How Robust are Robustness Checks?

Brenda Prallon*

November 11, 2026

Click here for the latest version

Abstract

Robustness checks are routine in empirical work, but there is no standard statistical procedure to formally measure what one can learn from them. I propose the *robustness tolerance*: the smallest distance b at which the hypothesis that every robustness-check estimand lies within b of the main-specification estimand is not rejected. It is a lower confidence bound for the maximal discrepancy, obtained by inverting a moment-inequality test, and it accounts for the sampling uncertainty and the correlation of the estimates. I show the robustness tolerance satisfies desirable theoretical properties. An application shows that, although assessing overall robustness is context-specific, the robustness tolerance guides those judgments and improves transparency.

1 Introduction

Robustness checks are commonly carried out in empirical research. In a survey of the articles published in *AEJ: Applied* between 2011 and 2026 that have a complete replication package, 177 of the 179 papers with classifiable specifications report at least one specification that is labelled as, or can be classified as, a robustness check, and the median paper reports about

*Cornell University. Department of Economics. *Email: bq45@cornell.edu*. AI was used for drafting, editing, proof-reading, and coding. I would like to thank Jörg Stoye, Francesca Molinari and José Luis Montiel Olea for their guidance. I would also like to thank Eric Auerbach, Levon Barseghyan, Edoardo Bollati, Federico Bugni, Ivan Canay, Jacob Dorn, Senan Hogan-Hennessy, Eleftheria Kelekidou, Yiqi Liu, Michael Lovenheim, Douglas Miller, Vítor Possebom, Seth Sanders, Yiwei Sun and seminar participants of the Northwestern University Econometrics Reading Group, Econometric Society World Congress 2025 and Sociedade Brasileira de Econometria 2025 for valuable feedback. All errors are my own.

five robustness specifications for every headline specification.[1] Examples include dropping potentially "problematic" observations, changing sample weights, or adding extra controls to address ex-post concerns about endogeneity, among others. Researchers' degrees of freedom in these choices may lead them to very different results even if using the same data (Breznau et al. (2022)). Robustness checks provide some insurance by slightly perturbing these choices and showing results do not change by "a lot".

The issue is "a lot" is usually not formally quantified; often researchers conclude results are "qualitatively similar". A naive but simple measure would take the maximum distance between robustness checks estimates and the main estimate; but it would not take into account any uncertainty in the estimators nor their correlation structure: it is a point estimate with no coverage statement, and a single imprecise check can make it arbitrarily large. This paper provides a measure that is intuitive, uses all the information in the covariance matrix, and is based on a formal test with statistical guarantees. The goal is to replace multiple appendix tables and informal statements about robustness by an actual statistic, easily interpretable in each context.

The proposed measure is the *robustness tolerance*. For the main specification estimand θ_0 and robustness checks estimands θ_j, $j = 1, \ldots, m$, all scalars, consider the test

$$H_0 : \max_{j \in \{1,\ldots,m\}} |\theta_0 - \theta_j| \leq b, \tag{1}$$

where $b \geq 0$ is also a scalar. The robustness tolerance is defined in words as the smallest threshold b for which H_0, the null of "all robustness checks falling within this distance

[1]The sample is the 188 *AEJ: Applied* articles with full data availability in the replication-package database of Paul Goldsmith-Pinkham (https://paulgp.com/replication-package-db); 181 could be processed and 179 have at least one gradeable specification. Specifications were extracted from each article and its online appendix by a large language model, which recorded every printed regression cell it judged in scope with its coefficient, standard error, sample size, and descriptions of outcome, treatment, sample, controls and fixed effects, and labelled each cell as a headline specification or a robustness check under a written rule: a headline specification is one the paper anchors its claim to (quoted in the abstract, introduced as the baseline or preferred specification, or used as the reference point of the robustness tables), and a robustness check is a variation that preserves the treatment and the outcome while changing controls, sample, estimator, fixed effects, or identification strategy; subgroup splits count as robustness checks only when the paper uses them to show that the effect survives, and the coefficient on a control never counts. The numerical content of every entry was re-verified against the published tables; the headline-versus-robustness labels were not independently reviewed, by a second pass or by hand. Counts are of printed specifications, not of distinct robustness strategies, so a single idea applied across many columns counts many times; the median paper reports 8 headline specifications and 4.9 robustness specifications per headline specification (means 16.2 and 9.2). At the level of individual claims, 135 of 906 claims with a headline estimate (15%) have no robustness specification attached to them, a figure that includes two papers in which every printed cell was labelled a headline specification.

from the main specification," cannot be rejected. It is formally defined in Section 2.

The main contribution of this paper is the framework: a preferred specification, robustness as the null, and a measure defined by the smallest threshold the data do not reject, which is a lower confidence bound for the maximal discrepancy and comes with a relative version that measures discrepancies in units of the main effect. Within it, I show that taking robustness as the null is what protects the author: a check whose estimand lies inside the interval spanned by the other checks cannot push the measure above the true maximal discrepancy, except with probability α, however imprecise the check; no measure that takes fragility as the null, that is, no valid upper bound on the maximal discrepancy, has this property, and an imprecise check drives such a bound to infinity. A check whose estimand lies outside the interval moves the tolerance once its departure is large relative to the standard error of its own discrepancy, which is what a referee asking for it is entitled to. For linear regressions whose specifications may differ in controls, samples and variable definitions, I provide a plug-in covariance estimator for the stacked coefficients, built from their influence functions, under which the procedure is uniformly valid.

The test which defines the robustness tolerance depends only on the estimated discrepancies and their covariance, and simulations show how the tolerance responds to these parameters. In an application, a check the authors single out as an exception moves the point estimate by 79% of the main effect; the tolerance shows the estimand moves by at least 9%. This illustrates how the robustness tolerance may signal robustness even when individual checks may not. The chosen set of robustness checks is a key component of the analysis, and is assumed to be pre-specified based on domain-specific knowledge.

This framing of robustness is not innocuous. It requires a preferred specification θ_0. This reflects empirical practice in part: researchers usually argue, based on domain-specific knowledge, in favor of a certain empirical strategy, which is identifying the true parameter of interest. Robustness checks are then provided as additional evidence to strengthen weaker parts of the argument, but are typically not carefully justified, or may use lower-quality data. Furthermore, having a preferred specification is coherent with studying point-identified parameters and reporting point-identified confidence intervals. In this setting, robustness is conceptually framed as the null, and the robustness tolerance quantifies evidence *against* it, with a one-sided test. However, it is sometimes not explicitly clear which regression to treat as the main one. Researchers (and referees) should decide whether they have sufficient arguments for a preferred specification, or whether they want to consider some form of model misspecification/uncertainty, which I discuss below.

The prevalence of robustness checks can be linked to concerns about model misspecification. For recent reviews, see Armstrong (2025), Andrews, Chen, and Tecchio (2026) and references therein. This paper does not consider the main specification θ_0 to be misspecified, or partially identified; it treats the robustness checks as estimating parameters θ_j that may differ from θ_0, which is what current empirical practice implicitly assumes, and asks how far from θ_0 the data show them to be. If the researcher is worried about misspecification (or uncertainty) in θ_0 and wishes to use robustness checks to expand the model in the sense of Armstrong (2025), θ_0 should be considered partially identified and methods for union bounds such as in Bei (2024) can be used to compute appropriate confidence intervals (intuitively, the null in Equation 1 would take the pairwise difference between all specifications and would have a flipped sign). Alternatively, the more recent Jaeger (2026) also provides measurements of robustness based on the range of the parameter values included, without a preferred specification; however, the object of interest is the range itself, and no explicit confidence interval is provided for a possibly partially identified parameter.

A robustness test has been proposed by Lu and White (2014), with a similar null from Equation 1 except taking $b = 0$. This is a strong requirement: all robustness checks estimands must be identifying the exact same as the main specification estimand[2]. The test is equivalent to a Hausman test, and the authors propose using a more efficient estimator upon failure to reject. However, much more realistically, robustness checks may not identify the true parameter of interest, but still be best linear predictors that are close to the truth: in other words, it is often the case where $b > 0$. To perform such a test with $b > 0$ (and hence obtain the robustness tolerance), I use the procedures described in Cox and Shi (2023). While other tests from the moment inequalities literature could potentially have been used[3], Cox and Shi (2023) is fast to compute[4], involves no tuning parameters, and includes, as a special case, the test of Lu and White (2014).

Robustness checks are, often, a handcraft art; not only do they depend on the question at hand, but often also on specific datasets (for example, in Section 4, I discuss an application that runs analyses across different weighting schemes due to a lack of documentation of survey weights)[5]. The innate particularities of robustness checks indicate their limitations

[2]For important cases where this assumption should hold, see Lu and White (2014).

[3]See Molinari (2020), Tamer (2010), Canay and Shaikh (2017) and Canay, Illanes, and Velez (2026) for extensive reviews.

[4]Similarly to building confidence intervals by test inversion, to compute the robustness tolerance, the test must be repeatedly evaluated with different values of b.

[5]Certain type of robustness checks can be systematic: see Giordano, Meager, and Broderick (2026) for

in measuring bias beyond specifications that can be observed. In fact, there is a large body of literature that has criticized the practice of informal robustness analyses as proxies for assessing omitted variable bias (Imbens (2003), Altonji, Elder, and Taber (2005), Oster (2019), Cinelli and Hazlett (2020), Chernozhukov, Cinelli, Newey, Sharma, and Syrgkanis (2026), Diegert, Masten, and Poirier (2026a), for a non-exhaustive list of examples). Although bias from unobservables is a very relevant topic to applied work, this paper focuses instead on the role of robustness checks in measuring bias from "observables": minor choices, from data cleaning to economic/statistical modeling, that are deemed not to be major drivers of the main result. Nonetheless, Section 5.2 shows that the robustness tolerance can be used to calibrate the sensitivity parameter of Diegert et al. (2026a): the value at which an omitted variable would move the main estimand by as much as the robustness checks are shown to move it. This value also expresses the scope of a robustness exercise, how far the checks move the estimand, in the units of a sensitivity parameter: the strength of an omitted variable relative to the included controls.

The remainder of the paper is divided as follows: Section 2 introduces the framework and formal definition for robustness tolerance, together with a relative version that normalizes the unit of measurement of robustness, the implementation, and the theoretical properties of the measure; Section 3 shows what the test responds to and studies the tolerance by simulation as the configuration, correlation and number of checks vary, including the case of a zero tolerance and the effect of adding a check; Section 4 shows one application and discusses the interpretation of the measure; Section 5 presents two extensions, a standardized robustness tolerance and the connection between the robustness tolerance and other formal sensitivity measures; and Section 6 concludes.

2 Set-up

2.1 The robustness tolerance

Consider a finite collection of $m+1$ specifications indexed by $j = 0, \ldots, m$. Let θ_j denote the population estimand associated with specification j, and let $\theta = (\theta_0, \theta_1, \ldots, \theta_m)'$. Let $j = 0$ denote the main specification, which is the specification for which the researcher argues about their identification strategy and is willing to attribute causal meaning to

the case of checking sensitivity to few observations.

θ_0[6]. This framework for robustness, with a preferred specification, is coherent with studies focusing on point-wise parameters and confidence intervals. The remaining specifications $j = 1, \ldots, m$ represent robustness checks. The robustness checks need not identify the same parameter as the preferred specification, causal or otherwise; the framework measures how far their estimands are from θ_0 without deciding whether a discrepancy is bias. However, we believe that robustness checks should be informative, such that θ_j is close to θ_0, $j \neq 0$. The question is: *how* close?

A natural estimand of interest is the maximum distance between the main specification estimand θ_0 and robustness-check estimands, θ_j:

$$b(\theta) := \max_{j \in \{1,\ldots,m\}} |\theta_j - \theta_0|. \tag{2}$$

While many estimators could be proposed for $b(\theta)$, there are a few desirable properties specific to this framework of robustness: with a preferred specification, the researcher commits to justified choices, while robustness checks are not usually as carefully justified – which is why, typically, it is expected that $\theta_j \neq \theta_0$. In this setting, an informative measure of robustness should satisfy consistency; however, it should also protect the researcher against an imprecisely estimated check whose population discrepancy lies within the range already spanned by the other checks. This protects the researcher when including imprecisely estimated checks asked by referees. The properties are formally stated in Section 2.2 (Properties 1 and 2), and the robustness tolerance is shown to satisfy both. The latter aligns with the framing *robustness as the null*[7].

With those properties in mind, for a threshold $b \geq 0$, consider the following level α test $\phi_n(b, \alpha)$ of the one-sided null hypothesis:

$$H_0(b) : b(\theta) \leq b. \tag{3}$$

Equivalently,

$$H_0(b) : A\theta \leq b\mathbf{1}_{2m}, \tag{4}$$

[6]The framework does not require θ_0 to have causal meaning, but causality is a common justification.

[7]Proposition 2 in Section 2.2 shows that no robustness measure with a flipped null can satisfy Property 2.

where

$$A = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Thus, each robustness check contributes the pair of inequalities $\theta_j - \theta_0 \leq b,\ \theta_0 - \theta_j \leq b$.

We are now ready for the main definition of this paper:

Definition 1 (Robustness tolerance)**.** For significance level α, define

$$b_{RT}(\alpha) := \inf \{b \geq 0 : \phi_n(b, \alpha) = 0\}. \tag{5}$$

In words, b_{RT} is the smallest threshold b for which H_0 is not rejected at level α. If $\phi_n(\cdot, \alpha)$ is monotonic in b, then $b_{RT}(\alpha)$ coincides with the lower limit of a one-sided $(1-\alpha)100\%$ confidence interval for $b(\theta)$; otherwise, $b_{RT}(\alpha)$ is a lower coverage bound with coverage at least $1 - \alpha$.

The definition presumes that the θ_j are measured in common units; a check that rescales the outcome or the treatment must be expressed in the units of the main specification before it enters.

I discuss more about the choice of test $\phi_n(b, \alpha)$ in Section 2.2. If b is very large, such that none of the inequalities are violated, any reasonable test would not reject H_0, but b might not be very informative. If $b = 0$, this means we are assuming every specification is identifying the exact same estimand, which may be too strong of a requirement. On the other hand, b_{RT} is defined as an adaptive quantity, similar in spirit to a p-value. It does not ask the researcher to simply take the largest distance between robustness checks and the main specification, nor does it require all robustness checks to be indistinguishable from the main estimate. It takes into account whether certain robustness checks have a larger variance, and how they correlate with the main estimate (and each other).

