\Phi_{1,2,T}(M^*) = \Phi_{1,2,T}(M^1) \quad \text{[by (iii), since \( M^2 := S_{S^*} = S_{S^1}, B =: M^1 \)]}.
\]

Thus, by its definition, we obtain that \( LM_{T,S^*,S^1}^{ES}(\theta) = \Phi_{1,2,T}(M^*) = \Phi_{1,2,T}(M^1) = LM_{T,S^*,S^1}^{ES}(\theta) \). \( \blacksquare \)

**Proof of Lemma 2:** Define \( \hat{\Omega}_T := \hat{G}_T^T(\theta) \hat{V}_T^{-1}(\theta) \hat{G}_T(\theta) \). Thus, the null space of \( R_{\hat{\Omega}_T}^{-1} \) is of dimension \( d_\theta - d_R \) since \( R \) is full row rank and \( \hat{\Omega}_T \) is nonsingular. Now, consider a \((d_\theta - d_R) \times d_\theta\) matrix \( S \) whose rows form the basis for the null space of \( R_{\hat{\Omega}_T}^{-1} \). Note that, this specific choice of \( S \) and, hence, the related quantities \( R_{S^1}^1 \) and \( S_{S^1}^1 \) obtained from it all depend on the given \( T \) in the statement of the lemma.

**Claim 1:** With this \( S \), we have a nonsingular \( A_S := [R^r, S^r]' \) in (3).

**Proof:** Suppose not. Then, the full row-rank of \( R = [R^r_1, \ldots, R^r_{d_R}]' \) implies that there exists a \((d_\theta - d_R) \times 1\) vector \( c \neq 0 \) such that \( R_1 = \sum_{j=2}^{d_R} a_j R_j + c' S \) for some scalar coefficients \( a_2, \ldots, a_{d_R} \). Since \( \hat{\Omega}_T^{-1} \) is positive definite, it means that for this \( c \neq 0 \), we have \( R_1 \hat{\Omega}_T^{-1} = \sum_{j=2}^{d_R} a_j R_j \hat{\Omega}_T^{-1} + c' S \hat{\Omega}_T^{-1} \). Post-multiply both sides by \( S' \) and note that the rows of \( S \) belong in the null space of \( R_{\hat{\Omega}_T}^{-1} \), i.e. \( R_j \hat{\Omega}_T^{-1} S' = 0 \) for \( j = 1, \ldots, d_R \). Hence, it follows that \( 0 = c' S \hat{\Omega}_T^{-1} S' \). Since \( S \hat{\Omega}_T^{-1} S' \) is positive definite (as \( \hat{\Omega}_T^{-1} \) is positive definite and as the rows of \( S \) are linearly independent), this is only possible if \( c = 0 \), which contradicts our supposition. Therefore, Claim 1 is true. \( \blacksquare \)

**Claim 2:** \( R_{\hat{\Omega}_T}^{-1} S' = 0 \) if and only if \( R^r S_{\hat{\Omega}_T} S^r_S = 0 \).
Proof: We use (25) repeatedly in this proof. Post-multiply $R_S^1\tilde{\Omega}_T S_S^1 = 0$ by $S$ to get $R_S^1\tilde{\Omega}_T S_S^1 S = 0$, i.e., $R_S^1\tilde{\Omega}_T (I_{d_{\theta}} - R_S^1 R) = 0$ by (25). Hence,

$$R = (R_S^1\tilde{\Omega}_T R_S^1)^{-1} R_S^1\tilde{\Omega}_T.$$  \hfill (29)

Similarly obtain $S = (S_S^1\tilde{\Omega}_T S_S^1)^{-1} S_S^1 \tilde{\Omega}_T$. Thus, $R\tilde{\Omega}_T^{-1} S' = (R_S^1\tilde{\Omega}_T R_S^1)^{-1} (R_S^1\tilde{\Omega}_T S_S^1) (S_S^1\tilde{\Omega}_T S_S^1)^{-1}$. Hence, $R\tilde{\Omega}_T^{-1} S' = 0$ if and only if $R_S^1\tilde{\Omega}_T S_S^1 = 0$, once again by using the positive definiteness of $\tilde{\Omega}_T$. □

Thus, using this specific choice of $S$ for which $R\tilde{\Omega}_T^{-1} S' = 0$ and hence $R_S^1\tilde{\Omega}_T S_S^1 = 0$, we obtain from the definition of $LM_{T,S}^{FS}(\theta)$ that $LM_{T,S}^{FS}(\theta) = T \times \left( \hat{V}_T^{-1/2} g_T \right)' P \left( \hat{V}_T^{-1/2} \hat{G}_T R_S^1 \right) \left( \hat{V}_T^{-1/2} g_T \right)$. On the other hand, (4) gives:

$$LM_T(\theta) = T \times \left( \hat{V}_T^{-1/2} g_T \right)' P \left( \hat{V}_T^{-1/2} \hat{G}_T \tilde{\Omega}_T^{-1} R_S^1 \right) \left( \hat{V}_T^{-1/2} g_T \right)$$

by using (29). However, since $(R_S^1\tilde{\Omega}_T R_S^1)^{-1}$ is nonsingular, we have, by the construction of the projection matrix $P(\cdot)$, that $P \left( \hat{V}_T^{-1/2} \hat{G}_T R_S^1 (R_S^1\tilde{\Omega}_T R_S^1)^{-1} \right) = P \left( \hat{V}_T^{-1/2} \hat{G}_T R_S^1 \right)$. Therefore, $LM_T(\theta) = T \times \left( \hat{V}_T^{-1/2} g_T \right)' P \left( \hat{V}_T^{-1/2} \hat{G}_T R_S^1 \right) \left( \hat{V}_T^{-1/2} g_T \right) = LM_{T,S}^{FS}(\theta)$ (see above for the last equality). The desired result now follows from Lemma 1 for any general choice of $S$ in (3) such that $[R', S']'$ is nonsingular. □

Remark 21: The particular choice of $S$ employed to facilitate the proof of Lemma 2 has an interesting interpretation. To see it, consider the analogous population version of $S$, i.e., $S$ such that $R\Omega^{-1} S' = 0$. Similar to the proof of Claim 1 above, it can be shown that $[R', S']'$ is nonsingular. Similar to the proof of Claim 2 above, it can be shown that $R\Omega^{-1} S' = 0$ if and only if $R_S^1\Omega S_S^1 = 0$, where the $R_S^1$ and $S_S^1$ correspond to this particular choice of $S$. Now, note from the discussion in Appendix B.2 that with this particular choice of $S$, the score for $\beta$, i.e., $l_{\beta,S,T}(\theta)$ is identical to the efficient score for $\beta$, i.e., $l_{\beta,\gamma_S,S,T}(\theta)$. In other words, this particular choice of $S$ in the re-parameterization (3) directly makes the scores for $\beta$ and $\gamma_S$ uncorrelated and, by asymptotic normality, asymptotically independent. A followup along this line in the case of nonlinear null hypotheses is the topic of our ongoing research.

B.4 $\bar{LM}_T(\tilde{\theta}_T) = LM_T(\tilde{\theta}_T)$

From (26)-(28) and the definition in (3) it follows that $\bar{LM}_T(\theta) = LM_T(\theta) + \Phi_{2,T}(S_S^1, \theta)$ for all $\theta$ where the underlying quantities are defined. (Note that, by $\Phi_{2,T}(S_S^1, \theta)$ we mean $\Phi_{2,T}(S_S^1, \theta)$ with $g_T$, $\hat{G}_T$ and $\hat{V}_T$ evaluated at $\theta$.) Now, by the definition of the $\tilde{\theta}_T$, i.e., $(R_S^1 r_0 + S_S^1 \gamma_T)$ where $\gamma_T$ is the GMM estimator of $\gamma$ under the restriction that $\beta = r_0$, it follows from the first order condition of the GMM optimization
problem that $\Phi_{2,T}(S_{S}^{1}, \tilde{\theta}_{T}) = 0$. This is because $\Phi_{2,T}(S_{S}^{1}, \theta)$ is simply a quadratic form of the first derivative of the GMM objective function with respect to $\gamma_{S}$, which is zero when evaluated at $\tilde{\theta}_{T}$. Thus, $\tilde{LM}_{T}(\tilde{\theta}_{T}) = LM_{T}(\tilde{\theta}_{T})$. ■

Supplemental Appendix C: Proofs and clarifications for Section 4

C.1 Two useful lemmas

Since we use (have used) the following result in Lemma 10 often, let us state it here for reference.

**Lemma 10** Let $X$ be an $a \times b$ matrix, and $P$ and $Q$ be $a \times a$ and $b \times b$ nonsingular matrices. Then, $\text{rank}(X) = \text{rank}(PX) = \text{rank}(XQ)$.

**Proof:** $\text{rank}(X) \geq \text{rank}(PX) \geq \text{rank}(P^{-1}PX) = \text{rank}(X) \geq \text{rank}(XQ) \geq \text{rank}(XQQ^{-1}) = \text{rank}(X)$. ■

Lemma 11 lists a set of intermediate results useful for proving the results in the main text.