The robustness tolerance is a measure of level; the interpretation of the magnitude of b_{RT} depends on the units in the study considered. I offer two alternative definitions that are useful in aiding interpretation. The first one measures discrepancies relative to

the magnitude of the main estimand, that is, the unit of robustness becomes the main specification coefficient. It is quite general and presented here; the second one is specialized to linear regressions and presented in Section 5.1.

Definition 2 (Relative robustness tolerance)**.** For $c \geq 0$, consider the null hypothesis

$$H_0^{\text{rel}}(c) : \max_{j \in \{1,\ldots,m\}} |\theta_j - \theta_0| \leq c|\theta_0|. \tag{6}$$

Because of the absolute value of θ_0, this null is not itself a polyhedron. It can, however, be written as the union of two polyhedra,

$$H_0^{\text{rel}}(c) = P_+(c) \cup P_-(c),$$

where

$$P_+(c) = \{\theta : A^+(c)\theta \leq \mathbf{0}_{2m}\}, \qquad P_-(c) = \{\theta : A^-(c)\theta \leq \mathbf{0}_{2m}\},$$

with

$$A^+(c) = \begin{pmatrix} -(1+c) & 1 & 0 & \cdots & 0 \\ 1-c & -1 & 0 & \cdots & 0 \\ -(1+c) & 0 & 1 & \cdots & 0 \\ 1-c & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(1+c) & 0 & 0 & \cdots & 1 \\ 1-c & 0 & 0 & \cdots & -1 \end{pmatrix} \text{ and } A^-(c) = \begin{pmatrix} -(1-c) & 1 & 0 & \cdots & 0 \\ 1+c & -1 & 0 & \cdots & 0 \\ -(1-c) & 0 & 1 & \cdots & 0 \\ 1+c & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -(1-c) & 0 & 0 & \cdots & 1 \\ 1+c & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Let $\phi_n^+(c,\alpha)$ and $\phi_n^-(c,\alpha)$ denote level-α tests of $P_+(c)$ and $P_-(c)$, respectively, with one denoting rejection. Define the test of $H_0^{\text{rel}}(c)$ as the intersection–union test

$$\phi_n^{\text{rel}}(c,\alpha) := \phi_n^+(c,\alpha)\phi_n^-(c,\alpha).$$

Thus, $H_0^{\text{rel}}(c)$ is rejected only if both $P_+(c)$ and $P_-(c)$ are rejected.

For significance level α, the relative robustness tolerance is

$$c_{RT}(\alpha) := \inf\left\{c \geq 0 : \phi_n^{\text{rel}}(c,\alpha) = 0\right\}. \tag{7}$$

Its population counterpart is

$$c(\theta) = \inf\{c \geq 0 : \theta \in P_+(c) \cup P_-(c)\} = \begin{cases} \dfrac{\max_{j=1,\ldots,m} |\theta_j - \theta_0|}{|\theta_0|}, & \theta_0 \neq 0, \\ 0, & \theta = 0, \\ +\infty, & \theta_0 = 0,\ \theta \neq 0. \end{cases}$$

The population counterpart makes evident that the relative robustness tolerance is uninformative when θ_0 is close to 0: its target is $+\infty$ at $\theta_0 = 0$, and Corollary 1 below shows that $c_{RT}(\alpha)$ remains valid but can be conservative there. The robustness tolerance $b_{RT}(\alpha)$ is unaffected, so when the main estimand is close to zero it is the measure to use, and it is in any case the measure to use when the units of θ_0 are themselves interpretable. For the relative tolerance the infimum in (7) need not be attained: at $c = 0$ the two branches coincide and the active rank is at most m, while for small $c > 0$ the wrong-sign branch can place the restricted estimator at the origin, with rank $m + 1$ and a larger critical value, so that $c = 0$ is rejected and every $c > 0$ near it is not. A zero relative tolerance therefore means that thresholds arbitrarily close to zero are not rejected, not that equality is. When θ_0 is well away from zero, the relative tolerance has a direct reading, which the following definition makes precise.

Definition 3 (Robustness with respect to magnitude)**.** The robustness checks are *consistent with robustness with respect to magnitude* if $c_{RT}(\alpha) \leq 1$, and the data *reject* robustness with respect to magnitude if $c_{RT}(\alpha) > 1$.

The rule in Definition 3 rejects at c exactly when every threshold at or below c is rejected, so it is a valid level-α test of $H_0^{\text{rel}}(c)$ that rejects only when the RCC test of $H_0^{\text{rel}}(c)$ does. The two can differ in the other direction: because the non-rejected set need not be an interval (Section 2.2), the RCC test of $H_0^{\text{rel}}(1)$ can reject while $c_{RT}(\alpha) \leq 1$. In that case the data do not reject some parameter value satisfying the magnitude hypothesis, since $H_0^{\text{rel}}(c) \subseteq H_0^{\text{rel}}(1)$ for $c \leq 1$, which is what "consistent with" means here.

The threshold $c = 1$ is the natural benchmark because $H_0^{\text{rel}}(1)$ states that every θ_j lies in the interval between 0 and $2\theta_0$: the hypothesis that no robustness check moves the estimand by more than the main effect itself. Because $c_{RT}(\alpha)$ is a lower confidence bound for $c(\theta)$, $c_{RT}(\alpha) > 1$ is a level-α rejection of this hypothesis, that is, evidence that at least one robustness check identifies an estimand outside that interval: either

one with the opposite sign to θ_0, or one with the same sign but more than twice its magnitude. The test does not distinguish between the two cases, and $c_{RT}(\alpha) > 1$ on its own is therefore not evidence of a sign reversal. A necessary condition for rejection holds in every sample: if $\max_j |\hat{\theta}_j - \hat{\theta}_0| \leq |\hat{\theta}_0|$, then $\hat{\theta}$ itself satisfies the inequalities of $P_+(1)$ or $P_-(1)$, the corresponding branch statistic is zero, and $c_{RT}(\alpha) \leq 1$; more generally, $c_{RT}(\alpha) \leq \max_j |\hat{\theta}_j - \hat{\theta}_0|/|\hat{\theta}_0|$ whenever $\hat{\theta}_0 \neq 0$. The same reading applies at any other threshold c, with $c|\theta_0|$ in place of $|\theta_0|$, when the context suggests a stricter or looser benchmark; judging overall robustness depends on how rich the set of robustness checks is and on the domain-specific knowledge of the researcher.

Robustness with respect to magnitude is different from robustness with respect to sign, and it is worth being explicit that the relative tolerance is not designed to assess it: $H_0^{\text{rel}}(1)$ bundles sign and scale, and its rejection can come entirely from checks of the same sign as θ_0. Tests of sign congruence, such as the one recommended by Miller, Molinari, and Stoye (2025) for two parameters, reject only if the two estimates have opposite signs, and in the uncorrelated case only if each is also individually significant. In the survey of *AEJ: Applied* articles described in the Introduction, a robustness check with the opposite sign of its headline estimate occurs in 13% of comparable claim-estimates (159 of 1,207, in 55 of 158 papers), and in only 0.5% of them (6 claim-estimates, in 4 papers) are both the headline estimate and the opposite-signed check reported as significant[8]. The necessary condition for rejecting robustness with respect to magnitude, $\max_j |\hat{\theta}_j - \hat{\theta}_0| > |\hat{\theta}_0|$, on the other hand, is met more often: in 24% of the same claim-estimates (289 of 1,207, in 93 of 158 papers), and most of the difference between the two figures comes from checks that keep the sign of the headline estimate but more than double its magnitude.[9] Whether the magnitude hypothesis is in fact rejected in those cases depends on the joint precision of the estimates, which the printed tables do not reveal and which $c_{RT}(\alpha)$ is designed to account for.

[8] A claim-estimate is a headline coefficient together with the robustness checks that share its outcome, treatment variable, and outcome transformation; groups whose headline specifications disagree in sign are excluded. Significance is as reported in the published tables.

[9] Under the stricter grouping that keeps only claim-estimates with a single headline specification, the shares are 12% and 19% (982 groups, 135 papers). In 41 claim-estimates both the headline estimate and a check deviating by more than its magnitude are reported as significant, against 6 for the opposite-sign criterion.

2.2 Implementation and Theoretical Properties

I implement the tests $\phi_n(b,\alpha)$, $\phi_n^+(c,\alpha)$ and $\phi_n^-(c,\alpha)$ using the refined conditional chi-squared (RCC) test of Cox and Shi (2023)[10]. The test is set up for moment inequalities. Their proposed test builds on Al Mohamad, Van Zwet, Cator, and Goeman (2020), and uses a quasi-likelihood statistic, which it compares to a chi-square distribution with degrees of freedom corresponding to the rank of the inequalities that are active at the restricted estimator; crucially for this application, and differently from Al Mohamad et al. (2020), the test allows for an intercept (b can be greater than 0). Notably, Cox and Shi (2023)'s test has uniformly asymptotically exact size and no tuning parameters, making it extremely computationally attractive. However, the test does not apply verbatim, as it is written for moment functions: my framework handles estimators that are consistent, but typically not unbiased, by requiring uniform convergence of the remainder at high level (Assumption 2 in Appendix A.1).

Let $\hat{\theta} = (\hat{\theta}_0, \ldots, \hat{\theta}_m)'$ denote an estimator of θ, and let $\hat{\Sigma}$ denote an estimator of the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$. I provide a suitable one for linear regressions in the context of robustness checks, where each $\hat{\theta}_j$ varies both in covariates and sub-samples used, in Section 2.3.

For a threshold $b \geq 0$, define the restricted estimator

$$\hat{\mu}(b) \in \arg\min_{\mu: A\mu \leq b\mathbf{1}_{2m}} n(\hat{\theta} - \mu)'\hat{\Sigma}^{-1}(\hat{\theta} - \mu), \tag{8}$$

and the corresponding quasi-likelihood ratio statistic

$$T_n(b) = n(\hat{\theta} - \hat{\mu}(b))'\hat{\Sigma}^{-1}(\hat{\theta} - \hat{\mu}(b)). \tag{9}$$

Let a_s' denote row s of A, $s = 1, \ldots, 2m$, and define the set of active inequalities at the restricted estimator as

$$\widehat{\mathcal{J}}(b) = \{s \in \{1, \ldots, 2m\} : a_s'\hat{\mu}(b) = b\}. \tag{10}$$

For any $\mathcal{J} \subseteq \{1, \ldots, 2m\}$, let $A_{\mathcal{J}}$ denote the submatrix of A containing the rows indexed

[10]While they also provide a version for subvector inference in conditional moment inequality models (and an example of how to use it in linear regressions), it is not directly applicable here since each specification has a different informational set to condition on.

by $\mathcal{J}$. Define

$$\hat{r}(b) = \mathrm{rk}\left(A_{\widehat{\mathcal{J}}(b)}\right). \tag{11}$$

The conditional chi-squared (CC) test compares $T_n(b)$ to the $(1-\alpha)$ quantile of a chi-squared distribution with $\hat{r}(b)$ degrees of freedom. The RCC test only differs from the CC test when $\hat{r}(b) = 1$. While I implement the RCC test in examples, I will skip the theoretical details, which can be found in Cox and Shi (2023)'s Supplementary Appendix A.1.

Computing $b_{RT}(\alpha)$ means inverting the test over b, and the inversion requires care because the set of thresholds at which the test does not reject need not be an interval. The statistic $T_n(b)$ is continuous and nonincreasing in b, since the feasible set in (8) grows with b, but the critical value is not: when a constraint stops binding as the interval expands, $\hat{r}(b)$ falls by one and the critical value drops, so a threshold can be rejected after a smaller one has been accepted.[11] The implementation therefore does not bisect on the rejection indicator. For each rank $r = m, \ldots, 1$ in turn it solves $T_n(b) = \chi^2_{r,1-\alpha}$ by bisection, which is well defined because T_n is monotone; between successive roots it checks on a fine grid whether the active rank jumps, which is the only way an earlier interval of non-rejection can open, and locates the jump if it does; and at rank one it iterates the refinement, whose critical value depends on b, to a fixed point and polishes by bisection on the rejection indicator. The value returned is the infimum of the thresholds at which the test does not reject.

Two consequences follow. First, $b_{RT}(\alpha) \leq b$ does not imply that the test accepts $H_0(b)$; what the tolerance reports is that every threshold below it is rejected. Second, the coverage of $b_{RT}(\alpha)$ as a lower bound for $b(\theta)$ follows from the inclusion $\{b_{RT}(\alpha) > b(\theta)\} \subseteq \{\phi_n(b(\theta), \alpha) = 1\}$, so it is at least $1-\alpha$ asymptotically, and exact only when the non-rejected set is an interval. The same procedure, applied to each branch separately with c in place of b, computes $c_{RT}(\alpha)$; there the active rank can reach $m+1$ for $c > 0$, since both inequalities of a pair bind only when the restricted estimator sets $\mu_0 = 0$, and the search starts from that rank.

The non-monotonicity is a property of the RCC test, not of the framework: any family of level-α tests of $H_0(b)$ can be inverted in the same way, and families with different power properties give different tolerances for the same target. Natural alternatives are the

[11] With $m = 2$, $\hat{\Delta} = (3.2, 1)'$, $\hat{\Sigma}_\Delta = I_2$, $n = 1$ and $\alpha = 0.05$, both constraints bind for $b < 1$ and the test does not reject on $[0.764, 1]$; for $b > 1$ only the first binds, the critical value falls from $\chi^2_{2,0.95}$ to the rank-one value, and the test rejects again on $(1, 1.344)$. Here $b_{RT}(\alpha) = 0.764$.

classical least-favorable test (Kudô, 1963; Wolak, 1987), which has exact size but responds to slack inequalities, and the conditional and hybrid tests of Andrews, Roth, and Pakes (2023), which condition on the inequality with the largest violation rather than on the set of active ones. Which family detects the alternatives that matter in a robustness exercise is a question about power that I leave for future work.

The test can be easily adapted to incorporate equality between the main estimate and robustness checks, if the researcher thinks that would be appropriate (see Lu and White (2014) for details of when this should hold). This is done by replacing the unit vector in Equation (4) to a binary vector where the appropriate entries to 0. In fact, a very interesting aspect of the Cox and Shi (2023) test is that, when the whole binary vector is equal to 0, it reduces to the robustness test shown in Lu and White (2014).

Theoretical properties

The framework takes the set of robustness checks as given, with m fixed; which checks are run and reported is a separate question, and one where selection is documented: Campbell, Brodeur, Dreber, Johannesson, Kopecky, Lusher, and Tsoy (2024) find that author-reported specifications survive far more often than those run by independent reproducers. The properties below therefore do not describe what happens when a check is "added" to a running analysis; they compare the measure across two fixed sets of checks that differ by one check, each with its own target. Let the two sets be $\{1, \dots, m\}$ and $\{1, \dots, m+1\}$, write $M_n^{(m)}$ and $M_n^{(m+1)}$ for a measure computed on each (the main specification is always included), and let $\Delta_j = \theta_j - \theta_0$ denote the discrepancy of check j and $\hat{\Delta}_j = \hat{\theta}_j - \hat{\theta}_0$ its estimate. The targets are $b^{(m)}(\theta) = \max_{j \le m} |\Delta_j|$ and $b^{(m+1)}(\theta) = \max\{b^{(m)}(\theta), |\Delta_{m+1}|\}$, the population quantity (2) for each set. Let $\sigma^2_{\Delta,j}$ denote the asymptotic variance of $\sqrt{n}(\hat{\Delta}_j - \Delta_j)$ and $\hat{\sigma}^2_{\Delta,j}$ the estimate of it implied by $\hat{\Sigma}$, $j = 1, \dots, m+1$, and let $\bar{\sigma}_\Delta = \max_{j \le m+1} \sigma_{\Delta,j}$.

Two parties have a stake in how the measure differs across the two sets.

Property 1 (Referee). A measure M_n has the referee property if, whenever the estimand of check $m+1$ lies outside the interval spanned by checks $1, \dots, m$, the measure computed with it exceeds the measure computed without it:

(i) for every fixed $F \in \mathcal{F}$ with $|\Delta_{m+1}| > b^{(m)}(\theta)$, $\mathbb{P}_F\big(M_n^{(m+1)} > M_n^{(m)}\big) \to 1$;

(ii) for every sequence $\{F_n\} \subset \mathcal{F}$ along which

$$\frac{\sqrt{n}\left(|\Delta_{m+1}(F_n)| - b^{(m)}(\theta(F_n))\right)}{\bar{\sigma}_\Delta(F_n)} \longrightarrow \infty,$$

$$\mathbb{P}_{F_n}\big(M_n^{(m+1)} > M_n^{(m)}\big) \to 1.$$

Property 2 (Researcher). A measure M_n has the researcher property if a check whose estimand lies inside the interval cannot make the paper look less robust than it is, whatever its precision: for every $F \in \mathcal{F}$ with $|\Delta_{m+1}| \le b^{(m)}(\theta)$,

$$\limsup_{n\to\infty} \mathbb{P}_F\big(M_n^{(m+1)} > b^{(m)}(\theta)\big) \le \alpha,$$

with no restriction on $\sigma_{\Delta,m+1}$. In particular, for every $K > b^{(m)}(\theta)$, $\limsup_n P_F(M_n^{(m+1)} > K) \le \alpha$ whatever $\sigma_{\Delta,m+1}$: a noisy check can carry the measure far above the truth only on an event of probability at most α.

Part (i) of Property 1 is what a referee is entitled to: if the requested specification identifies something further from the main estimand than the other checks do, the measure must eventually say so. Part (ii) puts a floor on how fast. It requires the measure to respond once the excess discrepancy is large relative to the sampling uncertainty of the differences, measured by the largest of their standard errors, uniformly over the class of distributions; it does not say that the measure may not respond earlier, and the condition is sufficient, not necessary. What (ii) rules out are measures whose response depends on the excess discrepancy in the units of θ rather than in standard errors, so that a precisely measured departure can go undetected for a long time. Property 2 is the author's side of the same exercise: an imprecise check that happens to sit inside the null should not be allowed to push the measure above the truth. For a measure that takes robustness as the null, this is validity, since a check inside the interval leaves $b(\theta)$ unchanged.

The robustness tolerance has both properties, and the argument yields a finite-sample rule for when the two tolerances differ. Both rest on bounds that hold for every realization of the data.

Lemma 1. For the RCC test of $H_0(b)$ with m checks and any positive definite $\hat{\Sigma}$, for every realization of the data,

$$\max_{j\le m}\left(|\hat{\Delta}_j| - \hat{\sigma}_{\Delta,j}\sqrt{\chi^2_{m,1-\alpha}/n}\right)_+ \;\le\; b_{RT}(\alpha) \;\le\; \max_{j\le m}|\hat{\Delta}_j|,$$

where $\chi^2_{m,1-\alpha}$ is the $(1-\alpha)$ quantile of the chi-squared distribution with m degrees of freedom and $(x)_+ = \max\{x, 0\}$.

Proof. See Appendix A.5. □

The lower bound is the tolerance one would obtain from the single check j alone, with the full-rank critical value; the RCC test can only do better. The next result is stated under the high-level conditions of Appendix A.1 (Assumption 2: $\hat{\theta}$ is asymptotically linear uniformly over $\mathcal{F}$ and $\hat{\Sigma}$ is uniformly consistent), imposed on the larger set of checks, which Section 2.3 verifies for linear regressions with the plug-in covariance estimator.

Proposition 1 (The robustness tolerance has the referee and researcher properties). Suppose Assumption 2 holds for the set of checks $\{1, \ldots, m+1\}$, and let $b^{(m)}_{RT}(\alpha)$ and $b^{(m+1)}_{RT}(\alpha)$ be the robustness tolerances of Definition 1 computed from checks $\{1, \ldots, m\}$ and $\{1, \ldots, m+1\}$. Then:

(a) $b_{RT}(\alpha)$ satisfies Property 1. In addition, for every sequence $\{F_n\} \subset \mathcal{F}$ with

$$\frac{\sqrt{n}\left(|\Delta_{m+1}(F_n)| - b^{(m)}(\theta(F_n))\right)}{\sigma_{\Delta,m+1}(F_n)} \longrightarrow \infty,$$

one has $\liminf_{n\to\infty} \mathbb{P}_{F_n}\left(b^{(m+1)}_{RT}(\alpha) > b^{(m)}_{RT}(\alpha)\right) \geq 1-\alpha$.

(b) $b_{RT}(\alpha)$ satisfies Property 2.

(c) (Consistency.) For every fixed $F \in \mathcal{F}$,

$$b^{(m)}_{RT}(\alpha) \xrightarrow{p} b^{(m)}(\theta), \qquad b^{(m+1)}_{RT}(\alpha) \xrightarrow{p} b^{(m+1)}(\theta).$$

Proof. See Appendix A.5. □

The additional statement in part (a) is Property 1(ii) with $\bar{\sigma}_\Delta$ replaced by $\sigma_{\Delta,m+1}$ and probability one replaced by $1-\alpha$: on the event, of probability at least $1-\alpha$, that $b^{(m)}_{RT}(\alpha)$ does not exceed its target $b^{(m)}(\theta)$, whether the two tolerances differ depends only on the discrepancy of check $m+1$ relative to the standard error of its own difference. The precision of the other checks enters only through the remaining event, on which $b^{(m)}_{RT}(\alpha)$ can exceed its target by an amount of the order of the largest sampling error among the differences. The properties are requirements, not what distinguishes the RCC test: simpler lower bounds, such as the Bonferroni bound $\max_{j\leq m}\left(|\hat{\Delta}_j| - z_{1-\alpha/2m}\,\hat{\sigma}_{\Delta,j}/\sqrt{n}\right)_+$,

have both and are consistent, since with m fixed any fixed critical value is eventually dominated by a diverging signal-to-noise ratio. What separates them is what they do with the covariance of the discrepancies in finite samples. The Bonferroni bound uses only the standard errors and pays for every check through α/m whether or not it binds; the RCC test uses the full $\hat{\Sigma}$, drops the checks that do not bind, and is sharper whenever correlated checks make some combination of the discrepancies precisely estimated, the situation Section 3.2 documents.

For the robustness tolerance, part (a) says that what matters for check $m+1$ is the standard error of its *difference* $\hat{\Delta}_{m+1} = \hat{\theta}_{m+1} - \hat{\theta}_0$, not the standard error of $\hat{\theta}_{m+1}$. Writing σ_j^2 for the asymptotic variance of $\sqrt{n}(\hat{\theta}_j - \theta_j)$ and ρ_{0j} for the asymptotic correlation between $\hat{\theta}_j$ and $\hat{\theta}_0$,

$$\sigma_{\Delta,j}^2 = \sigma_j^2 + \sigma_0^2 - 2\rho_{0j}\,\sigma_0\sigma_j,$$

so two checks with the same σ_j can have very different $\sigma_{\Delta,j}$. When ρ_{0j} is close to one, as for a check on the same sample that adds a control nearly collinear with the controls already included, so that the two estimates are nearly the same, $\sigma_{\Delta,j}^2$ is close to $(\sigma_j - \sigma_0)^2$, which can be far below either σ_0^2 or σ_j^2; when $\rho_{0j} = 0$, as for a check estimated on a sample that does not overlap with the main one, $\sigma_{\Delta,j}^2 = \sigma_0^2 + \sigma_j^2$ and the difference is noisier than either estimate. Since the rate in part (a) is in units of $\sigma_{\Delta,m+1}$, referees should prioritize checks whose departure from the main specification is precisely measured.

If instead the check is known to be slack, it adds nothing to the target, and what it costs depends on the precision of its difference: Section 3 shows that a precisely measured slack check is asymptotically ignored by the test, while an imprecisely measured one is not, and can lower the tolerance.

The same questions can be asked of a measure that takes fragility as the null, that is, of an upper confidence bound for $b(\theta)$ obtained by inverting tests of $b(\theta) \geq b$. Such a measure cannot have the researcher property.

Proposition 2 (Every valid upper bound fails Property 2). Consider the normal model $\hat{\theta} \sim N(\theta, \Sigma/n)$ with Σ known and positive definite, checks $\{1, \ldots, m+1\}$, and let U_n be any upper confidence bound for $b(\theta)$: $\mathbb{P}_\theta(U_n \geq b(\theta)) \geq 1 - \alpha$ for every θ.

(a) For every θ and every n, $\mathbb{P}_\theta\big(U_n > b(\theta)\big) \geq 1 - \alpha$. In particular, if $|\Delta_{m+1}| \leq b^{(m)}(\theta)$, then $\mathbb{P}_\theta\big(U_n > b^{(m)}(\theta)\big) \geq 1 - \alpha$ for every n, so for $\alpha < 1/2$ the measure U_n fails Property 2.

(b) Let $\Sigma_{m+1,m+1} \to \infty$ with the correlation matrix Ω fixed. Then, for every $K < \infty$

and every θ,

$$\liminf_{\Sigma_{m+1,m+1}\to\infty} \mathbb{P}_\theta(U_n \geq K) \geq 1-\alpha.$$

In particular, under $\Delta_{m+1} = 0$, where $b^{(m+1)}(\theta) = b^{(m)}(\theta)$, U_n is not bounded in probability.

Proof. See Appendix A.5. □

Part (a) is close to immediate: a valid upper bound sits above the truth by construction, and Property 2 asks the measure to sit below it. It is stated to make the incompatibility exact: Property 2 is the defining property of a lower bound, and no measure that takes fragility as the null can have it. Part (b) is the substantive asymmetry. For a lower bound, a very noisy check is indistinguishable from a check on the boundary of the interval, and the tolerance exceeds the truth only on an event of probability at most α; for an upper bound it is indistinguishable from a check at infinity, and validity forces the bound to follow it with probability at least $1-\alpha$. The point comparison, $\max_j |\hat{\Delta}_j|$, the maximal estimated discrepancy, is a confidence bound in neither direction, but it fails Property 2 as well, since it exceeds $b^{(m)}(\theta)$ whenever the estimate of the largest discrepancy overshoots, an event of probability approaching at least one half; and it shares the behavior in part (b): under $\Delta_{m+1} = 0$ its term for check $m+1$ is of order $\sigma_{\Delta,m+1}/\sqrt{n}$, which diverges. This is the formal reason for framing robustness as the null.

2.3 Theoretical Validity for Linear Regressions

For the remainder of the paper, I focus on the case in which the estimands are coefficients from linear projections, which is a common case in Economics. In this section I define the covariance estimator $\hat{\Sigma}$ and write low level conditions on Ordinary Least Squares (OLS) estimators that are sufficient for the RCC test to be uniformly valid. These conditions verify, for linear regressions, the high-level conditions of Appendix A.1 (Assumption 2), under which Proposition 4 establishes the validity of the RCC test for any asymptotically linear estimator and for arbitrary families of polyhedral null hypotheses; the results of Section 2.2 are stated at that level of generality, and other estimators can be accommodated by verifying Assumption 2 directly. I introduce notation based on the traditional Frisch–Waugh–Lovell (FWL) decomposition.

Let $\mathcal{F}$ denote a class of distributions for the data. For each $F \in \mathcal{F}$, expectations and variances below are taken with respect to F.

For specification $j = 0, \ldots, m$, let D_{ji} denote the regressor of interest, and let Z_{ji} denote the vector of controls. It is often the case that robustness checks use slightly different sub-samples of the data, due to missing data patterns or data combination limitations; to address that, let $d_{ji} \in \{0, 1\}$ indicate whether observation i is included in specification j.

Define $X_{ji} = (D_{ji}, Z_{ji}')'$, and let $\beta_j = (\theta_j, \gamma_{yj}')'$ denote the coefficients of the population linear projection satisfying

$$E_F\left[d_{ji} X_{ji}\left(y_{ji} - X_{ji}'\beta_j\right)\right] = 0.$$

Write the population residual as $\varepsilon_{ji} = y_{ji} - X_{ji}'\beta_j$. Notice that not only controls Z_{ji} are allowed to vary with j, but also y_{ji}, D_{ji}, to accommodate robustness checks that possibly use different proxies for these variables. Let

$$Q_{Zj} = E_F[d_{ji} Z_{ji} Z_{ji}']$$

and define the population residual from projecting D_{ji} on Z_{ji} by

$$u_{ji} = D_{ji} - Z_{ji}'\gamma_{Dj}, \qquad \gamma_{Dj} = Q_{Zj}^{-1} E_F[d_{ji} Z_{ji} D_{ji}].$$

Define

$$\sigma_j^2 = \mathrm{Var}_F\left(\frac{d_{ji} u_{ji} \varepsilon_{ji}}{E_F[d_{ji} u_{ji}^2]}\right).$$

Let Σ denote the covariance matrix of the vector with jth element $\frac{d_{ji} u_{ji} \varepsilon_{ji}}{E_F[d_{ji} u_{ji}^2]}$, and let $V = \mathrm{diag}(\Sigma),\ \Omega = V^{-1/2} \Sigma V^{-1/2}$.