**Lemma 11** Let assumptions $O$ and $N$ hold. Consider a sequence $\{\theta_{T} = R_{S}^{1}r_{0} + S_{S}^{1}\gamma_{S,T} : T \geq 1\}$ where $r_{0}$ satisfies (14) and $\{\gamma_{S,T} : T \geq 1\}$ is such that $\theta_{T}$ satisfies (15). Then, the following results hold as $T \to \infty$:

(a) $\hat{V}_{T}(\theta_{T}) \xrightarrow{P} V(\theta^{0}) \equiv V$.

(b) $\hat{V}_{Gg,T}(\theta_{T}) \xrightarrow{P} V_{Gg}(\theta^{0}) \equiv V_{Gg}$.

(c) $\tilde{G}_{T}(\theta_{T})\Pi_{\rho_{g}}D_{T,\rho_{g}} \xrightarrow{P} G^{*}$ where $\Pi_{\rho_{g}}, D_{T,\rho_{g}}$ and $G^{*}$ are as defined in (18), (19) and (21) respectively.

(d) $\sqrt{T}\tilde{g}_{T}(\theta_{T}) = \sqrt{T}\tilde{g}_{T}(\theta^{0}) + G^{*}\mu_{T,\theta} + o_{p}(1)$ where $G^{*}$ and $\mu_{T,\theta}$ are as defined in (21) and (15) respectively.

(e) $[\hat{V}_{1,g,T}(\theta_{T})\hat{g}_{T}^{-1}(\theta_{T})\tilde{g}_{T}(\theta_{T}), \ldots, \hat{V}_{d_{g},g,T}(\theta_{T})\hat{g}_{T}^{-1}(\theta_{T})\tilde{g}_{T}(\theta_{T})] \Pi_{\rho_{g}}D_{T,\rho_{g}} = o_{p}(1)$ (a $d_{g} \times d_{g}$ matrix).

(f) $\tilde{G}_{T}(\theta_{T})\Pi_{\rho_{g}}D_{T,\rho_{g}} \xrightarrow{P} G^{*}$ where $\Pi_{\rho_{g}}, D_{T,\rho_{g}}$ and $G^{*}$ are as defined in (18), (19) and (21) respectively.

**Proof:** (a) and (b) follow by assumption N8 since $\theta_{T} = \theta^{0} + o_{p}(1)$.

(c) We prove it working term-by-term in the following decomposition:

$$
\tilde{G}_{T}(\theta_{T})\Pi_{\rho_{g}}D_{T,\rho_{g}}
= [\tilde{G}_{T}(\theta_{T}) - \hat{G}_{T}(\theta^{0})] \Pi_{\rho_{g}}D_{T,\rho_{g}} + \sqrt{T} \left[ G_{T}(\theta^{0}) - \frac{\Lambda_{T}}{\sqrt{T}} \rho_{g}(\theta^{0}) \right] \frac{\Pi_{\rho_{g}}D_{T,\rho_{g}}}{\sqrt{T}} + \frac{\Lambda_{T}}{\sqrt{T}} \rho_{g}(\theta^{0}) \Pi_{\rho_{g}}D_{T,\rho_{g}}. \tag{30}
$$

From the definitions in (18) and (19) it follows that $\Pi_{\rho_{g}}D_{T,\rho_{g}} = o(\sqrt{T})$ by N3, and hence using N6 it follows that the second term on the right hand side (RHS) of (30) is $o_{p}(1)$. On the other hand, (20) and (21) imply that the third term on the RHS of (30) converges to $G^{*}$ by construction.
To complete the proof, now we show that the first term on the RHS of (30) is $o_p(1)$. We deviate from Antoine and Renault (2012) in the treatment of this term, and the result thus obtained has implications in terms of the allowable weakness of identification [see the part of Remark 9 that led to footnote 9]. Let $G_{T,i}(\theta) := \frac{\partial}{\partial \theta_i} g_T(\theta)$ denote the $i$-th column of $G_T(\theta)$ for $i = 1, \ldots, d_\theta$ (recall that $\theta = (\theta_1, \ldots, \theta_{d_\theta})'$).

Therefore, with a bad but common abuse of notation in denoting the mean values element by element, we obtain by a mean value expansion of $G_{T,i}(\theta_T)$ around $G_{T,i}(\theta^0)$ for $i = 1, \ldots, d_\theta$ that:

$$
\left[ G_{T}(\theta_T) - G_{T}(\theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} \\
= \left[ \{ \frac{\partial}{\partial \theta'} G_{T,1}(\theta_T(\theta_1)) \} (\theta_T - \theta^0), \ldots, \{ \frac{\partial}{\partial \theta'} G_{T,d_\theta}(\theta_T(\theta_{d_\theta})) \} (\theta_T - \theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} \\
= \left[ \{ \frac{\partial}{\partial \theta_1} G_{T}(\theta_T(\theta_1)) \} (\theta_T - \theta^0), \ldots, \{ \frac{\partial}{\partial \theta_{d_\theta}} G_{T}(\theta_T(\theta_{d_\theta})) \} (\theta_T - \theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} \\
= \left[ \{ \frac{\partial}{\partial \theta_1} G_{T}(\theta_T(\theta_1)) \} (\theta_T - \theta^0), \ldots, \{ \frac{\partial}{\partial \theta_{d_\theta}} G_{T}(\theta_T(\theta_{d_\theta})) \} (\theta_T - \theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} \\
= \left[ \{ \frac{\partial}{\partial \theta_i} G_{T}(\theta_T(\theta_i)) \} (\theta_T - \theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0}
$$

(31)

by twice interchanging the order in which the derivatives are taken in each of the $d_\theta$ columns. Note that, for $i = 1, \ldots, d_\theta$, we used $\theta_T(\theta_i)$ (such that $||\theta_T(\theta_i) - \theta^0|| \leq ||\theta_T - \theta^0||$) to denote the mean value in the first equality of the above equation. Recalling that $\mu_{T,\theta} = \sqrt{T} D_{T,\rho_0}^{-1} \Pi_{\rho_0}' (\theta_T - \theta^0)$ by (15), define $U_{T,i}$ for $i = 1, \ldots, d_\theta$ as the $d_q \times d_\theta$ matrix with

$$
\left\{ \frac{\partial}{\partial \theta_i} G_{T}(\theta_T(\theta_i)) \right\} (\theta_T - \theta^0) = \left\{ \frac{\partial}{\partial \theta_i} G_{T}(\theta_T(\theta_i)) \right\} \frac{\sqrt{T}}{\lambda_{T,i}} \frac{\Pi_{\rho_0} D_{T,\rho_0} \lambda_{T,j_i}}{\sqrt{T}} \frac{\lambda_{T,i}}{\sqrt{T} \lambda_{T,j_i}}
$$

in the $i$-th column and zero everywhere else. [See Remark 9 for $\lambda_{T,j_i}$.] Therefore, (31) implies that:

$$
\left[ G_{T}(\theta_T) - G_{T}(\theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} = \sum_{i=1}^{d_\theta} U_{T,i} \Pi_{\rho_0} D_{T,\rho_0}
$$

and thus,

$$
\left\| \left[ G_{T}(\theta_T) - G_{T}(\theta^0) \right] \Pi_{\rho_0} D_{T,\rho_0} \right\| \\
\leq \sum_{i=1}^{d_\theta} \left\| U_{T,i} \right\| \times \left\| \Pi_{\rho_0} D_{T,\rho_0} \right\| \\
\leq \sum_{i=1}^{d_\theta} \left\| \frac{\partial}{\partial \theta_i} G_{T}(\theta_T(\theta_i)) \frac{\sqrt{T}}{\lambda_{T,i}} \right\| \times \left\| \frac{\Pi_{\rho_0} D_{T,\rho_0} \lambda_{T,j_i}}{\sqrt{T}} \right\| \times \left\| \frac{\lambda_{T,i}}{\sqrt{T} \lambda_{T,j_i}} \right\| \\
\leq \sum_{i=1}^{d_\theta} \sup_{\theta} \left\{ \left\| \frac{\sqrt{T}}{\lambda_{T,i}} \frac{\partial}{\partial \theta_i} \frac{\lambda_T}{\sqrt{T} \rho_0(\theta)} \right\| + \left\| \frac{\sqrt{T}}{\lambda_{T,i}} \frac{\partial}{\partial \theta_i} \left[ G_{T}(\theta) - \frac{\lambda_T}{\sqrt{T} \rho_0(\theta)} \right] \right\| \right\} \times \left\| \frac{\lambda_{T,i}}{\sqrt{T} \lambda_{T,j_i}} \right\| \times \left\| \frac{\Pi_{\rho_0} D_{T,\rho_0} \lambda_{T,j_i}}{\sqrt{T}} \right\| \\
= o_p(1)
$$

since, on the third line from above, the order of magnitude of the terms (from left to right) inside the sum is respectively: (i) $\sup_{\theta} \left\| \frac{\sqrt{T}}{\lambda_{T,i}} \frac{\partial}{\partial \theta_i} \frac{\lambda_T}{\sqrt{T} \rho_0(\theta)} \right\| = O(1)$ by N3 and N4; (ii) $\sup_{\theta} \left\| \frac{\sqrt{T}}{\lambda_{T,i}} \frac{\partial}{\partial \theta_i} \left[ G_{T}(\theta) - \frac{\lambda_T}{\sqrt{T} \rho_0(\theta)} \right] \right\| = o_p(1)$ by N3 and N7(a), (iii) $||\mu_{T,\theta}|| = O_p(1)$ by (15); (iv) $\left\| \frac{\Pi_{\rho_0} D_{T,\rho_0} \lambda_{T,j_i}}{\sqrt{T}} \right\| = O(1)$ by N3, (18) and (19);
and (v) $\frac{\lambda_{T,j_i}}{\lambda_{T,j_1}} = o(1)$ by N7(b) [also see Remark 9].

(d) A mean value expansion (with similar abuse of notation as above to denote the mean value $\tilde{\theta}_T$) gives $\sqrt{T} g_T(\theta_T) = \sqrt{T} g_T(\theta^0) + G_T(\tilde{\theta}_T) \sqrt{T} (\theta_T - \theta^0) = \sqrt{T} g_T(\theta^0) + G_T(\tilde{\theta}_T) \Pi_{p_0} D_{T,p_0} \mu_{T,\theta} = \sqrt{T} g_T(\theta^0) + G^* \mu_{T,\theta} + o_p(1)$ where the second equality uses (15) and the last one uses the result from Lemma 11(c).

(e) The result follows by Lemma 11 (a), (b), (d) since $\Pi_{p_0} D_{T,p_0} = o(\sqrt{T})$ by N3 and $\sqrt{T} g_T(\theta^0) = O_p(1)$.

(f) The result follows by Lemma 11 (c) and (e). \hfill \blacksquare

C.2 Clarification and details regarding footnote 9:

We briefly illustrate the said tradeoff by showing that if one strengthens the smoothness assumption by extending it to the second derivative, then this allows to weaken the rate assumption in N7(b). Since, at this point the definition of $\lambda_{T,j_i}$ is already stated, and since Lemma 11 works with $\lambda_{T,j_i}$ instead of $\lambda_{T,1}$, the discussion below uses $\lambda_{T,j_i}$. All we do hold if assumptions are maintained in terms of $\lambda_{T,1}$.

From Lemma 11, it is clear that the discussion here is pertinent mainly to part Lemma 11 (c) [see the proof of Lemma 11 (d)-(f)]. Indeed, the only part of (c) that needs attention is where we show that $[G_T(\theta_T) - G_T(\theta^0)] \Pi_{p_0} D_{T,p_0}$, i.e., the first term on the RHS of (30), is $o_p(1)$. So, let us focus on this.

**Remark 22:** If $g(Z_i; \theta)$ is linear in $\theta$, as in linear instrumental variables models, then this is trivially true since $[G_T(\theta) - G_T(\theta^0)] \equiv 0$ for all $\theta$. So, let us focus on a $g(Z_i; \theta)$ that is nonlinear in $\theta$.

Now, to accommodate for more smoothness we extend assumption N6 as N6’ to the second derivative, and replace N7 by N7’ as follows. (Assumptions N1-N5 and N8 remain the same.)

**Assumption N6’:** (a-one-time assumption for this clarification only)

(a) $\frac{\partial}{\partial \rho} \psi_T(\theta^0) = \sqrt{T} \left[ G_T(\theta^0) - \frac{\lambda_{T,j_1}}{\sqrt{T}} \rho_0(\theta^0) \right] = O_p(1)$. (This was the original N6.)

(b) For $i = 1, \ldots, d_\theta$: $\frac{\partial}{\partial \rho} \frac{\partial}{\partial \rho} \psi_T(\theta^0) = \sqrt{T} \left[ G_T(\theta^0) - \frac{\lambda_{T,j_1}}{\sqrt{T}} \rho_0(\theta^0) \right] = O_p(1)$. (This is the extension.)

**Assumption N7’:** (a-one-time assumption for this clarification only)

(a) $\rho(\theta)$ is thrice continuously differentiable in $\theta \in \mathcal{N}(\theta^0)$. $g(z; \theta)$ is thrice differentiable in $\theta \in \mathcal{N}(\theta^0)$ for each $z \in \mathbb{R}^{d_z}$ and $\sup_{\theta \in \mathcal{N}(\theta^0)} \left\| \frac{\partial}{\partial \theta_k} \left[ G_T(\theta) - \frac{\lambda_{T,j_1}}{\sqrt{T}} \rho_0(\theta) \right] \right\| = o_p(\lambda_{T,j_i}/\sqrt{T})$ for $i, k = 1, \ldots, d_\theta$.

(b) $\lambda_{T,j_i}$ from (19) satisfies $\lambda_{T,j_i}^3 / \lambda_{T,j_1} \rightarrow \infty$ as $T \rightarrow \infty$.

Comparing assumption N7 with N7’ reveals the tradeoff in terms of parts (a) and (b) of these assumptions. We note that similar tradeoffs can be generated by working with higher order order derivatives.

For clarity, specify further structure but without loss of generality. First, for $i = 1, \ldots, d_\theta$, define $\Pi_{p_0,i}$ and $D_{T,p_0,i}$ by the UBT-Construction like that in (18) and (19), but this time, by taking

$$W_T = \left[ \frac{\partial}{\partial \theta_i} \rho_{0,i}(\theta^0), \ldots, \frac{\partial}{\partial \theta_i} \rho_{0,i}(\theta^0) \right] = \left( I^* \frac{\partial}{\partial \theta_i} \rho_{0,i}(\theta^0) \right)'$$
(instead of \( W_T = [\rho_{\theta,1}(\theta^0), \ldots, \rho_{\theta,d}(\theta^0)] = (I^* \rho_\theta(\theta^0))' \) not depending on \( T \) in the UBT-Construction.
The corresponding quantities with full column-rank, and thus also the elements of \( D_{T,\rho_{\theta,i}} \) will change.
Indeed, no full-rank conditions are required, and instead, for the purpose of this proof, the only properties we will require are: For \( i = 1, \ldots, d_\theta \),

\[
\begin{align*}
\left( I^* \left\{ \frac{\partial}{\partial \theta_i} I^* \frac{\Lambda_T}{\sqrt{T}} I^* I^* \rho_\theta(\theta^0) \right\} \Pi_{\rho_{\theta,i}} D_{T,\rho_{\theta,i}} \right) &= O(1), \\
\Pi_{\rho_{\theta,i}} D_{T,\rho_{\theta,i}} &= o(\sqrt{T})
\end{align*}
\]

and these will not change since (32) holds by the construction of \( D_{T,\rho_{\theta,i}} \), while (33) follows from N3.

Start from (31). All we do below is to tease out further structure in the non-zero (i.e., the \( i \)-th) column of \( U_{T,i} \) (defined below (31)) so that assumption N7’ could be effectively used to show that \( \left[ \tilde{G}_T(\theta_T) - G_T(\theta^0) \right] \Pi_{\rho_\theta} D_{T,\rho_\theta}, \) i.e., the first term on the RHS of (30) is \( o_p(1) \). With this purpose in mind, for each \( i = 1, \ldots, d_\theta \), consider a further mean value expansion (with similar abuse of notation, and this time using \( \theta_T(\theta^k) \) to denote the mean value such that \( ||\theta_T(\theta^k) - \theta^0|| \leq ||\theta_T(\theta_i) - \theta^0|| \leq ||\theta_T - \theta^0|| \) for \( k = 1, \ldots, d_\theta \):

\[
\begin{align*}
\left\{ \frac{\partial}{\partial \theta_i} \tilde{G}_T(\theta_T(\theta_i)) \right\} (\theta_T - \theta^0) &= \left\{ \frac{\partial}{\partial \theta_i} \tilde{G}_T(\theta^0) \right\} (\theta_T - \theta^0) \\
&+ \left[ \left\{ \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_i} \tilde{G}_T(\theta_T(\theta_i)) \right\} (\theta_T(\theta_i) - \theta^0), \ldots, \left\{ \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_i} \tilde{G}_T(\theta_T(\theta_i^d)) \right\} (\theta_T(\theta_i) - \theta^0) \right] (\theta_T - \theta^0)
\end{align*}
\]

by similar (to above) interchange in the order of the derivatives. Since \( \mu_{T,\theta} = \sqrt{T} D_{T,\rho_\theta}^{-1} \Pi_{\rho_\theta}(\theta_T - \theta^0) \) by (15), it follows that:

\[
\begin{align*}
\left\{ \frac{\partial}{\partial \theta_i} \tilde{G}_T(\theta^0) \right\} (\theta_T - \theta^0) &= \left( I^* \left\{ \frac{\partial}{\partial \theta_i} I^* \frac{\Lambda_T}{\sqrt{T}} I^* I^* \rho_\theta(\theta^0) \right\} \Pi_{\rho_{\theta,i}} D_{T,\rho_{\theta,i}} \right) \mu_{T,\theta} \frac{1}{\sqrt{T}} \\
&= \frac{\partial}{\partial \theta_i} \sqrt{T} \left( \tilde{G}_T(\theta^0) - \frac{\Lambda_T}{\sqrt{T}} \rho_\theta(\theta^0) \right) \left( \Pi_{\rho_{\theta,i}} D_{T,\rho_{\theta,i}} \right) \mu_{T,\theta} \frac{1}{\sqrt{T}}
\end{align*}
\]

for \( i = 1, \ldots, d_\theta \). Define the \( d_\theta \times d_\theta \) matrices \( U_{a,T,i} \) and \( U_{b,T,i} \) such that all their columns are zeros, except for the \( i \)-th column, which for them is \( u_{a,T,i} \) and \( u_{b,T,i} \) respectively. Do this for all \( i = 1, \ldots, d_\theta \).