The covariance estimator $\hat{\Sigma}$ is the sample analogue of Σ. Let $\hat{\beta}_j$ denote the OLS coefficient vector from regressing y_{ji} on X_{ji} over the observations with $d_{ji} = 1$, let $\hat{\varepsilon}_{ji} = y_{ji} - X_{ji}'\hat{\beta}_j$ denote the corresponding residual, and let $\hat{u}_{ji}$ denote the residual from regressing D_{ji} on Z_{ji} over the same observations. Writing E_n for the sample average over $i = 1, \ldots, n$, define

$$\hat{\psi}_{ji} = \frac{d_{ji} \hat{u}_{ji} \hat{\varepsilon}_{ji}}{E_n[d_{ji} \hat{u}_{ji}^2]}, \qquad \hat{\psi}_i = (\hat{\psi}_{0i}, \ldots, \hat{\psi}_{mi})', \tag{12}$$

and

$$\hat{\Sigma} = E_n\left[\hat{\psi}_i \hat{\psi}_i'\right]. \tag{13}$$

By the Frisch–Waugh–Lovell theorem, $\hat{\psi}_{ji}$ is the estimated influence function of $\hat{\theta}_j$. The diagonal of $\hat{\Sigma}$ therefore collects the usual heteroskedasticity-robust variance estimators of

the $m+1$ specifications, and the off-diagonal entries extend them to pairs of specifications: an observation used in both specifications contributes the product of its two influence functions, and an observation used in only one of them contributes zero. The estimator requires neither resampling nor tuning parameters, and Lemma 3 in Appendix A.3 shows that it is uniformly consistent under the same conditions that deliver the asymptotic normality of $\hat{\theta}$, stated next. In the application of Section 4, where standard errors are clustered, $E_n[\hat{\psi}_i\hat{\psi}_i']$ is replaced by its cluster analogue, $n^{-1}\sum_g \big(\sum_{i\in g}\hat{\psi}_i\big)\big(\sum_{i\in g}\hat{\psi}_i\big)'$, where g indexes clusters. Appendix A.6 describes an alternative covariance estimator based on the pairs bootstrap. It is consistent for the same limit under additional conditions, so that the results below also hold with it in place of (13), and it may be convenient for estimators other than least squares.

Assumption 1. There exist constants $\epsilon > 0$, $\delta > 0$, and $K < \infty$ such that the following conditions hold uniformly over $F \in \mathcal{F}$:

(a) The vectors

$$W_i = (y_{0i}, \ldots, y_{mi}, X_{0i}', \ldots, X_{mi}', d_{0i}, \ldots, d_{mi})'$$

are i.i.d. across the n observations $i = 1, \ldots, n$.

(b) For every $j = 0, \ldots, m$, the dimension of Z_{ji} is fixed across n, and

$$\lambda_{\min}(Q_{Zj}) \geq \epsilon, \qquad E_F[d_{ji}u_{ji}^2] \geq \epsilon,$$

where $\lambda_{\min}(Q_{Zj})$ denotes the smallest eigenvalue of Q_{Zj}.

(c) For every $j = 0, \ldots, m$,

$$E_F\big[\|d_{ji}Z_{ji}\|_2^4\big] \leq K, \qquad E_F\big[d_{ji}|u_{ji}|^4\big] \leq K, \qquad E_F\big[|\varepsilon_{ji}|^2\|d_{ji}Z_{ji}\|_2^2\big] \leq K,$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

(d) For every $j = 0, \ldots, m$,

$$\sigma_j^2 \geq \epsilon, \qquad E_F\left[\left|\frac{d_{ji}u_{ji}\varepsilon_{ji}}{E_F[d_{ji}u_{ji}^2]\sigma_j}\right|^{2+\delta}\right] \leq K, \qquad \det(\Omega) \geq \epsilon.$$

Assumption 1(a) is the i.i.d sampling condition. Assumption 1(b) contains the usual conditions for the population linear projections, imposed uniformly over $\mathcal{F}$: the first ensures

that the projection of the treatment on the controls is well defined and does not become ill conditioned along any sequence of distributions, while the second requires the treatment to retain variation after the controls are partialled out, so that the Frisch–Waugh–Lovell denominator $E_F[d_{ji}u_{ji}^2]$ is bounded away from zero. Assumption 1(c) are standard moment conditions; they are the same order of moments as in the standard asymptotic theory of least squares.

Assumption 1(d) concerns the influence functions of the coefficients of interest, whose covariance matrix is Σ. The first condition requires the asymptotic variance of each $\hat{\theta}_j$ to be bounded away from zero uniformly; it is needed because the analysis standardizes each coordinate by its own asymptotic standard deviation. The second condition is a $(2+\delta)$ moment bound on the standardized influence function. It is the Lyapunov condition for the central limit theorem to hold uniformly over $\mathcal{F}$, and it is the type of condition typically imposed in the moment inequality literature; in particular it corresponds to the moment condition in Assumption 1 of Cox and Shi (2023). The third condition requires the correlation matrix Ω of the influence functions to be bounded away from singularity uniformly. The conditions in Assumption 1 are those needed for the validity of the RCC test given a uniformly consistent covariance estimator, and they are also sufficient for the plug-in estimator (13) to be uniformly consistent (Lemma 3 in Appendix A.3). No further conditions are needed.

Proposition 3. Suppose Assumption 1 holds, and let $\phi_n^{\mathrm{RCC}}(b,\alpha)$ be the RCC test of $H_0(b)$ constructed from $\hat{\theta}$ and the plug-in covariance estimator $\hat{\Sigma}$ in (13). Then, for $\alpha \in (0, 1/2)$:

(a) The RCC test is uniformly asymptotically valid:

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{\substack{b\geq 0:\\ A\theta(F)\leq b\mathbf{1}_{2m}}} E_F\left[\phi_n^{\mathrm{RCC}}(b,\alpha)\right] \leq \alpha.$$

(b) For every fixed $F \in \mathcal{F}$, let

$$b_F = b(\theta(F)) = \max_{j=1,\ldots,m} |\theta_j(F) - \theta_0(F)|.$$

Then

$$\lim_{n\to\infty} E_F\left[\phi_n^{\mathrm{RCC}}(b_F,\alpha)\right] = \alpha,$$

Proof. See Appendix A.2. □

Proposition 3 holds for any covariance estimator satisfying the uniform consistency condition of Lemma 3; the plug-in estimator is the one that requires no conditions beyond Assumption 1. Appendix A.6 verifies the condition for the pairs-bootstrap covariance estimator under additional assumptions.

Corollary 1 (Relative robustness tolerance). Suppose Assumption 1 holds, and let $\phi_n^{\mathrm{rel}}(c,\alpha) = \phi_n^+(c,\alpha)\phi_n^-(c,\alpha)$ be the intersection–union test of Definition 2 with $\phi_n^\pm$ the RCC tests of Section 2.2, constructed from $\hat{\theta}$ and $\hat{\Sigma}$ in (13). Let $c(\theta)$ denote the population relative tolerance defined after Definition 2. Then, for $\alpha \in (0, 1/2)$:

(a) The intersection–union test is uniformly asymptotically valid over F and c jointly:

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{\substack{c\geq 0:\\ \max_{j=1,\ldots,m} |\theta_j(F)-\theta_0(F)|\leq c|\theta_0(F)|}} E_F\left[\phi_n^{\mathrm{rel}}(c,\alpha)\right] \leq \alpha.$$

(b) The relative robustness tolerance is an asymptotically uniformly valid lower confidence bound for the population relative tolerance:

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}} P_F(c_{RT}(\alpha) > c(\theta(F))) \leq \alpha,$$

where the event is empty when $c(\theta(F)) = +\infty$.

(c) For every fixed $F \in \mathcal{F}$ with $\theta_0(F) \neq 0$, or with $\theta(F) = \mathbf{0}$,

$$\lim_{n\to\infty} E_F\left[\phi_n^{\mathrm{rel}}\big(c(\theta(F)),\alpha\big)\right] = \alpha.$$

Proof. See Appendix A.4. □

Part (c) is pointwise in F and excludes the case $\theta_0(F) = 0$ with $\theta(F) \neq \mathbf{0}$, in which $c(\theta(F)) = +\infty$ and there is nothing to cover. Along sequences F_n with $\sqrt{n}\,\theta_0(F_n)/\sigma_0(F_n)$ bounded and $\theta(F_n) \not\to \mathbf{0}$, that is, with the main estimand within a bounded number of standard errors of zero, the two cones $P_+(c)$ and $P_-(c)$ meet only near the origin, the wrong-sign branch no longer rejects with probability approaching one, and the intersection–union may be conservative. This is the familiar degeneracy of inference on a ratio when the denominator approaches zero, and it is the reason the relative tolerance should be reported together with $b_{RT}(\alpha)$, which is unaffected.

3 Simulations and Illustrations

This section studies the finite-sample behavior of the robustness tolerance by simulation. Section 3.1 establishes what the test responds to: the estimated discrepancies and their covariance matrix, instead of levels of $\hat{\theta}$. This observation dictates the design of the experiments and the units in which their results should be read. The remaining subsections then vary, one at a time, the configuration of the true discrepancies, the correlation among their estimates, and the number of checks, and document when the tolerance is close to the true maximum discrepancy, when it is far from it, and why. All simulations are run in the Gaussian limit experiment of Assumption 2: $\hat{\theta}$ is drawn from $N(\theta, \Sigma)$ with Σ known, which is the model the test faces asymptotically, and the test is applied with $n = 1$. All results use 1,000 replications and $\alpha = 0.05$.

3.1 A test of discrepancies

Section 2.2 stated the properties of the tolerance in terms of the discrepancies $\Delta_j = \theta_j - \theta_0$, their estimates $\hat{\Delta}_j$, and the standard errors $\sigma_{\Delta,j}$ of the latter, and Lemma 1 bounded the tolerance by functions of these alone. This is not a convenience of exposition: the null hypothesis $H_0(b)$ restricts the discrepancies and nothing else, and so does the test. Collect the estimated discrepancies in $\hat{\Delta} = (\hat{\Delta}_1, \ldots, \hat{\Delta}_m)' = L\hat{\theta}$, with $L = [-\mathbf{1}_m,\ I_m]$, and let $\hat{\Sigma}_\Delta = L\hat{\Sigma}L'$ be the implied estimator of the asymptotic variance of $\sqrt{n}(\hat{\Delta} - \Delta)$, whose diagonal entries are the $\hat{\sigma}^2_{\Delta,j}$ of Section 2.2.

Lemma 2 (Reduction to the discrepancies). For every $b \geq 0$ and every positive definite $\hat{\Sigma}$,

$$T_n(b) = \min_{\zeta \in \mathbb{R}^m:\ \|\zeta\|_\infty \leq b} n(\hat{\Delta} - \zeta)'\hat{\Sigma}_\Delta^{-1}(\hat{\Delta} - \zeta),$$

and the active set and rank $\hat{r}(b)$ in (10)–(11) coincide with those of the minimizer $\hat{\zeta}(b)$ in the interval $\|\zeta\|_\infty \leq b$. Therefore, $b_{RT}(\alpha)$ is a function of $(\hat{\Delta}, \hat{\Sigma}_\Delta)$ alone.

Proof. The constraint $A\mu \leq b\mathbf{1}_{2m}$ in (8) reads $\|L\mu\|_\infty \leq b$, so the feasible set is the union over ζ with $\|\zeta\|_\infty \leq b$ of the subspaces $\{\mu : L\mu = \zeta\}$. For fixed ζ, the minimum of $n(\hat{\theta} - \mu)'\hat{\Sigma}^{-1}(\hat{\theta} - \mu)$ over $\{\mu : L\mu = \zeta\}$ equals $n(L\hat{\theta} - \zeta)'(L\hat{\Sigma}L')^{-1}(L\hat{\theta} - \zeta)$, the standard formula for a quadratic form minimized subject to linear equality constraints, and $L\hat{\mu}(b) = \hat{\zeta}(b)$ at the optimum. The active inequalities are those with $|\hat{\zeta}_j(b)| = b$, in both formulations. □

The lemma is specific to $H_0(b)$. The relative null $H_0^{rel}(c)$ in Definition 2 restricts $|\theta_j - \theta_0|$ relative to $|\theta_0|$, so it involves the level of the estimates and not only their discrepancies; c_{RT} is accordingly not covered by what follows, and this section is about b_{RT}.

Two consequences of the lemma organize the design. First, the tolerance is invariant to the common level of the estimates: adding the same constant to every $\hat{\theta}_j$, or to every θ_j, changes nothing, so the simulations set $\theta_0 = 0$ without loss of generality. Second, two designs that differ in the covariance matrix of the estimates but share Σ_Δ produce the same distribution of b_{RT}. The simulations therefore parametrize Σ_Δ directly and set its diagonal to one, so that both the true maximum discrepancy, $\bar{\Delta} = \max_j |\Delta_j| = b(\theta)$, and the tolerance are measured in standard errors of a discrepancy.[12]

This scale determines what the figures can show. With a single check, the RCC test of $H_0(b)$ does not reject if and only if

$$|\hat{\Delta}_1| \leq b + c_\alpha(b/\hat{\sigma}_{\Delta,1})\,\hat{\sigma}_{\Delta,1}, \qquad c_\alpha(t) = z_{1-\alpha\Phi(t)},$$

where the critical constant depends on the threshold through the refinement: the slack of the mirror inequality at the restricted estimator is b in units of $\hat{\sigma}_{\Delta,1}$, and the refinement sets the level to $2\alpha\Phi(b/\hat{\sigma}_{\Delta,1})$. Thus c_α equals $z_{1-\alpha/2} = 1.96$ at $b = 0$ and decreases to $z_{1-\alpha} = 1.64$ as $b/\hat{\sigma}_{\Delta,1}$ grows. The tolerance is zero when $|\hat{\Delta}_1| \leq 1.96\,\hat{\sigma}_{\Delta,1}$ and otherwise solves the equation above; it lies between $(|\hat{\Delta}_1| - 1.96\,\hat{\sigma}_{\Delta,1})_+$ and $(|\hat{\Delta}_1| - 1.64\,\hat{\sigma}_{\Delta,1})_+$, and equals the latter to three decimals once $|\hat{\Delta}_1|$ exceeds five standard errors.[13] I refer to the large-signal value $z_{1-\alpha} = 1.64$ as the one-check gap: the distance, in standard errors of the discrepancy, between the truth and the tolerance on average when a single check is used and the discrepancy is well away from zero. The formula has two readings. Away from zero, the gap approaches 1.64 on average, whatever the value of $\bar{\Delta}$; whether it looks large or small depends only on how many standard errors of a discrepancy $\bar{\Delta}$ is worth. Near zero, the tolerance is *censored*: $b_{RT} = 0$ whenever $|\hat{\Delta}_1| \leq c_\alpha\hat{\sigma}_{\Delta,1}$, which is the event that the test does not reject $b = 0$. When the true discrepancy is within a couple of standard errors of zero this event has substantial probability, and the mean of b_{RT} is pulled toward zero by it. This is not a feature of the definition but of the power of the test at $b = 0$; with several checks it depends on their configuration and correlation in a way that Section 3.3 makes exact. The ratio $\bar{\Delta}/\sigma_\Delta$ is also the only place where a sample size would appear.