On the other hand, for the notation-abused quantity \( \theta_T(\theta_i) \), define \( \mu_{T,\theta(i)} := \sqrt{T} D_{T,\rho_\theta}^{-1} \Pi_{\rho_\theta}(\theta_T(\theta_i) - \theta^0) \) where \( ||\mu_{T,\theta(i)}|| \leq ||\mu_{T,\theta}|| \) by construction for \( i = 1, \ldots, d_\theta \) (recall that \( \mu_{T,\theta} = \sqrt{T} D_{T,\rho_\theta}^{-1} \Pi_{\rho_\theta}(\theta_T - \theta^0) \) by (15)). Now for each \( i = 1, \ldots, d_\theta \), define the \( d_\theta \times d_\theta \) matrices \( U_{c,T,i,k} \) for \( k = 1, \ldots, d_\theta \) such that all the
columns of $U_{c,T,i,k}$ are zeros, except for the $k$-th column which is:

\[
\begin{bmatrix}
\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_k} G_T(\theta_T(\theta^0_i))
\end{bmatrix} (\theta_T(\theta_i) - \theta^0_i) = \begin{bmatrix}
\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_k} G_T(\theta_T(\theta^0_i)) \sqrt{\lambda_T}
\end{bmatrix} \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \mu_{T,\theta(i)} \frac{\lambda_{T,j_1}}{\sqrt{T} \lambda_{T,j_1}}.
\]

Therefore, it follows that $U_{T,i}$ (defined below (31)) can be written as:

\[
U_{T,i} = U_{a,T,i} + U_{b,T,i} + \left( \sum_{k=1}^{d_0} U_{c,T,i,k} \right) (\theta_T - \theta^0) = U_{a,T,i} + U_{b,T,i} + \left( \sum_{k=1}^{d_0} U_{c,T,i,k} \right) \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \mu_{T,\theta} \frac{1}{\lambda_{T,j_1}}.
\]

And, therefore,

\[
\| [G_T(\theta_T) - G_T(\theta^0)] \Pi_{p_0} D_{T,p_0} \| \\
\leq \sum_{i=1}^{d_0} \| U_{T,i} \| \times \| \Pi_{p_0} D_{T,p_0} \| \\
\leq \sum_{i=1}^{d_0} \| U_{a,T,i} \| \times \| \Pi_{p_0} D_{T,p_0} \| + \sum_{i=1}^{d_0} \| U_{b,T,i} \| \times \| \Pi_{p_0} D_{T,p_0} \| \\
+ \sum_{i=1}^{d_0} \sum_{k=1}^{d_0} \| U_{c,T,i,k} \| \times \left\| \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \left\| \frac{\mu_{T,\theta}}{\lambda_{T,j_1}} \times \| \Pi_{p_0} D_{T,p_0} \|. 
\]

Since $\| u_{a,T,i} \| = O(1/\sqrt{T})$ by its definition and using (32), it follows that $\sum_{i=1}^{d_0} \| U_{a,T,i} \| \times \| \Pi_{p_0} D_{T,p_0} \| = o_p(1)$ by using (33). Since $\| u_{b,T,i} \| = O(1/\sqrt{T})$ by its definition and using N6’ and (33), it follows that $\sum_{i=1}^{d_0} \| U_{b,T,i} \| \times \| \Pi_{p_0} D_{T,p_0} \| = o_p(1)$ by using (33). Finally, note that $\sum_{i=1}^{d_0} \sum_{k=1}^{d_0} \| U_{c,T,i,k} \| \times \left\| \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \left\| \frac{\mu_{T,\theta}}{\lambda_{T,j_1}} \times \| \Pi_{p_0} D_{T,p_0} \| = o_p(1)$ since, collecting similar terms together,

\[
\left\| \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \right\| \times \left\| \frac{\mu_{T,\theta}}{\lambda_{T,j_1}} \times \| \Pi_{p_0} D_{T,p_0} \| \\
\leq \sup_{\theta} \left\| \frac{\lambda_T}{\sqrt{T}} \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_k} \rho_T(\theta) \times \sqrt{T} \lambda_{T,j_1} \right\| + \left\| \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_k} \left[ G_T(\theta_T) - G_T(\theta_T(\theta^0)) \right] \frac{\lambda_T}{\sqrt{T}} \mu_{T,\theta} \frac{1}{\lambda_{T,j_1}} \right\| \\
= O_p(1) \times O(1) \times O_p(1) \times O_p(1) \times o(1)
\]

term by term: (i) $\sup_{\theta} \left\| \frac{\lambda_T}{\sqrt{T}} \frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta_k} \rho_T(\theta) \times \sqrt{T} \lambda_{T,j_1} \right\| = O(1)$ by using N3 and N7’(a); (ii) $\left\| \frac{\Pi_{p_0} D_{T,p_0} \lambda_{T,j_1}}{\sqrt{T}} \right\|^3 = O(1)$ by using N3, (18) and (19); (iii) $\| \mu_{T,\theta(i)} \| = O_p(1)$ by using (15) and the definition of $\mu_{T,\theta(i)}$; (iv) $\| \mu_{T,\theta} \| = O_p(1)$ by using (15); and (v) $\frac{\lambda_T}{\lambda_{T,j_1}} = o(1)$ by using N7’(b). Thus $[G_T(\theta_T) - G_T(\theta^0)] \Pi_{p_0} D_{T,p_0} = o_p(1)$.

This completes the announced demonstration in the clarification for footnote 9. ■
C.3 Proof of the results from Section 4

Proof of Lemma 3: The proof is based on the original work of Antoine and Renault (2012), Andrews and Guggenberger (2014), Andrews and Cheng (2014) and Cheng (2015), with suitable adjustments that are required by our setup. Let \( \hat{G}_T := \hat{G}_T(\theta^0) \), \( \hat{V}_T := \hat{V}_T(\theta^0) \). By M1 and M2, \( \hat{V}_T \) is positive definite with probability approaching one as \( T \to \infty \). Thus, if defined, let \( \hat{V}_T^{-1/2} \) be such that \( \hat{V}_T^{-1/2} \hat{V}_T^{-1/2} = \hat{V}_T^{-1} \) and let \( \tilde{g}_T := \hat{V}_T^{-1/2} \tilde{g}_T(\theta^0) \) and \( H_T := \hat{V}_T^{-1/2} \hat{G}_T \). Then, for \( T \) sufficiently large, (4) gives:

\[
L M_T(\theta^0) = T \tilde{g}_T' P \left( H_T \{ H_T' H_T \}^{-1} \right) \tilde{g}_T \\
= T \tilde{g}_T' P \left( H_T B_T Y_T \left\{ \left( H_T B_T Y_T \right)' \left( H_T B_T Y_T \right) \right\}^{-1} Y_T B_T' R' \Pi_T^* D_T^* \right) \tilde{g}_T
\]

where \( Y_T := \text{diag}(1/\delta_{T,1}, \ldots, 1/\delta_{T,p}, \sqrt{T}1_{d_\theta-p}) \), a \( d_\theta \times d_\theta \) diagonal matrix, nonsingular for any given \( T \). 

(1c is the \( 1 \times c \) vector \((1, \ldots, 1)\), \( Y_T \) is diag\((1/\delta_{T,1}, \ldots, 1/\delta_{T,p})\) if \( d_\theta = p \) and is diag\((\sqrt{T}1_{d_\theta-p})\) if \( p = 0 \).

For a given \( T \), \( \Pi_T^* \) and \( D_T^* \) are \( d_R \times d_R \) nonsingular matrices defined as follows.

Step 1: Definition of \( \Pi_T^* \) and \( D_T^* \), and the asymptotic behavior of \( Y_T B_T' R' \Pi_T^* D_T^* \)

Under assumption M3(a) we can, without loss of generality, partition the set of elements \( \delta_{T,1}, \ldots, \delta_{T,p} \) into \( m-1 \) groups containing \( p_1, p_2, \ldots, p_{m-1} \) elements respectively as \( (\delta_{T,1}, \ldots, \delta_{T,p_1}) \), \( (\delta_{T,p_1+1}, \ldots, \delta_{T,p_2}) \), \ldots, \( (\delta_{T,p_{m-2}+1}, \ldots, \delta_{T,p_{m-1}}) \) where \( p_j \geq 0 \) and \( \bar{p}_j := \sum_{k=1}^{j} p_k \) for \( j = 1, \ldots, m-1 \) and \( m \in \{1, \ldots, p+1\} \) (let \( p_m := d_\theta - p \); and when \( p = 0 \) let \( m = 1 \); and also, by construction, \( \bar{p}_{m-1} = p \) and \( \bar{p}_m = d_\theta \), such that:

\[
\delta_{T,\bar{p}_j} \neq o(\delta_{T,\bar{p}_j-p_{j+1}}) \text{ for } j = 1, \ldots, m-1, \text{ and } \delta_{T,p_{j+1}} = o(\delta_{T,\bar{p}_j}) \text{ for } j = 1, \ldots, m-2. \tag{34}
\]

Taking \( W_T := RB_T = [W_{T,1}, \ldots, W_{T,m}] \) where \( W_{T,j} := RB_{(\bar{p}_j-p_j+1, \bar{p}_j)} \) for \( j = 1, \ldots, m \), define \( \Pi_T^* = [\Pi_{T,1}^*, \ldots, \Pi_{T,m}^*] \) as the \( \Pi \) matrix from the UBT-Construction in Appendix A.1.1. \( B_T \) is orthogonal for each \( T \) and also \( B_T \to B \), which is nonsingular by M3(c). Therefore, by Lemma 10, quantities such as \( q_T \) and \( c^*_T j_i, \) in the UBT-Construction have well defined limits as \( T \to \infty \). Denote these limits as \( q \) and \( c^*_j \), respectively, i.e., by dropping the subscript \( T \), and note that \( \sum_{i=1}^{q} c^*_j = d_R \).