[12] A covariance matrix of the estimates that induces a given Σ_Δ always exists: $\Sigma = \sigma_0^2\mathbf{1}\mathbf{1}' + \mathrm{diag}(0, \Sigma_\Delta)$ gives $L\Sigma L' = \Sigma_\Delta$ for any $\sigma_0^2 > 0$, and by Lemma 2 the value of σ_0^2 is irrelevant.

[13] With $\hat{\sigma}_{\Delta,1} = 1$ and $\alpha = 0.05$, the tolerance is 0.060, 0.730, 1.307 and 3.355 at $|\hat{\Delta}_1| = 2$, 2.5, 3 and 5, that is, the estimate less 1.94, 1.77, 1.69 and 1.65.

The limit experiment has none; a sample of size n is approximated by the limit experiment with covariance Σ/n, in which a discrepancy of $\bar{\Delta}$ in the units of the estimand is $\sqrt{n}\,\bar{\Delta}/\sigma_\Delta$ standard errors. Holding $\sigma_\Delta = 1$ and varying $\bar{\Delta}$ is therefore the same as holding the discrepancy fixed and varying $\sqrt{n}$.

I report two values of $\bar{\Delta}$. At $\bar{\Delta} = 2.5$ the discrepancies are only two and a half standard errors from zero, and the configuration of the checks and the correlation of their estimates decide how often the tolerance is censored and how far from the truth it is when it is not. At $\bar{\Delta} = 10$ censoring did not occur in any replication, its probability being of the order of $\Phi(-8)$, and the mean tolerance is within two standard errors of the truth in every configuration; configuration and correlation decide only the size of that gap. Runs at $\bar{\Delta} = 5$ and 20 confirm what the one-check formula says: the gap $\bar{\Delta} - b_{RT}$ has, up to events of negligible probability, the same distribution at every $\bar{\Delta}$ large enough that censoring does not occur: the draws are shared across $\bar{\Delta}$, and once the refinement's slack is large the critical values coincide, so the $\bar{\Delta} = 20$ figure is the $\bar{\Delta} = 10$ figure shifted up by ten, line by line, and $b_{RT}/\bar{\Delta} \to 1$ at the rate $1/\sqrt{n}$ in every cell. This is the consistency of Proposition 1(c) in the Gaussian experiment. With several checks, the gap between the truth and the tolerance could in principle grow with their number and depend on their correlation. Lemma 1 bounds how much: the tolerance is within $\sqrt{\chi^2_{m,1-\alpha}}$ standard errors of the largest estimated discrepancy, 4.1 at $m = 9$ against a one-check gap of 1.64, and since the estimate of the largest discrepancy exceeds it in expectation, the same bound holds for the expected gap to the truth in the Gaussian experiment. The simulations show that the actual gap stays close to the one-check gap, because the test drops the checks that are not binding: the number of checks adds no new scale to the picture, only the configuration and correlation effects documented next.

Since the correlation of the discrepancies plays this role, it is worth being explicit about where it comes from, because it is not the correlation of the estimates. Every discrepancy contains the same $\hat{\theta}_0$, so the covariance of two of them is

$$\mathrm{Cov}(\hat{\Delta}_j, \hat{\Delta}_k) = \mathrm{Cov}(\hat{\theta}_j, \hat{\theta}_k) - \mathrm{Cov}(\hat{\theta}_j, \hat{\theta}_0) - \mathrm{Cov}(\hat{\theta}_k, \hat{\theta}_0) + \mathrm{Var}(\hat{\theta}_0),$$

so the shared baseline estimate pushes the correlation of the discrepancies up, and it is non-positive only if the two checks covary with the baseline by more, in total, than they covary with each other plus the baseline's own variance. When the three estimates are independent, for instance because the checks use disjoint subsamples, the correlation is $\mathrm{Var}(\hat{\theta}_0)/\{(\sigma_0^2 + \sigma_j^2)(\sigma_0^2 + \sigma_k^2)\}^{1/2} > 0$ whatever the variances. The value $\rho_\Delta = 0$ used

below is the limiting case in which the two covariances with the baseline exactly offset its variance. Section 3.2 gives examples of robustness checks at each end of the range.

3.2 Configuration and correlation

Figures 1 and 2 report the mean tolerance across 1,000 replications for $m = 1, \ldots, 9$ checks whose discrepancies are equicorrelated with correlation $\rho_\Delta \in \{0, 0.5, 0.9, 0.95\}$ and unit standard errors, at $\bar{\Delta} = 2.5$ and $\bar{\Delta} = 10$. Three configurations of the true discrepancies are considered: all m checks deviate by $\bar{\Delta}$ with alternating signs; all deviate by $\bar{\Delta}$ with the same sign; one deviates by $\bar{\Delta}$ and the other $m - 1$ have discrepancy zero. The draws are the same across the two figures up to the shift in $\bar{\Delta}$.

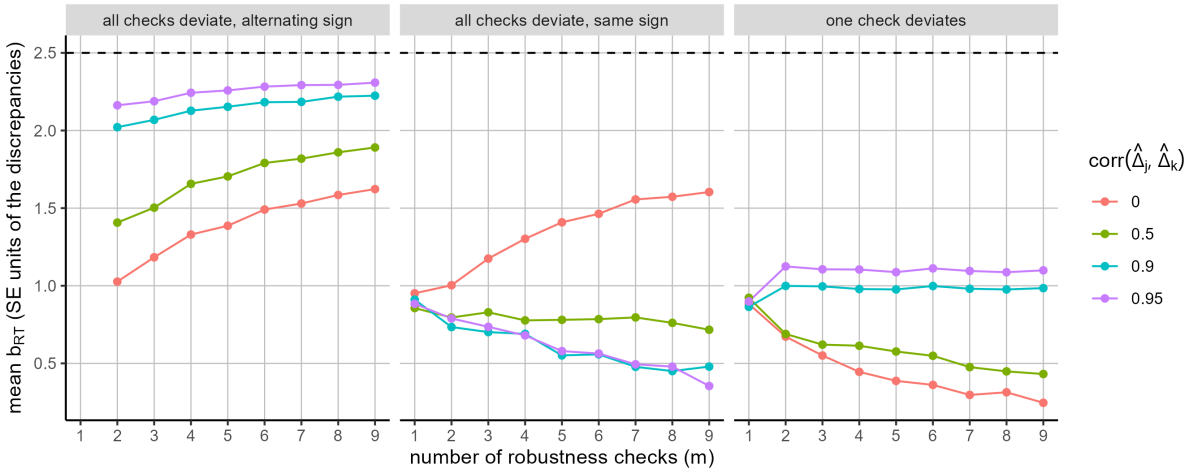


Figure 1: Mean b_{RT} by number of checks and correlation of the discrepancies, $\bar{\Delta} = 2.5$. Unit standard errors of the discrepancies, 1,000 replications, $\alpha = 0.05$. The dashed line is $\bar{\Delta}$.

The three panels of Figure 1 differ because correlation among the discrepancies changes what the data measure precisely. When the discrepancies are correlated, differences between checks, $\theta_j - \theta_k$, are estimated with standard error of order $\sqrt{1 - \rho_\Delta}$, while the average discrepancy, $\mathbf{1}'\Delta/m$, is estimated with standard error of order one, however many checks there are. A configuration is therefore well determined to the extent that it is made of differences between checks. Alternating signs are pure differences between checks: the tolerance rises with ρ_Δ and with m, from 1.03 at $\rho_\Delta = 0$ and $m = 2$ to 2.31 at $\rho_\Delta = 0.95$ and $m = 9$, against a truth of 2.5. Same-sign deviations are pure average, with no differences between checks: under correlation, m checks that all move by $\bar{\Delta}$ in the same direction are one piece of evidence seen m times, so the tolerance does not rise with m but falls, to 0.36 at $\rho_\Delta = 0.95$ and $m = 9$, a gap of 2.1 against a one-check gap of about 1.6;

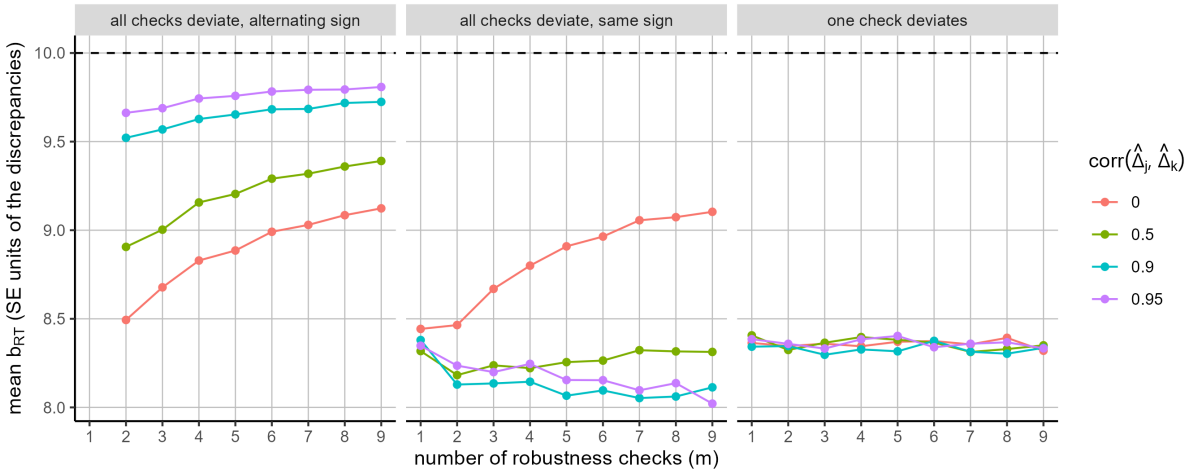


Figure 2: Mean b_{RT} by number of checks and correlation of the discrepancies, $\bar{\Delta} = 10$. Same design and draws as Figure 1.

only when the discrepancies are uncorrelated do additional checks add information, and the tolerance rises from 0.95 to 1.60. A single deviating check differs from the others, and when they are correlated with it the slack checks act as controls: the tolerance is flat in m at about 1.0–1.1 for $\rho_\Delta \geq 0.9$. When they are uncorrelated they carry no information about the deviating check and are not yet ignored at this distance from zero, and the tolerance falls from 0.88 at $m = 1$ to 0.25 at $m = 9$. Section 3.3 shows that this decline is mostly the probability of a zero.

Figure 2 is the same experiment ten times further from zero, and all of the above survives only as the distance between the tolerance and the truth. Every line lies between 8.0 and 9.8. The one-check panel is flat at 8.3–8.4 for every m and ρ_Δ, that is, at $\bar{\Delta} - c_\alpha$, because the slack checks are now dropped by the test whatever their correlation; the same-sign panel sits at 8.0–8.4 for $\rho_\Delta \geq 0.5$ and rises to 9.1 at $\rho_\Delta = 0$; the alternating panel runs from 8.5 to 9.8.

Which correlation applies is determined by how the checks are constructed, and it is known from $\hat{\Sigma}$ before any discrepancy is examined. The application of Section 4 shows the range. Nested specifications that add controls to one another produce discrepancies correlated at 0.84–0.89, and specifications that change the set of fixed effects at 0.54–0.67: each estimate shares most of its sampling variation with the main estimate, and the discrepancies inherit that. Alternative weighting schemes on the same sample produce discrepancies correlated at 0.00–0.06 (except for two nested weights, at 0.55), with standard errors as large as those of the estimates themselves. The reason is that an estimate's

sampling variation comes from the observations that receive large weight: weighting by the number of workers lets a few large jobs drive the estimate, while the main specification weights jobs equally, so the two estimates are computed from the same rows but their sampling errors have little in common, and nothing cancels in the difference. Alternative definitions of the treatment fall in between, at 0.40. Using the same data is therefore not enough for the discrepancies to be precise or correlated; the checks must also draw on the same observations in the same way. Which regime a set of checks is in is a property of its design, read off $\hat{\Sigma}$, and it is known before any discrepancy is seen.

3.3 The case of $b = 0$

At $b = 0$ the null is a point, every constraint is active whatever the precision of the check, $\hat{r}(0) = m$, and the test of $H_0(0)$ is the joint test of m equalities, $T_n(0) = \hat{\Delta}'\hat{\Sigma}_\Delta^{-1}\hat{\Delta}$ against $\chi^2_{m,1-\alpha}$. In the limit experiment $T_n(0)$ is noncentral chi-squared with m degrees of freedom and noncentrality $\lambda = \Delta'\Sigma_\Delta^{-1}\Delta$, so

$$P(b_{RT} = 0) = P\big(\chi^2_m(\lambda) \le \chi^2_{m,1-\alpha}\big), \tag{14}$$

one minus the power of the joint test. The formula is exact in the Gaussian experiment despite the non-monotonicity of Section 2.2: at $b = 0$ every constraint is active and the rank is already m, so the critical value cannot rise as b increases, and $T_n(b)$ is continuous in b, so rejection at 0 implies rejection on a neighborhood of 0; hence $b_{RT} = 0$ if and only if $H_0(0)$ is not rejected. Figure 3 plots (14) at $\bar{\Delta} = 1.5$ and 2.5 next to the simulated frequency at $\bar{\Delta} = 2.5$.