Define \( D_T^* = \text{diag}(\delta_{T,\bar{p}_1 c^*_1, \ldots, \delta_{T,\bar{p}_m c^*_m}) \) where we use the notation \( \delta_{T,\bar{p}_{m-1}+1} = \cdots = \delta_{T,\bar{p}_m} = T^{-1/2} \) to allow for the possibility that \( j_q = m \). \( D_T^* \) is a \( d_R \times d_R \) nonsingular diagonal matrix for each \( T \).

Therefore, as \( T \to \infty \), it follows by M3(a) and (34), and then again using Lemma 10, that

\[
W^{*'} = \lim_{T \to \infty} Y_T B_T' R' \Pi_T^* D_T^*
\]

is a finite, non-random, \( d_\theta \times d_R \) matrix with full column-rank \( d_R \).\(^{16}\)

\(^{16}\)To see its full column-rank, use arguments similar to those below (19) along with M3(a) to obtain that for \( W^{*'} \), its
The rest of the proof is completely based on Andrews and Guggenberger (2014).

Step 2: Asymptotic behavior of $H_TB_T\Upsilon_T$

Under (9), $\|\Delta_T\| \leq c \times \hat{c}$ for some $c > 0$ by M2. Then, it follows that:

$$V_T^{-1/2} \tilde{G}_T B_T \Upsilon_T = V_T^{-1/2} \hat{G}_T \left[ B_{T,(1:p)} \Delta^{-1}_{T,(1:p)} ; \sqrt{T} B_{T,(p+1:d_R)} \right]$$
$$= V_T^{-1/2} G_T \left[ B_{T,(1:p)} \Delta^{-1}_{T,(1:p)} ; \sqrt{T} B_{T,(p+1:d_R)} \right]$$
$$+ V_T^{-1/2} \sqrt{T} \left( \tilde{G}_T - G_T \right) \left[ B_{T,(1:p)} \left( \sqrt{T} \Delta_{T,(1:p)} \right)^{-1} ; B_{T,(p+1:d_R)} \right].$$

By the orthogonality of $B_T$, it follows from the relation $V_T^{-1/2} G_T = C_{T,(1:d_R)} \Delta_T B_T'$ (obtained from (9)) and M3, that the first term on the right hand side of the above equation converges to $[C_{(1:p)}, C_{(p+1:d_R)} L].$

On the other hand, M1 and M2 give $\sqrt{T} \left( \tilde{G}_T - G_T \right)$ $\xrightarrow{d}$ devec$_d \left( \psi G - V_{Gg} V^{-1} \psi \right) = O_p(1)$ which, crucially, is independent of $\psi$. Also M3 implies that $[B_{T,(1:p)} \left( \sqrt{T} \Delta_{T,(1:p)} \right)^{-1} ; B_{T,(p+1:d_R)}] \rightarrow [0, B_{(p+1:d_R)}]$ as $T \rightarrow \infty$. Thus, by M1, for the second term on the right hand side of the above equation, we now have that $V_T^{-1/2} \sqrt{T} \left( \tilde{G}_T - G_T \right) \left[ B_{T,(1:p)} \left( \sqrt{T} \Delta_{T,(1:p)} \right)^{-1} ; B_{T,(p+1:d_R)} \right] \xrightarrow{d} \left[ 0, V^{-1/2} \text{devec}_d \left( \psi G - V_{Gg} V^{-1} \psi \right) B_{(p+1:d_R)} \right].$

Since M2 implies that $V_T^{-1/2} V_T^{1/2} \leadsto I_{d_g}$, it now follows, by combining the two terms, that

$$H_TB_T\Upsilon_T = \tilde{V}_T^{-1/2} \hat{G}_T B_T \Upsilon_T = \left( \tilde{V}_T^{-1/2} V_T^{1/2} \right) V_T^{-1/2} \tilde{G}_B T B_T \Upsilon_T \xrightarrow{d} G^*$$

(36)

where $G^* := [C_{(1:p)}, C_{(p+1:d_R)} L + V^{-1/2} \text{devec}_d \left( \psi G - V_{Gg} V^{-1} \psi \right) B_{(p+1:d_R)}],$ as defined in M3(d).

Step 3: Asymptotic behavior of $LM_T(\theta^0)$

Therefore, $P(H_TB_T\Upsilon_T \left( \left( H_TB_T \Upsilon_T \right)' \left( H_TB_T \Upsilon_T \right) \right)^{-1} Y_T B_T' R'T \Pi_T D_T) \xrightarrow{d} P(G^* (G^*G^*)^{-1} W^*)$, a finite matrix with full column-rank $d_R$ almost surely by (35), (36) and Lemma 10. Now, since M1 and M2 imply that $\sqrt{T} g_T \xrightarrow{d} V^{-1/2} \psi \sim N(0, I_{d_g})$, and since we have already noted the independence between $\psi$ and $G^*$, it follows that $LM_T(\theta^0) \xrightarrow{d} \chi^2_{d_R}$. □

Proof of Proposition 4: Let $\{\phi_{\gamma_S,T} : T \geq 1\}$ denote the sequence of indicator variables where $\phi_{\gamma_S,T} = 0$ if $CI_T(\gamma_S; \epsilon)$ contains $\gamma^0_S$, and $\phi_{\gamma_S,T} = 1$ otherwise. It is given that $CI_T(\gamma_S; \epsilon)$ has asymptotic coverage columns from $(d_R - \sum_{i=1}^q c^*_{j_i})$ to $(d_R - \sum_{i=1}^q c^*_{j_i} + c^*_{j_i})$ for $i = 1, \ldots, q$ are represented by the $d_g \times c^*_{j_i}$ matrix:

$$\begin{cases} 
\left[ (\delta_{p_1} \text{diag}(\delta_{p_1}^{-1} \cdots \delta_{p_1}^{-1}) B'_{(p_1)} R' \Pi_{j_1})', 0' \right]' & \text{if } j_1 = 1, \\
0', (\delta_{j_1} \text{diag}(\delta_{j_1}^{-1} \cdots \delta_{j_1}^{-1}) B'_{(j_1)} R' \Pi_{j_1})', 0' \right]' & \text{otherwise}
\end{cases}$$

(\text{where the 0 denotes sub-matrices of zeros with number of rows, which can be zero, such that the number of rows of the corresponding big matrix is } d_g). \text{ The non-zero blocks in such sets of columns (one block per set of columns) are: (i) at mutually non-overlapping positions (sets of rows); (ii) are finite by M1, M3(a); (iii) of full column-rank by Lemma 10, which tells that pre-multiplication by the nonsingular matrix } \delta_{j_1} \text{diag}(\delta_{j_1}^{-1} \cdots \delta_{j_1}^{-1}) \text{ does not change the rank of } B'_{(j_1)} R' \Pi_{j_1}. \text{ The latter has full column-rank } c^*_{j_i} \text{ for } i = 1, \ldots, q \text{ by (i) in the UBT-Construction. Therefore, full column-rank } d_R \text{ of } W^{*'} \text{ follows by noting that } \sum_{i=1}^q c^*_{j_i} = d_R. \text{ Note that the additional structure in M3(a), that Andrews and Guggenberger (2014) did not require, was imposed here precisely for this step [see sentence 4 in Remark 6].}
(1 − 𝜖) when 𝐻₀ is true. Hence, lim_{𝑇→∞} 𝑃_{𝑇}(φ_{S,T} = 0) ≥ (1 − 𝜖) where 𝑃_{𝑇}(⋅) denotes the probability of an event under 𝐹_{𝑇} constrained by assumptions O and M1-M3 and when 𝛽₀ = 𝑟₀, equivalently, when 𝑅θ₀ = 𝑟₀ [see Remark 23 and footnote 1 in that order]. Therefore, by construction:

\[
\lim_{𝑇→∞} 𝑃_{𝑇} \left( \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) ≤ LM_{T}(θ₀) \right) ≥ \lim_{𝑇→∞} 𝑃_{𝑇}(φ_{S,T} = 0) ≥ 1 − 𝜖, \tag{37}
\]

since for any 𝑇 ≥ 1, the event \{φ_{S,T} = 0\} ⊆ the event \{inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) ≤ LM_{T}(θ₀) \}. Recalling that the definition in (6) allows for the convention that inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) = ∞ if CI_{T}(γ_S;τ) is empty. Now, let \{φ_{β,T} : 𝑇 ≥ 1\} denote the sequence of indicator variables where φ_{β,T} = 1 if inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) > χ_{dr}^2 (1 − 𝛼), and φ_{β,T} = 0 otherwise. Therefore,