Two features of the figure carry over to practice. First, with a single deviating check and uncorrelated checks with zero discrepancy, λ is $\bar{\Delta}^2$ whatever m, while the critical value grows with m: each slack check adds one unit of noise to $T_n(0)$ in expectation and about 1.5–2 units to $\chi^2_{m,1-\alpha}$, so the zero becomes more likely as checks are added. This is the decline in the one-check panel of Figure 1. Second, correlation changes λ, not the critical value, and it does so by configuration. For one deviating check, correlated slack checks pin down the common level, and for $\bar{\Delta} = 2.5$, λ rises from $\bar{\Delta}^2 = 6.25$ at $\rho_\Delta = 0$ to 111 at $\rho_\Delta = 0.95$ and $m = 9$. For same-sign deviations, $\lambda = m\bar{\Delta}^2/(1 - \rho_\Delta + m\rho_\Delta)$, which tends to $\bar{\Delta}^2/\rho_\Delta$ as m grows. This is the case to name: correlated checks that all move the estimate in the same direction by a modest amount, as when a sequence of added controls shifts every estimate the same way, are the configuration in which a zero tolerance

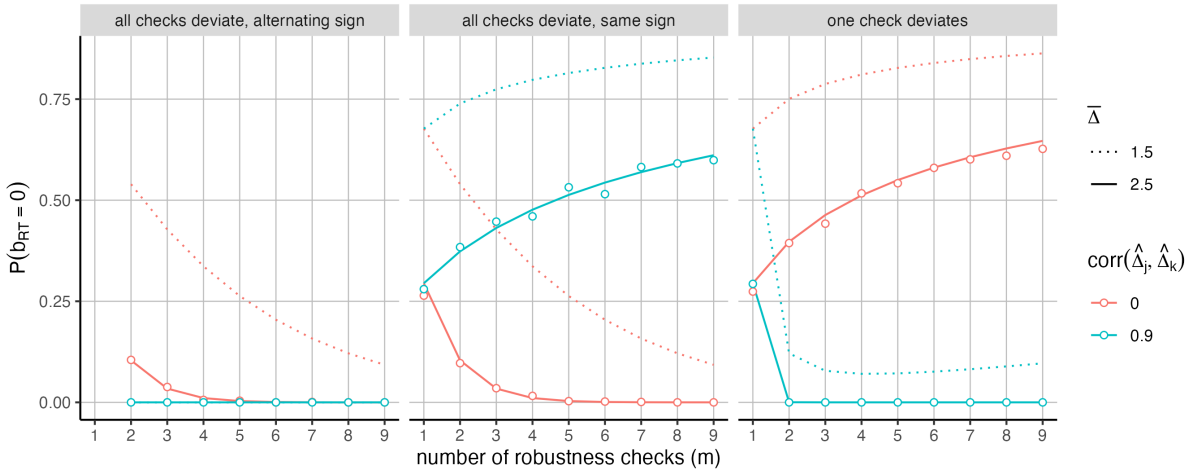


Figure 3: Probability of a zero tolerance. Lines: $P(\chi^2_m(\lambda) \leq \chi^2_{m,1-\alpha})$ with $\lambda = \Delta'\Sigma_\Delta^{-1}\Delta$, equation (14), at $\bar{\Delta} = 2.5$ (solid) and $\bar{\Delta} = 1.5$ (dotted); circles: simulated frequency of $b_{RT} = 0$ at $\bar{\Delta} = 2.5$, 1,000 replications. Unit standard errors of the discrepancies, equicorrelation ρ_Δ, $\alpha = 0.05$. At $\bar{\Delta} = 5$ the probability is below 0.04 in every cell.

is least informative. In both cases the zero is a statement about power. The joint test of m equalities has little power against one deviating coordinate when the others add only noise, and little power against a common shift when the checks are correlated. A zero tolerance says that the data do not reject equality of the estimands; it does not say that they are equal, and in these configurations it would not, even with discrepancies a few standard errors from zero.

3.4 Adding a check

In Figures 1 and 2 every check has the same standard error. Section 2.2 made a claim about checks whose standard errors differ: within one set, a precisely measured check with zero discrepancy is ignored by the test, and an imprecisely measured one is not. Figure 4 adds to a base set of three checks, one deviating by $\bar{\Delta}$ and two with zero discrepancy, a fourth check with zero discrepancy and standard error 0.2, 1 or 5, independent of the others, and computes the tolerance with and without it on the same draws.

The precise check ($\sigma_{\Delta,4} = 0.2$) is inactive at the solution: it contributes nothing to the statistic or to the degrees of freedom, and the test is the same as without it. Away from zero ($\bar{\Delta} = 10$) this holds for all correlations; near zero ($\bar{\Delta} = 2.5$) it fails only when the base checks are uncorrelated and the tolerance is often censored. The noisy ($\sigma_{\Delta,4} = 5$) check is never slack in its own standard errors near zero, and it lowers the tolerance, in most draws,

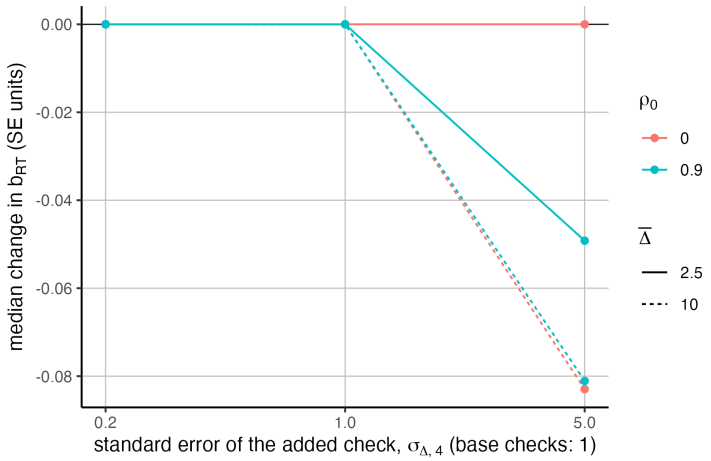


Figure 4: Adding one check with zero discrepancy. Base set of three checks with equicorrelation ρ_0 and unit standard errors, one deviating by $\bar{\Delta}$; the added check has discrepancy zero, standard error $\sigma_{\Delta,4}$, and is independent of the base discrepancies. Median change in the tolerance when the check is added, in standard errors of the base discrepancies. 1,000 replications, $\alpha = 0.05$.

by two routes. Near zero it is active, adds a term of order one to the statistic and a degree of freedom to the critical value, and the critical value grows faster; the tolerance falls in about three quarters of the draws at $\bar{\Delta} = 2.5$ when the base checks are correlated, and about half when they are not, where it is often zero with or without the check. Away from zero it is usually inactive, but the refinement of the RCC test sets the critical value by how far the inactive constraints are from binding in their own standard errors, and at $\bar{\Delta} = 10$ a check with standard error 5 is within two of its own standard errors of binding: the tolerance falls in nearly every draw, by about a tenth of a standard error at the median. The effect fades as $\bar{\Delta}$ grows and the noisy check's slack, in its own standard errors, with it. The figure reports the median change; the mean is pulled up by the draws, of probability about α, in which the noisy check's own estimate falls far outside the interval and raises the tolerance by an amount proportional to its standard error. This is the finite-sample content of the claim in Section 2.2: a check inside the interval costs nothing when its discrepancy is precisely measured, and typically lowers the tolerance when it is not. As there, the comparison is between two fixed sets of checks, each with its own target; the framework takes the set as given and the level α is spent once on it.

4 Application: "Wage Stagnation and the Decline of Standardized Pay Rates, 1974–1991"

Massenkoff and Wilmers (2023) examine how transitioning from wage-setting practices based on job titles and seniority to managerial discretion affects wages; they find wage declines, particularly among the lowest-paid workers. In a context of declining worker bargaining power, they conclude that the transition to discretionary pay amplified blue-collar workers' real wage losses during the 1980s. While they do not claim standardized pay rates to be the primary driver of wage stagnation, the argument is that the mechanism had acted as a barrier against market pressures during the shift from high to low worker bargaining power. The authors use microdata from the Wage Fixing Authority Survey (WFAS), which contains information on pay levels and pay-setting practices for 50,000 workplaces surveyed between 1974 and 1991.

I replicate[14] Table 2 of the paper, which is the main table exploring the effects of nonstandardized pay on wages, and the two sets of explicit robustness checks the authors report for its specifications (3) and (4): alternative weighting schemes (their Table A.2) and alternative definitions of flexible pay (their Table A.3). The replicated tables are in Appendix A.8, Tables 2, 3 and 4; every estimate they contain also appears in the figures below. Specification (3) has occupation-year-city-industry fixed effects and specification (4) adds occupation-establishment fixed effects. Because the authors are not explicit about which of the two is the preferred specification, and run both sets of checks for each, I compute the tolerances separately for each.

Throughout, the covariance matrix of the stacked estimates is the plug-in estimator (13) of Section 2.3, with the influence functions summed within establishments, the level at which the authors cluster; its diagonal reproduces the clustered standard errors reported in the tables. Table 1 collects the tolerances, and Figures 5–8 show, for four of the cases, the estimates with their confidence intervals and the discrepancies with theirs, together with the tolerance and the point comparison $\max_j |\hat{\Delta}_j|$. The figures are the application counterpart of Section 3: what the reader sees in a robustness table is the top panel; what the test sees is the bottom one.

[14] Results vary slightly, because I use R instead of Stata; in R, I use package `fixest v0.11.2` to estimate the high-dimensional fixed effects specifications, using function `feols` with argument `fixef.rm = "singleton"`, to drop singleton fixed effects. The algorithm is such that this is done in one pass. The authors use Stata's `reghdfe`, which instead removes singletons recursively. This is the cause of the difference in the number of observations in columns (4) and (5). But estimators do not have any visible differences. See https://github.com/lrberge/fixest/issues/478 for more details on the algorithm.

4.1 Alternative weighting schemes

In the main tables each job (establishment-by-occupation) is one equal unit. Table A.2 reweights observations by the number of employees at the pay level, by the survey weights, by weights from the Current Population Survey, and by the number of employees after excluding the largest 5% of jobs. The authors conclude: *"The results are largely consistent with the job-weight results presented in the main tables. The exception is the employee-weighted job-by-establishment fixed effect model. This is driven entirely by the large jobs (above 31 workers) that make up 5 percent of the observations but 54 percent of the employment-weighted sample."*

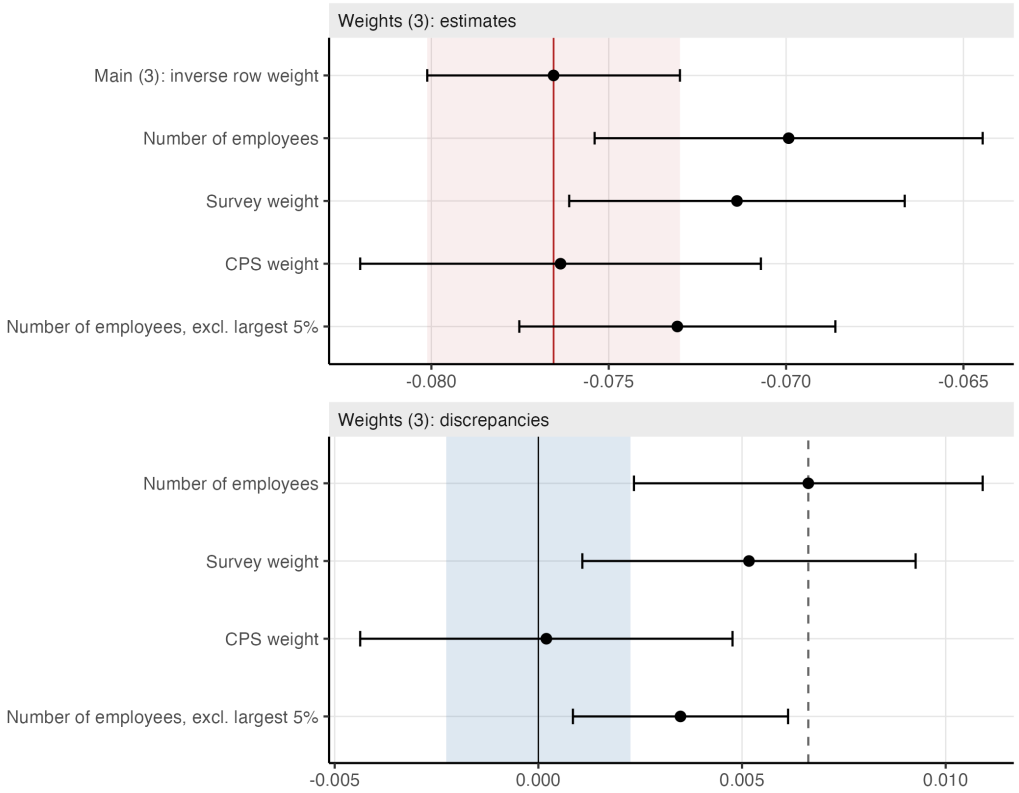


Figure 5: Alternative weighting schemes, centered at column (3) of Table 2. Top: estimates with ±1.96 standard errors; the red line and band are the main estimate and its interval. Bottom: discrepancies $\hat{\Delta}_j$ with ±1.96 standard errors of the discrepancy; the solid line marks zero discrepancy, the shaded interval is $[-b_{RT}, b_{RT}]$ with $b_{RT} = 0.0023$, and the dashed line is the point comparison $\max_j |\hat{\Delta}_j| = 0.0066$.

Figure 5 centers the comparison at specification (3). In the top panel, every robustness estimate's interval overlaps the main estimate's, and three of the four point estimates lie inside it. The bottom panel shows what the test sees. Three of the four discrepancies

exclude zero, at between two and a half and three standard errors, so the data reject $b = 0$; what they establish is $b_{RT} = 0.0023$, with the employee-weighted and survey-weighted checks binding: about a third of the point comparison, and about 3% of the main effect ($c_{RT} = 0.030$). The discrepancies are as imprecise as the estimates themselves and nearly uncorrelated with one another, because reweighting changes which observations carry the estimate and, in this case, a reweighted estimate shares little sampling variation with the main one. These checks are in the regime of Figure 1, where the discrepancies are only a few standard errors from zero and the tolerance falls well short of the point comparison $\max_j |\hat{\Delta}_j|$.