\[
𝑃_{𝑇}(φ_{β,T} = 0) = 𝑃_{𝑇} \left( \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) ≤ χ_{dr}^2 (1 − 𝛼) \right)
\[
≥ 𝑃_{𝑇} \left( \left\{ \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) ≤ LM_{T}(θ₀) \right\} \cap \left\{ LM_{T}(θ₀) ≤ χ_{dr}^2 (1 − 𝛼) \right\} \right)
\[
= 1 − 𝑃_{𝑇} \left( \left\{ \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) > LM_{T}(θ₀) \right\} \cup \left\{ LM_{T}(θ₀) > χ_{dr}^2 (1 − 𝛼) \right\} \right)
\[
≥ 1 − \left( 𝑃_{𝑇} \left( \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) > LM_{T}(θ₀) \right) + 𝑃_{𝑇} \left( LM_{T}(θ₀) > χ_{dr}^2 (1 − 𝛼) \right) \right),
\]

where the second line follows by the definition of φ_{β,T}, the third line by the construction of the two-step projection test in (6), the fourth line by De Morgan’s law, and the fifth line by Bonferroni’s inequality. Taking limits on both sides gives:

\[
\lim_{𝑇→∞} 𝑃_{𝑇}(φ_{β,T} = 0) ≥ 1 − \lim_{𝑇→∞} 𝑃_{𝑇} \left( \inf_{τ₀∈CI_{T}(γ_S;τ)} LM_{T} \left( A_{S}^{-1}(r_0' , γ_0') \right) > LM_{T}(θ₀) \right)
\[
− \lim_{𝑇→∞} 𝑃_{𝑇} \left( LM_{T}(θ₀) > χ_{dr}^2 (1 − 𝛼) \right)
\[
≥ 1 − (1 − 𝜖 + 𝛼),
\]

where the last line follows by (37) and Lemma 3. ■

Remark 23: Since the way it is stated in the statement of the proposition, the coverage probability of CI_{T}(γ_S;τ) is (1 − 𝜖), possibly under a larger class of distributions than 𝐹_{𝑇} constrained by the assumptions O and M1-M3. This is the reason behind the inequality lim_{𝑇→∞} 𝑃_{𝑇}(φ_{S,T} = 0) ≥ (1 − 𝜖). However, the confidence sets CI_{T}(γ_S;τ), e.g., CI_{T}^{SW}(γ_S;𝑟₀, 𝜖) defined in (10), that we actually mention [see Remark 2] are asymptotically similar and hence, for them, the above inequality will hold as an equality. ■

Proof of Lemma 5: (a) Utilizing the nonsingular matrices \(\Pi_{r₀}, D_{T,r₀}, \Pi_{R} \) and \(D_{T,R} \) in (18),(19), (22)
and (23) respectively, recall from (4) that $LM_T(\theta)$ can be written as:

$$LM_T(\theta) = T \times \left( \tilde{V}_T^{-1/2}(\theta) \tilde{g}_T(\theta) \right) P \left( \hat{H}_T(\hat{\theta}_T) \hat{H}_T(\hat{\theta}_T) \right)^{-1} D_{T,p_0} \Pi' R' \Pi_R D_{T,R} \left( \tilde{V}_T^{-1/2}(\theta) \tilde{g}_T(\theta) \right)$$

where $\hat{H}_T(\theta) := \tilde{V}_T^{-1/2}(\theta) \tilde{G}_T(\theta) \Pi_{p_0} D_{T,p_0}$. Now note that for $\theta_T$ defined in (15), we have:

(i) $\tilde{V}_T^{-1/2}(\theta_T) \xrightarrow{P} V^{-1/2}$ by N8 [see Lemma 11(a)] and $V^{-1/2}(\theta_T) \xrightarrow{} V^{-1/2}$ by definition [also see N8];

(ii) $\tilde{V}_T^{-1/2}(\theta_T) \sqrt{T} \tilde{g}_T(\theta_T) = V^{-1/2}[\sqrt{T} \tilde{g}_T(\theta^0)] + G^* \mu_{T,\theta} + o_p(1)$ by (i) and Lemma 11(c); and the RHS is $O_p(1)$ by N2, N8, and the definition of $\mu_{T,\theta}$ in (15);

(iii) $\hat{H}_T(\theta_T) \xrightarrow{P} V^{-1/2}G^*$ by (i) and Lemma 11(f).

(iv) $D_{T,p_0} \Pi' R' \Pi_R D_{T,R} \rightarrow R'$ by (24).

Note that the limiting quantities in (i), (iii) and (iv) are also finite and full column-rank by construction.

Therefore, from (i)-(iv) it follows that

$$LM_T(\theta_T) = \left( V^{-1/2}[\sqrt{T} \tilde{g}_T(\theta^0)] + G^* \mu_{T,\theta} \right)' P \left( V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^* \right) \left( V^{-1/2}[\sqrt{T} \tilde{g}_T(\theta^0)] + G^* \mu_{T,\theta} \right) + o_p(1).$$

Hence, the result of part (a) follows by recalling that $R^* \mu_{T,\theta} \xrightarrow{P} \mu_\beta$ [see below (15)] and noting that:

$$P \left( V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^* \right) V^{-1/2} G^* \mu_{T,\theta} = V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^* \left( R^* (G'^* V^{-1} G^*)^{-1} R'^* \right)^{-1} R^* (G'^* V^{-1} G^*)^{-1} G^* (G'^* V^{-1} G^*)^{-1} R'^* \left( R^* (G'^* V^{-1} G^*)^{-1} R'^* \right)^{-1} R^* \mu_{T,\theta} = V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^* \left( R^* (G'^* V^{-1} G^*)^{-1} R'^* \right)^{-1} \mu_\beta + o_p(1).$$

Importantly, the same probability limit holds for all $\theta_T$ satisfying (15), and hence the proof.

(b) Define $\nu := V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^* \left( R^* (G'^* V^{-1} G^*)^{-1} R'^* \right)^{-1} \mu_\beta$. From (a), now it follows by N2 and the full column-rank of $V^{-1/2} G^* (G'^* V^{-1} G^*)^{-1} R'^*$ [see N8, (21) and (24)] that $LM_T(\theta_T) \xrightarrow{d} \chi^2_{d_R}$ with non-centrality parameter given by $\nu' \nu = \mu_\beta^2 \left( R^* (G'^* V^{-1} G^*)^{-1} R'^* \right)^{-1} \mu_\beta$.

\textbf{Proof of Lemma 6:} Define the sequence $\{\gamma_T^\dagger : T \geq 1\}$ such that:

$$\gamma_T^\dagger := \arg \inf_{\gamma_0 \in \text{CI}_T(\gamma_S; \epsilon)} LM_T \left( A_S^{-1}(\gamma_0^0, \gamma_0^1) \right).$$

By condition (16) on $\text{CI}_T(\gamma_S; \epsilon)$, it then follows that $\gamma_T^\dagger$ gives $\theta_T^\dagger = R_S^1 r_0 + S_1^1 \gamma_T^\dagger$, for which $\sqrt{T} D_{T,p_0}^1 \Pi' R' (\theta_T^\dagger - \theta^0) = O_p(1)$, i.e., (15) holds since (14) also holds. Therefore, the final result follows by Lemma 5. \qed
Proof of Lemma 7: The lemma defines the supremum in case of an empty $C_{T}^{SW}(\gamma_{S};r_{0},\epsilon)$ in a way that allows us to ignore those cases in the sequel, with the caveat from the discussion above the lemma.

Now, the proof follows in three steps. In Step 1 we show that $C_{T}^{SW}(\gamma_{S};r_{0},\epsilon)$ shrinks to $\gamma_{S}^{0}$ in probability; more precisely, that the distance between $\gamma_{S}^{0}$ and $C_{T}^{SW}(\gamma_{S};r_{0},\epsilon)$ converges in probability to zero. Using this, in Step 2 we obtain that this rate cannot be slower than $\lambda_{T,1}$. Using this, in Step 3 we obtain the final result. Details, once stated, are not repeated in the subsequent steps.