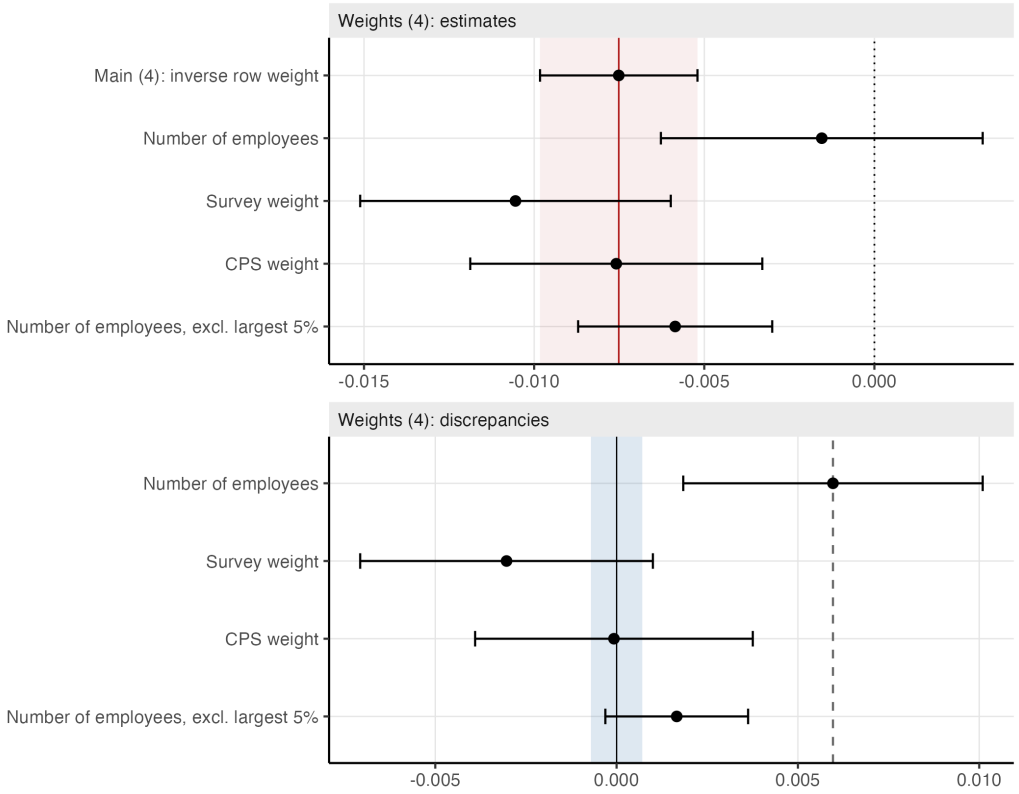


Figure 6: Alternative weighting schemes, centered at column (4) of Table 2. Top: estimates with ±1.96 standard errors; the red line and band are the main estimate and its interval, and the dotted line marks zero. Bottom: discrepancies $\hat{\Delta}_j$ with ±1.96 standard errors of the discrepancy; the solid line marks zero discrepancy, the shaded interval is $[-b_{RT}, b_{RT}]$ with $b_{RT} = 0.0007$, and the dashed line is the point comparison $\max_j |\hat{\Delta}_j| = 0.0060$.

Figure 6 centers the comparison at specification (4), whose main effect is a tenth of the previous one and six standard errors from zero. The employee-weighted check is the authors' exception: its estimate is about a fifth of the main effect, with a discrepancy of 2.8 standard errors, or 79% of the main effect by the point comparison. The other

three discrepancies straddle zero. The test rejects $b = 0$ on the strength of the one check and gives $c_{RT} = 0.09$ for the set of four. The employee-weighted check on its own gives $c_{RT} = 0.31$: the other three discrepancies straddle zero but, with standard errors as large as the binding one's, they are not slack in their own units, remain active at the solution, and raise the critical value, the mechanism of Section 3.4. The authors' set of checks, which is the one reported, establishes a departure of a tenth of the main effect; the check they single out establishes, by itself, a departure of three tenths. Excluding the largest 5% of jobs, the authors' explanation of the exception, removes the binding check: the remaining discrepancies are all within two standard errors of zero, the joint test of equality does not reject, and both the absolute and relative tolerances are zero. Centered at (3), the same exclusion roughly halves the tolerance.

4.2 Alternative definitions of flexible pay and the full set

Table A.3 replaces the treatment, pay varying with merit, a combination of merit and seniority, or other nonseniority reasons, by two narrower definitions. The authors write that *"this coding decision has little impact on the wage effect estimates: flexible or nonstandardized pay effects are consistently negative across variable definitions."* Figure 7 centers the comparison at (3). The narrow-merit discrepancy, 0.0121, is 6.2 standard errors from zero, and its interval lies entirely outside the main estimate's in the top panel: here the eye and the test agree. The tolerance is $b_{RT} = 0.0089$, three quarters of the point comparison and 12% of the main effect ($c_{RT} = 0.118$), with only the narrow-merit check binding; the merit-or-combination check, at 1.5 standard errors, is inactive. "Little impact" is therefore a departure of at least a tenth of the effect, which is also what the top panel shows once the intervals are drawn. Centered at (4), both discrepancies are within two standard errors of zero, and the tolerance is zero.

Running the four weighting checks and the two definition of pay checks as a single set of six shows what adding checks whose discrepancies lie inside the interval does. Centered at (3), the tolerance is 0.0087 against 0.0089 from the two definition checks alone: the four weighting discrepancies lie inside the interval and none is active at the solution; the small drop is the refinement of Section 3.4 at work, since with one binding check the critical value is set by the slack of the nearest inactive one in its own standard error, and the nearest weighting check is not far. Centered at (4), where every discrepancy is within a few standard errors of zero, all six constraints are active and the tolerance is 0.0001: $b = 0$ is rejected, but only just.

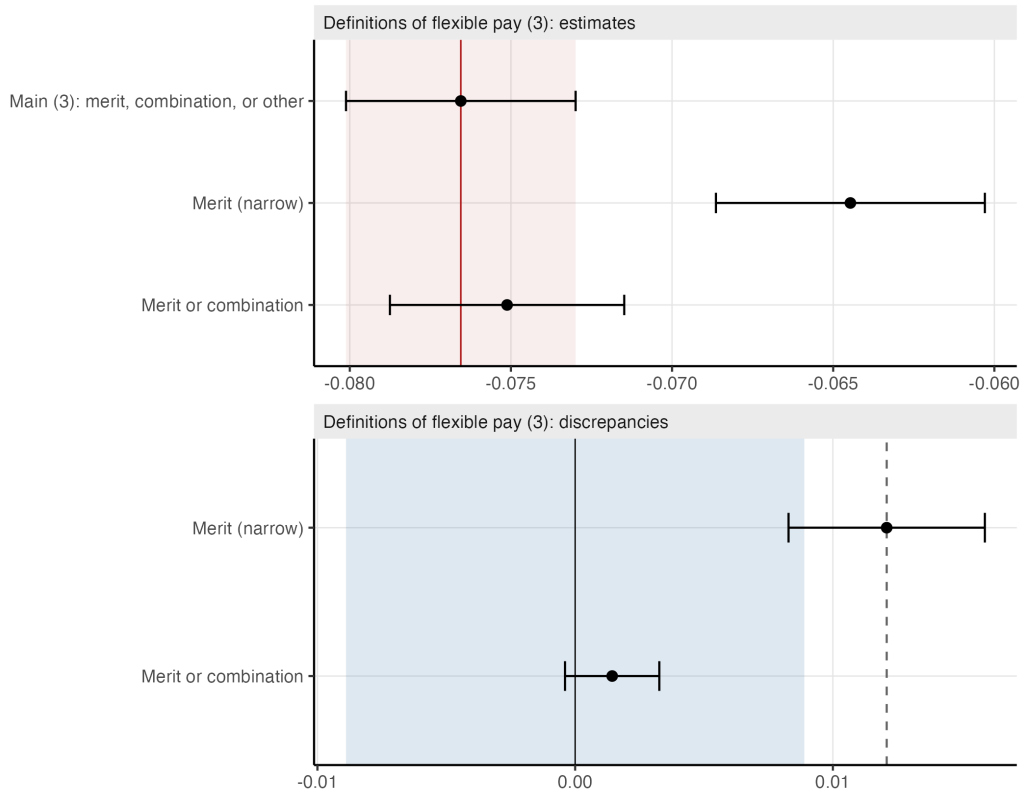


Figure 7: Alternative definitions of flexible pay, centered at column (3) of Table 2. Top: estimates with ±1.96 standard errors; the red line and band are the main estimate and its interval. Bottom: discrepancies $\hat{\Delta}_j$ with ±1.96 standard errors of the discrepancy; the solid line marks zero discrepancy, the shaded interval is $[-b_{RT}, b_{RT}]$ with $b_{RT} = 0.0089$, and the dashed line is the point comparison $\max_j |\hat{\Delta}_j| = 0.0121$.

4.3 The baseline specifications as a benchmark

The baseline specifications of Table 2 are not robustness checks. Columns (1) and (2) omit the fixed effects that the identification argument requires, so they are known to be wrong, and the discrepancies between them and the preferred specifications are differences in estimands by construction. Computing the tolerance across them therefore gives a reference point for the explicit checks: this is what a large tolerance looks like in the units of this application. Figure 8 centers the comparison at (4). The discrepancies centered at (4) give robustness tolerance of $b_{RT} = 0.132$ log points, and relative robustness tolerance $c_{RT} = 14$, so the data reject robustness with respect to magnitude (Definition 3) by a wide margin, while every estimate keeps its sign: the rejection comes entirely from checks of the same sign as the main effect, the case Section 2 warned the relative tolerance does not distinguish.

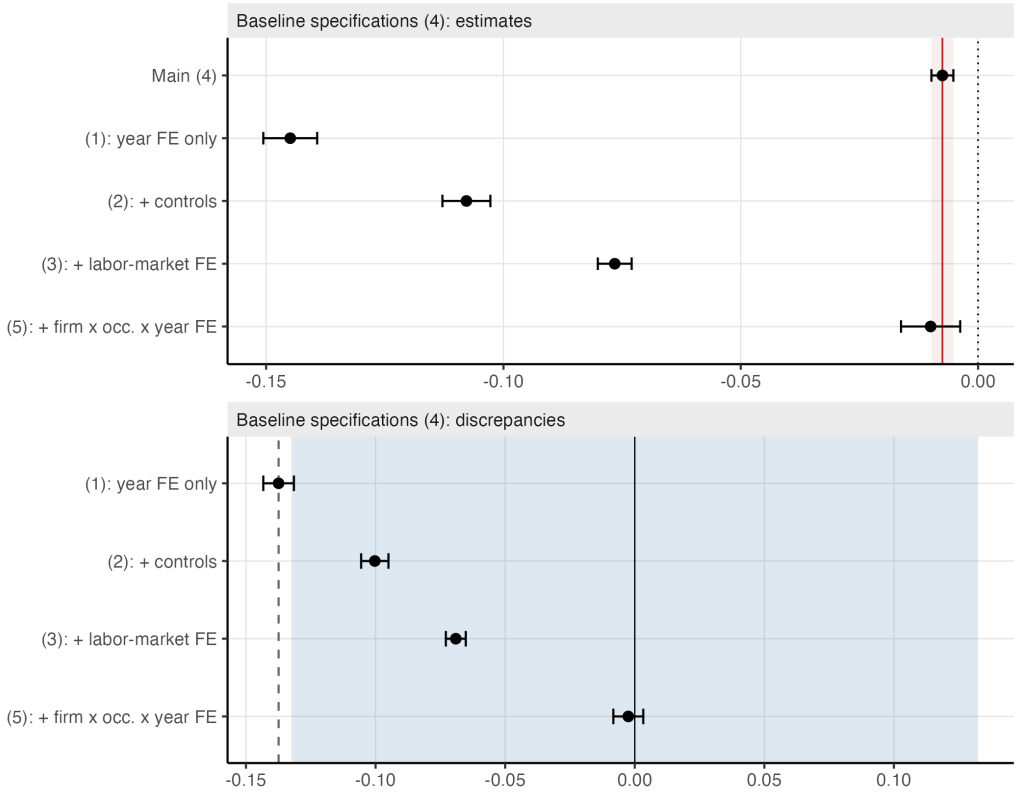


Figure 8: Baseline specifications, centered at column (4) of Table 2. Top: estimates with ±1.96 standard errors; the red line and band are the main estimate and its interval, and the dotted line marks zero. Bottom: discrepancies $\hat{\Delta}_j$ with ±1.96 standard errors of the discrepancy; the solid line marks zero discrepancy, the shaded interval is $[-b_{RT}, b_{RT}]$ with $b_{RT} = 0.132$, and the dashed line is the point comparison $\max_j |\hat{\Delta}_j| = 0.137$.

4.4 Reading the table

Table 1 reports, for each case, the robustness tolerance, the relative robustness tolerance, the point comparison, and the number of checks. Three features deserve comment.

First, the tolerance and the point comparison are different objects, and the table shows how different. The tolerance is below the point comparison in every case, as it must be: $b = \max_j |\hat{\Delta}_j|$ is never rejected, since the estimate itself satisfies the constraints. How far below depends on how precisely the discrepancies are measured, not on how large they are: the gap is a few percent of the point comparison in the baseline comparison, where the discrepancies are tens of standard errors from zero, and most of it in the weighting comparisons, where they are one to three standard errors from zero.

Second, two of the eight tolerances across explicit checks are zero and a third is zero for practical purposes. A zero means that the test does not reject $b = 0$: the discrepancies are

	b_{RT}	c_{RT}	$\max_j \lvert\hat{\Delta}_j\rvert$	m
Wt. (3)	0.0023	0.03	0.0066	4
Wt. (4)	0.0007	0.09	0.0060	4
Wt. (3) −5%	0.0012	0.02	0.0052	3
Wt. (4) −5%	0.0000	0.00	0.0030	3
Merit (3)	0.0089	0.12	0.0121	2
Merit (4)	0.0000	0.00	0.0012	2
All (3)	0.0087	0.12	0.0121	6
All (4)	0.0001	0.02	0.0060	6
Baseline (3)	0.0652	0.87	0.0690	4
Baseline (4)	0.1325	14.34	0.1374	4

Table 1: Robustness tolerance, relative tolerance, point comparison and number of checks for the regressions from tables A.2 and A.3 of Massenkoff and Wilmers (2023), centered at columns (3) and (4) of Table 2. "All" refers to specifications in both tables A.2 and A.3 together. −5% refers to excluding large jobs. "Baseline" refers to the five specifications of Table 2 with the named column as the main specification. Plug-in covariance clustered at the establishment level.

jointly within sampling error of zero, and the data are consistent with all the specifications sharing the same estimand. It does not mean that they do. A test of equality in the spirit of Lu and White (2014) would not reject in these cases either, and would say no more.