Step 1: Cauchy-Schwartz inequality gives: $\|\rho(\theta)\| \leq \left\|\frac{\Lambda_{T}}{\sqrt{T}}\rho(\theta)\right\| \leq \left\|\frac{1}{\left\|\frac{\Lambda_{T}}{\sqrt{T}}\right\|}\|\rho(\theta)\|\right\| \geq \frac{\|\rho(\theta)\|}{\sqrt{T}\sum_{j=1}^{l} \frac{k_{j}}{\Lambda_{T,j}}}$

[see N3]. Recall that $d_{g} = \sum_{j=1}^{l} k_{j}$. Define the sequence $\Lambda_{T} = \min\{\lambda_{T,1}, \ldots, \lambda_{T,1}\}$ for $T \geq 1$. Hence, the above and (13) give that: $\left\|\frac{\Lambda_{T}}{\sqrt{T}} E_{T} [\tilde{g}_{T}(\theta)]\right\| = \left\|\frac{\Lambda_{T}}{\sqrt{T}} \frac{\rho(\theta)}{\sqrt{d_{g}}}\right\| \geq \frac{\|\rho(\theta)\|}{\sqrt{d_{g}}}$. Take any constant $\varepsilon > 0$. Consider the outside of the open ball around $\gamma_{S}^{0}$. Then, it follows by N1 and the above that:

$$\lim_{T} \inf_{\beta \in B, \gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \left\|\frac{\sqrt{T}}{\lambda_{T}} E_{T} [\tilde{g}_{T}(R_{S}^{1}\beta + S_{S}^{1}\gamma)]\right\| = \inf_{\beta \in B, \gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \frac{\|\rho(\theta)\|}{\sqrt{d_{g}}} > 0.$$

Taken together with N8 ($\inf_{\theta \in \Theta} \min[\text{eigen values}(V^{-1}(\theta))] > 0$), this gives:

$$\lim_{T} \inf_{\beta \in B, \gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \left\|V^{-1/2}(R_{S}^{1}\beta + S_{S}^{1}\gamma) \frac{\sqrt{T}}{\lambda_{T}} E_{T} [\tilde{g}_{T}(R_{S}^{1}\beta + S_{S}^{1}\gamma)]\right\| = 0.$$

Now, note that:

$$V^{-1/2}(\theta) \frac{\sqrt{T}}{\lambda_{T}} \tilde{g}_{T}(\theta) = V^{-1/2}(\theta) \frac{\sqrt{T}}{\lambda_{T}} (\tilde{g}_{T}(\theta) - E_{T}[\tilde{g}_{T}(\theta)]) + V^{-1/2}(\theta) \frac{\sqrt{T}}{\lambda_{T}} E_{T}[\tilde{g}_{T}(\theta)].$$

By N2, N3 and N8, the first term on the RHS is $o_{p}(1)$ uniformly in $\theta \in \Theta$. Therefore, by using the uniform consistency of $\hat{V}_{T}^{-1}(\theta)$ for $V^{-1}(\theta)$ from N8, and the definition of $Q_{T}(\theta)$ from (11), it follows that

$$\lim_{T} \inf_{\beta \in B, \gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \left(\frac{\Lambda_{T}^{2}}{\sqrt{T}} \times T \times Q_{T}(R_{S}^{1}\beta + S_{S}^{1}\gamma) > c \text{ for some } c > 0\right) = 1,$$

and hence,

$$\lim_{T} \inf_{\beta \in B, \gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \left(T \times Q_{T}(R_{S}^{1}\beta + S_{S}^{1}\gamma) > c \text{ for all } c < \infty\right) = 1,$$

and hence,

$$\lim_{T} \inf_{\gamma \in \Gamma_{S} : \|\gamma - \gamma_{S}^{0}\| \geq \varepsilon} \left(T \times Q_{T}(R_{S}^{1}r_{0} + S_{S}^{1}\gamma) > c \text{ for all } c < \infty\right) = 1. \quad (38)$$

The second line follows since $\lim_{T} \Lambda_{T} = \infty$ by N3. The third line follows since $r_{0} \in B$ for large $T$. 

41
Since \( \varpi > 0 \) is arbitrary, by the definition of \( CI_{T}^{SW}(\gamma_{S}; r_{0}, \epsilon) \) in (10) where the critical value is a fixed, finite positive number for a given \( \epsilon < 1 - \alpha \), it follows from (38) that:

\[
\sup_{\gamma_{0} \in CI_{T}^{SW}(\gamma_{S}; r_{0}, \epsilon)} \|\gamma_{0} - \gamma_{S}^{0}\| = o_{p}(1).
\]

Step 2: Take any constant \( \varpi > 0 \). Define \( \{\Gamma_{T}(\varpi) : T \geq 1\} \), shrinking at rate slower than \( \lambda_{T,1} \), as:

\[
\Gamma_{T}(\varpi) := \{ \gamma \in \Gamma_{S} : a_{T}\|\gamma - \gamma_{S}^{0}\| \leq \varpi \text{ and } \lambda_{T,1}\|\gamma - \gamma_{S}^{0}\| \geq b_{T} \text{ for some positive sequences } \{a_{T} : T \geq 1\} \text{ and } \{b_{T} : T \geq 1\} \text{ with } a_{T} \to \infty, b_{T} \to \infty \text{ as } T \to \infty \}.
\]

Consider a sequence \( \{\gamma_{T} : T \geq 1\} \) such that \( \gamma_{T} = \arg \inf_{\gamma \in \Gamma_{T}(\varpi)} T \times Q_{T}(R_{S}^{1}r_{0} + S_{S}^{1}\gamma) \) for each \( T \geq 1 \). (The sequence need not be unique.) Hence, \( \|\gamma_{T} - \gamma_{S}^{0}\| = o(1) \) (although \( \lim_{T} \lambda_{T,1}\|\gamma_{T} - \gamma_{S}^{0}\| = \infty \)). Also, \( \lim_{T} \lambda_{T,1}\|r_{0} - \beta_{T}^{0}\| = \infty \) by N3 and (14). Therefore, \( \theta_{T} := R_{S}^{1}r_{0} + S_{S}^{1}\gamma_{T} \in \mathcal{N}(\theta_{0}) \) (as in N4) for large \( T \). This gives, by a mean value expansion of \( \rho(\theta_{T}) \) around \( \rho(\theta_{0}) \) with mean value \( \bar{\theta}_{T} \) (element by element), that:

\[
\lambda_{T,1}A_{T}^{-1}\sqrt{T}E_T[\bar{g}_{T}(\theta_{T})] = 0 + \left[ \rho_{0}(\bar{\theta}_{T})R_{S}^{1}\lambda_{T,1}(r_{0} - \beta_{T}^{0}) + \rho_{0}(\bar{\theta}_{T})S_{S}^{1}\lambda_{T,1}(\gamma_{T} - \gamma_{S}^{0}) \right]
\]

by using N1. N4 and Lemma 10 imply that the terms inside the squared brackets on the RHS are full column-rank. Hence, the second term on the RHS is \( O(1) \) whereas the third term diverges to \( \pm \infty \). Since \( \lim \inf_{T} A_{T}/\lambda_{T,1} > 0 \) by N3, it follows that \( \lim_{T} \sqrt{T}\|E_T[\bar{g}_{T}(\theta_{T})]\| = \infty \). Hence, using the definitions of \( \gamma_{T} \) and \( \Gamma_{T}(\varpi) \) in conjunction with the result of Step 1, the same arguments as in Step 1 now give:

\[
\sup_{\gamma_{0} \in CI_{T}^{SW}(\gamma_{S}; r_{0}, \epsilon)} \lambda_{T,1}\|\gamma_{0} - \gamma_{S}^{0}\| = O_{p}(1).
\]

Step 3: Equipped with the result from Step 2, now we further refine this rate as follows. Define

\[
\Gamma_{T} := \{ \gamma \in \Gamma_{S} : \lambda_{T,1}\|\gamma - \gamma_{S}^{0}\| < b_{T} \text{ for any positive sequence } \{b_{T} : T \geq 1\} \text{ with } b_{T} \to \infty, \}
\]

\[
\|\sqrt{T}D_{T,\rho_{0}}^{-1}E_T[\bar{g}_{T}(r_{0} - \beta_{T}^{0}) + S_{S}^{1}(\gamma - \gamma_{S}^{0})]\| \geq a_{T} \text{ for some positive sequence } \{a_{T} : T \geq 1\} \text{ with } a_{T} \to \infty, \text{ and where } r_{0} \text{ is as defined in (14)} \}.
\]

Consider a sequence \( \{\gamma_{T} : T \geq 1\} \) such that \( \gamma_{T} = \arg \inf_{\gamma \in \Gamma_{T}} T \times Q_{T}(R_{S}^{1}r_{0} + S_{S}^{1}\gamma) \) for each \( T \geq 1 \). Either \( \gamma_{T} \notin \Gamma_{T} \), in which case Step 2 gives the result, or \( \gamma_{T} \in \Gamma_{T} \) and hence \( \lim_{T} \lambda_{T,1}\|\gamma_{T} - \gamma_{S}^{0}\| < \infty \).

\(^{17}\)This argument is used at the end of all the three steps in the proof of this lemma. A rigorous version of this argument is presented in the proof of Lemma 13.2 in Andrews (2017a). We focus on establishing an appropriate (for us) version of what Andrews (2017a) assumes as the global strong-identification condition for the nuisance parameter. See footnote 6.
Also, \( \lim_T \lambda_{T,1} \| r_0 - \beta^0 \| < \infty \) by N3 and (14). Therefore, for \( \theta_T := R_2^{-1} r_0 + S_1^{1/2} \gamma_T \), it follows that \( \lim \lambda_{T,1} \| \theta_T - \theta^0 \| < \infty \), giving \( \theta_T \in \mathcal{N}(\theta^0) \) (as in N4) for large \( T \). Furthermore, since \( \mu \beta \neq 0 \), \( D_{T,\rho_T}^{-1} \) and \( \Pi_{\rho_T}' \) are nonsingular, and \( R_2^1 \) and \( S_1^1 \) are full column-rank, it follows that \( \eta_T := \sqrt{T} D_{T,\rho_T}^{-1} \Pi_{\rho_T}' (\theta_T - \theta^0) \neq 0 \), and, by the definition of \( \Gamma_T \), that \( \lim_T \| \eta_T \| = \infty \).

To proceed, first we use (13) and N1, and by a mean value expansion of \( \rho(\theta_T) \) around \( \rho(\theta^0) \) with mean value \( \tilde{\theta}_T \) (element by element), we obtain that:

\[
\sqrt{T} \tilde{g}_T(\theta_T) = \sqrt{T} (\tilde{g}_T(\theta_T) - E_T[\tilde{g}_T(\theta_T)]) + \Lambda_T \left\{ \rho(\theta^0) + \rho_0(\theta^0)(\theta_T - \theta^0) + \left[ \rho_0(\theta_T) - \rho_0(\theta^0) \right](\theta_T - \theta^0) \right\} \\
= \psi_T(\theta_T) + \left( \frac{\Lambda_T}{\sqrt{T}} \rho_0(\theta^0) \Pi_{\rho_T} D_{T,\rho_T} \right) \eta_T + \phi_T(\theta_T)
\]

using \( \rho(\theta^0) = 0 \) [see N1], and where \( \psi_T(\theta_T) \), as defined in N2, is \( O_p(1) \) using N8, while

\[
\phi_T(\theta_T) := \Lambda_T \left[ \rho_0(\theta_T) - \rho_0(\theta^0) \right](\theta_T - \theta^0) = O_p \left( \frac{\lambda_{T,1}}{\lambda_{T,1}^2} \right) = o_p(1).
\]

The first equality for \( \phi_T(\theta) \) follows since \( \rho(\theta) \) is twice continuous differentiable in \( \mathcal{N}(\theta^0) \) [see N7(a)], and using N3 and the fact that \( \lim \lambda_{T,1} \| \theta_T - \theta^0 \| < \infty \) (noted above). The second one uses N7(b).\(^{18}\) Hence:

\[
\sqrt{T} \left\| \tilde{g}_T(\theta_T) - \psi_T(\theta_T) + G^{*} \frac{\eta_T}{\| \eta_T \|} \right\| = O_p \left( \frac{1}{\| \eta_T \|} \right).
\]

Hence, as in Step 1 and 2, by using N8, N1, the finiteness and full column-rank of \( G^{*} \) [see (21)], and that \( \eta_T \neq 0 \) while \( \lim_T \eta_T = \infty \), it follows that \( T \times D_{T,\rho_T}(\theta_T) \) diverges to \( \infty \) in probability. Therefore, using the definitions of \( \gamma_T, \theta_T \) and \( \Gamma_T \) in conjunction with the result of Step 2, the same arguments as in Step 1 now give the final result of the lemma:

\[
\sup_{\gamma_0 \in CH_T^{SW}(\gamma_0, r_0, \epsilon)} \sqrt{T} \left\| D_{T,\rho_T}^{-1} \Pi_{\rho_T}' \left( (R_2^1(r_0 - \beta^0) + S_1^1(\gamma_0 - \gamma_0^0)) \right) \right\| \leq O_p(1).
\]

**Proof of Proposition 8:** The proof is omitted since it is exactly same as that of Theorem 3.2(ii). ■

\(^{18}\)Recall from Remark 12 that under the Stock and Wright (2000) setup, \( \Pi_{\rho_T} = I_{d_0} \) and \( D_{T,\rho_T} = \sqrt{T} \Lambda_{T,1}^{-1} \). Hence, \( \eta_T = \Lambda_T (\theta_T - \theta^0) \). Hence, similar to above, an expansion of \( \sqrt{T} \tilde{g}_T(\theta_T) \) under Stock and Wright (2000)’s setup gives:

\[
\sqrt{T} \tilde{g}_T(\theta_T) = \psi_T(\theta_T) + \left( \tilde{\rho}_0(\theta^0) \frac{\Lambda_T}{\sqrt{T}} \Pi_{\rho_T} D_{T,\rho_T} \right) \eta_T + \phi_T(\theta_T) = \psi_T(\theta_T) + \tilde{\rho}_0(\theta^0) \eta_T + \phi_T(\theta_T)
\]

where, using the same arguments as in the main text but *without using* N7(b), it follows that:

\[
\phi_T(\theta_T) := |\tilde{\rho}_0(\theta_T) - \tilde{\rho}_0(\theta^0)| \tilde{\lambda}_T(\theta_T - \theta^0) = O_p \left( \frac{1}{\lambda_{T,1}^2} \right) \eta_T, \text{ i.e., } ||\phi_T(\theta_T)|| = O_p \left( \frac{||\eta_T||}{\lambda_{T,1}} \right).
\]

Therefore,\( \frac{\sqrt{T}}{||\eta_T||} \tilde{g}_T(\theta_T) = \frac{1}{||\eta_T||} \psi_T(\theta_T) + G^{*} \frac{\eta_T}{||\eta_T||} + O_p \left( \frac{1}{\lambda_{T,1}} \right). \) Since weak identification is not allowed, i.e., \( \tilde{\lambda}_{T,1} \rightarrow \infty \), and since \( G^{*} \) is full column-rank, this means that the rest of the steps in the main text of the proof can follow without change.
Supplemental Appendix D: Unrestricted-by-$H_0$ plug-in is not advisable

In reference to the key feature (F3) in Section 2, we conduct the same simulation study as in Section 4.2.3. However, instead of the standard plug-in tests, now we consider the unrestricted version of the plug-in tests that replace $\gamma_S$ in $LM_T(r_0, \gamma_S)$ by its unrestricted-by-$H_0$ CU-GMM estimator $\tilde{\gamma}_S := S\tilde{\theta}$, where $\tilde{\theta} := \arg\min_{\theta \in \Theta} Q_T(\theta)$ and $Q_T(\cdot)$ is as defined in (11). Here, $\tilde{\gamma}_S$ is the unrestricted LIML estimator.

Plots similar to Figure 1 are now reported in Figure 2. As evident from Figure 2, the intuition in (F3) is confirmed by the simulation results for the unrestricted plug-in tests. These tests are no longer invariant to $S$, and now behave in a way that resembles the behavior of the corresponding $S$-dependent infeasible tests when $H_0$ is false. This means that the unrestricted plug-in test has better power when the $\gamma_S$ with worse identification strength is used for the nuisance parameter. However, this comes at the cost of severe over-rejection of the true $H_0$ under specifications (ii) and (iii), where the standard plug-in test actually did not display over-rejection in Section 4.2.3 [see Figure 1]. To be clear, neither specification falls under the scope of our results, although Theorem 6 of Guggenberger and Smith (2005) indicates that the standard plug-in test (as in Section 4.2.3) would have had correct asymptotic size under specification (iii), i.e., when $\lambda_{T,1} = 1$ and $\lambda_{T,2} = \sqrt{T}$ and hence the nuisance parameters are strongly identified.

To avoid clutter, we do not report in Figure 2 the results for the unrestricted version of the two-step projection test. (They are available from the author.) But it should be mentioned that we find its empirical power to be extremely poor except under specification (vi) and, to a small extent, under specification (v). The problem of poor power can be roughly attributed to the all too frequent occurrences of very large and even unbounded\textsuperscript{19} first-step confidence set when the sample size $T$ is small. It happens because this confidence set does not take advantage of the false $H_0$ by imposing it. The problem of poor power does not go away even in simulations with very large $T$, under the three specifications (i)-(iii) that involve at least one weakly identified component (and hence, outside the scope of Section 4.2).

However, we do find that this problem of poor power essentially disappears if, e.g., $T = 2000$, for all the three specifications (iv)-(vi) that are under the scope of our results in Section 4.2. In other words, under these three specifications (iv)-(vi), as in Figure 1, the two-step test behaves similar to the less powerful infeasible and plug-in tests that use the better identified $\gamma_S^0$ as the nuisance parameter. Notwithstanding, the intuition from (F3) is not applicable to the two-step test since it may remain invariant to $S$;\textsuperscript{20} and, therefore, resembling the corresponding $S$-dependent infeasible test is very unlikely.

In summary, as noted in (F3), we do not recommend this strategy for the two-step projection test.

\textsuperscript{19}Since this a linear model, following convention, we did not impose compactness of $\Theta$ (or $\mathcal{B}$ and $\Gamma_S$) in our computation.

\textsuperscript{20}This happens if $\gamma_0 := \arg\inf_{\gamma \in \Gamma_T(\gamma_S, \epsilon)} LM_T(A_S^{-1}(r_0, \gamma)')$ and $r_0$ are such that $T \times Q_T(A_S^{-1}(r_0, \gamma_0)') \leq \chi^2_{d_0}(1 - \epsilon)$.
Figure 2: Empirical rejection probabilities of the infeasible test (infeas) in (8), and the unrestricted version of the plug-in test (LIML (unres)) based on the unrestricted CU-GMM (in this case, unrestricted LIML) estimator for $\gamma_S$. Two choices of $\gamma_S$, i.e., $\gamma_S = \theta_1$ and $\gamma_S = \theta_2$ are employed for the infeasible test and the unrestricted version of the plug-in test. For all tests, we take $\alpha = .045$. Results are based on 10,000 Monte Carlo trials. Horizontal axis: deviation of $H_0$ in (2) from truth [also see (14)]. Title: Identification strength that corresponds to specifications (i)-(vi) respectively.