Third, the absolute and relative tolerances answer different questions, and the contrast between columns (3) and (4) shows why both are reported. For the weighting checks the absolute tolerances are 0.0023 and 0.0007 log points, similar in magnitude; relative to main effects that differ by a factor of ten, they are 3% and 9% of the effect. The absolute tolerance says how much the estimand moves across the checks in the units of the outcome; the relative tolerance says how much of the main effect that movement is. When the main effect is close to zero the ratio loses meaning and only the absolute tolerance remains informative. Every set of explicit checks is consistent with robustness with respect to magnitude, with c_{RT} at most 0.12 against the threshold of one; among the baseline specifications, centered at (3) the checks remain consistent with it, at $c_{RT} = 0.87$, and centered at (4) they reject it. The two baseline values bracket the threshold because the main effect at (4) is a tenth of that at (3) while the discrepancies from the omitted fixed effects are of the same order in both.

As a comparison, a Bonferroni lower bound, $\max_j(|\hat{\Delta}_j| - z_{1-\alpha/2m}\hat{\sigma}_{\Delta,j})_+$, which uses the standard errors of the discrepancies but not their covariance and pays for every check whether or not it binds, is below b_{RT} in every case but one: the exception is All (4), where the six discrepancies are all within a few standard errors of zero, every constraint is active

at the tolerance, and the joint test pays six degrees of freedom to detect one deviating check; the Bonferroni bound, at 0.0004 against 0.0001, pays only through its multiplier. This is the near-zero regime of Section 3.3, where the joint test is at its weakest, and it is the one place in the application where ignoring the covariance costs nothing.

The threshold of one in Definition 3 is a benchmark, and in many applications a stricter one might be appropriate; in this application c_{RT} is small for every set of explicit checks, so the data are consistent with robustness at thresholds well below one. Because the acceptance set of the relative test need not be an interval (Section 2.2), $c_{RT} \leq 1$ and non-rejection of $H_0^{\text{rel}}(1)$ by the RCC test itself can in principle differ; here they do not: the direct test of $H_0^{\text{rel}}(1)$ rejects for the baseline specifications centered at (4) and for no other case. The tolerance does not settle what the threshold should be, but it makes the debate about it transparent: whatever threshold a field settles on, the comparison is between that number and c_{RT}, and the data's contribution to it is separated from the judgment.

5 Extensions

This section presents two extensions. The first is a third normalization of the robustness tolerance, in standardized-effect units, which is specific to linear regressions. The second connects the robustness tolerance to formal sensitivity measures.

5.1 Standardized robustness tolerance

The linear-regression setting allows for the introduction of a second alternative measurement of robustness, which is especially useful to compare robustness across studies that have similar objects of interest and sets of robustness checks, but different measurement units (for example, y is income measured in dollars vs. euros). It standardizes robustness in standardized effect units:

Definition 4 (Standardized robustness tolerance)**.** Let

$$s_y = \text{sd}(y_{0i} \mid d_{0i} = 1), \qquad s_D = \text{sd}(D_{0i} \mid d_{0i} = 1),$$

and define the standardized estimands

$$\theta_j^s = \theta_j \frac{s_D}{s_y}, \qquad j = 0, \ldots, m.$$

For $\kappa \geq 0$, consider the null hypothesis

$$H_0^s(\kappa) : \max_{j=1,\ldots,m} |\theta_j^s - \theta_0^s| \leq \kappa. \tag{15}$$

Equivalently,

$$H_0^s(\kappa) : A\theta^s \leq \kappa \mathbf{1}_{2m}.$$

Let $\phi_n^s(\kappa, \alpha)$ denote a level-α test of $H_0^s(\kappa)$, with one denoting rejection. For significance level α, define the standardized robustness tolerance as

$$\kappa_{RT}(\alpha) := \inf \left\{ \kappa \geq 0 : \phi_n^s(\kappa, \alpha) = 0 \right\}. \tag{16}$$

The corresponding population discrepancy is

$$\max_{j=1,\ldots,m} |\theta_j^s - \theta_0^s| = \frac{s_D}{s_y} \max_{j=1,\ldots,m} |\theta_j - \theta_0|.$$

Thus, $\kappa_{RT}(\alpha)$ measures robustness in standardized-effect units, where coefficients are expressed in outcome standard deviations per treatment standard-deviation.

The relative and standardized robustness tolerances provide different normalizations. To illustrate, consider two studies with standardized main coefficients of 0.20 and 0.05, respectively. Suppose that, in both studies, a comparable set of robustness checks produces a maximum change of 0.02 when coefficients are expressed in standardized-effect units. The standardized population discrepancies are then equal, and the standardized tolerances similar if the two studies estimate them with comparable precision, reflecting that the robustness checks generate changes of similar magnitude relative to the variation in the underlying variables. The corresponding relative changes, however, are 10% and 40%. The relative robustness tolerance therefore regards the second study as less robust because the same standardized change is large relative to its smaller main effect.

Neither normalization is inherently preferable. The relative robustness tolerance uses the magnitude of the main estimand as the benchmark and is therefore informative about the importance of specification changes relative to the reported effect itself. The standardized robustness tolerance instead uses the variation in the outcome and treatment as the benchmark. This distinction can be particularly useful when comparing studies that investigate similar questions and implement similar robustness checks: two studies may display similar sensitivity in standardized-effect units even when that sensitivity represents

very different fractions of their respective main effects.

The test $\phi_n^s(\kappa, \alpha)$ is the RCC test of Section 2.2 applied to the standardized estimator $\hat{\theta}^s = (\hat{s}_D/\hat{s}_y)\hat{\theta}$, with a covariance estimator that also accounts for the estimation of the normalization s_D/s_y: the plug-in estimator built from the influence function of $\hat{\theta}^s$, given in Appendix A.7 together with Corollary 2, the analogue of Proposition 3 for the standardized tolerance.

5.2 Calibrating Sensitivity Parameters and the Scope of Robustness Checks

Formal sensitivity analysis for omitted variable bias (Altonji et al. (2005), Oster (2019), Cinelli and Hazlett (2020), Chernozhukov et al. (2026), Diegert et al. (2026a)) asks how strong selection on unobservables would have to be to change a conclusion, and leaves the researcher to decide what strength is plausible. Despite substantial advances in the interpretability of sensitivity parameters, that decision remains the hard part of the exercise, and the literature offers benchmarks for it, such as the importance of the included controls, rather than values. The sensitivity literature cautions that bias from observable specification choices does not bound bias from unobservables: an omitted variable may move the estimand by more than any perturbation the researcher can run. I turn that caution into a calibration requirement, which a researcher may or may not accept: that the bias the sensitivity model allows for unobservables be at least as large as the discrepancy the robustness checks establish, $b^{\text{UN}} \geq b(\theta)$. A researcher who imposes it, and who has a lower confidence bound $b_{RT}(\alpha)$ for $b(\theta)$, must then allow a sensitivity parameter large enough that the implied worst-case bias is at least $b_{RT}(\alpha)$. Because the implied bias is increasing in the parameter, this is a lower bound on the parameter. I develop the bound for the sensitivity parameter of Diegert et al. (2026a), which is scalar and whose interpretation is stable across specifications in the sense of Diegert, Masten, and Poirier (2026b).[15]

Consider the main specification $j = 0$. I follow their notation, except that I use ν instead of r to avoid confusion with the degrees of freedom in (11), and rename coefficients

[15] Diegert et al. (2026b) show that the parameter used here, together with the parameter of Cinelli and Hazlett (2020), satisfies two axioms, consistency and monotonicity, under which a sensitivity parameter has the same meaning as the set of included controls changes, whereas the parameter of Oster (2019) does not. Mapping the robustness tolerance into a parameter that fails these axioms would inherit their pathologies.

with the superscripts "med" and "long" for the medium and long regressions, respectively. The main specification is the medium regression, $\theta_0^{\text{med}} = \theta_0$, but the parameter of interest is now θ_0^{long}, which is only partially identified. Let X_0 denote the calibration covariates: following Diegert et al. (2026a, Section 3.4), the included controls relative to which selection on unobservables is measured. Controls that are partialled out of every variable but not used for calibration, such as fixed effects, are absorbed in the notation, all variables being residuals from the projection on them. All quantities below are computed on the main specification's sample, with its weights, so that $\theta_0^{\text{med}} = \theta_0$. The medium regression is

$$y_0 = \alpha_0^{\text{med}} + \theta_0^{\text{med}} D_0 + (\beta_0^{\text{med}})' X_0 + e_0^{\text{med}}. \tag{17}$$

Let W be a vector of unobserved variables of dimension p_W, such that the researcher is interested in θ_0^{long} in the regression

$$y_0 = \alpha_0^{\text{long}} + \theta_0^{\text{long}} D_0 + (\beta_0^{\text{long}})' X_0 + \gamma' W + e_0^{\text{long}}. \tag{18}$$

Next, consider the regressions of D_0 on X_0, and of D_0 on X_0 and W:

$$D_0 = \alpha_{D_0}^{\text{med}} + (\xi^{\text{med}})' X_0 + e_{D_0}^{\text{med}}, \tag{19}$$

$$D_0 = \alpha_{D_0}^{\text{long}} + (\xi^{\text{long}})' X_0 + \pi' W + e_{D_0}^{\text{long}}, \tag{20}$$

where $\alpha_{D_0}^{\text{med}}$, $\alpha_{D_0}^{\text{long}}$ are constants and $e_{D_0}^{\text{med}}$, $e_{D_0}^{\text{long}}$ are the projection residuals. Diegert et al. (2026a) call $\sqrt{\text{var}(\pi' W)}$ selection on unobservables and $\sqrt{\text{var}((\xi^{\text{long}})' X_0)}$ selection on observables; since "observables" refers to robustness checks in this paper, I call the latter selection on included covariates. Their sensitivity parameter bounds the ratio of the two,

$$\frac{\sqrt{\text{var}(\pi' W)}}{\sqrt{\text{var}((\xi^{\text{long}})' X_0)}} \leq \bar{\nu}_D,$$

and a value of $\bar{\nu}_D$ is what the researcher must choose: $\bar{\nu}_D = 1$ allows unobservables to matter as much for selection into treatment as all included controls together, $\bar{\nu}_D = 0.5$ half as much. By Theorem 1 of Diegert et al. (2026a), under this restriction and with the controls allowed to be arbitrarily endogenous, $\theta_0^{\text{long}} \in [\theta_0^{\text{med}} - b^{\text{UN}}(\bar{\nu}_D),\ \theta_0^{\text{med}} + b^{\text{UN}}(\bar{\nu}_D)]$, where

$$b^{\text{UN}}(\bar{\nu}_D) = \begin{cases} \sqrt{\frac{\text{var}\left(e_0^{\text{med}}\right)}{\text{var}\left(e_{D_0}^{\text{med}}\right)} \cdot \frac{\bar{\nu}_D^2 R_{D\sim X_0}^2}{1 - R_{D\sim X_0}^2 - \bar{\nu}_D^2}}, & \text{if } \bar{\nu}_D^2 < 1 - R_{D\sim X_0}^2, \\ \infty, & \text{otherwise,} \end{cases} \tag{21}$$

with $R^2_{D\sim X_0}$ the R^2 of the regression in (19), a within R^2 when controls have been partialled out. The half-width b^{UN} is the largest bias from unobservables the model permits at $\bar{\nu}_D$; it is increasing in $\bar{\nu}_D$, zero at $\bar{\nu}_D = 0$, and infinite once $\bar{\nu}_D^2$ reaches $1 - R^2_{D\sim X_0}$.

When $0 < R^2_{D\sim X_0} < 1$ and both residual variances are positive and finite, the conditions of Theorem 1 of Diegert et al. (2026a), b^{UN} is continuous and strictly increasing on $[0, (1 - R^2_{D\sim X_0})^{1/2})$ with range $[0, \infty)$, so for any distance b there is a unique $\bar{\nu}_D$ at which the largest bias the model permits equals b. Evaluating that inverse at the robustness tolerance gives the following.

Definition 5 (Scope of the robustness checks)**.** The scope of a set of robustness checks, in the units of Diegert et al. (2026a), is

$$\hat{\bar{\nu}}_D(\alpha) := \sqrt{\frac{\left(b_{RT}(\alpha)^2 \cdot \frac{\mathrm{var}\left(e^{\mathrm{med}}_{D_0}\right)}{\mathrm{var}\left(e^{\mathrm{med}}_{0}\right)}\right)\left(1 - R^2_{D\sim X_0}\right)}{R^2_{D\sim X_0} + b_{RT}(\alpha)^2 \cdot \frac{\mathrm{var}\left(e^{\mathrm{med}}_{D_0}\right)}{\mathrm{var}\left(e^{\mathrm{med}}_{0}\right)}}}, \tag{22}$$

the value of $\bar{\nu}_D$ that solves $b^{\mathrm{UN}}(\bar{\nu}_D) = b_{RT}(\alpha)$.

In words, $\hat{\bar{\nu}}_D(\alpha)$ is the smallest sensitivity allowance, selection on unobservables relative to selection on included covariates, at which the worst-case bias the model permits is as large as the discrepancy the robustness checks establish. It is not a value at which an unobservable moves the estimand by that amount; the model bounds the bias, it does not identify it. It is a measure of how much of the specification space the robustness exercise explored, expressed on the scale the sensitivity literature uses to describe what cannot be explored. A set of checks that moves the estimand little, over a narrow set of perturbations, has small scope; this is the counterpart of the discussion in Section 2.2, where the tolerance was shown to be a statement about the set of checks actually run. The population quantities $\mathrm{var}(e^{\mathrm{med}}_{D_0})/\mathrm{var}(e^{\mathrm{med}}_{0})$ and $R^2_{D\sim X_0}$ are replaced by their sample counterparts from the main specification. Since $b_{RT}(\alpha)$ is a lower confidence bound for the maximal discrepancy and the two population quantities are plugged in, $\hat{\bar{\nu}}_D(\alpha)$ does not itself carry a coverage statement; it is a calibrated value, not a confidence bound.

Its statistical content is this. The map $b \mapsto \hat{\bar{\nu}}_D$ is increasing, so on the event $b_{RT}(\alpha) \leq b(\theta)$, of probability at least $1-\alpha$, $\hat{\bar{\nu}}_D(\alpha)$ is at most the allowance that the same calculation, with the sample values of $\mathrm{var}(e^{\mathrm{med}}_{D_0})/\mathrm{var}(e^{\mathrm{med}}_{0})$ and $R^2_{D\sim X_0}$, assigns to the true discrepancy. The sampling variation of those two quantities is not accounted for, so $\hat{\bar{\nu}}_D(\alpha)$ is a calibrated

value and not a confidence bound; a bound would require joint inference on the discrepancy and the two nuisance quantities, which I do not pursue. A researcher who has a value of $\bar{\nu}_D$ in mind can compare it with $\hat{\bar{\nu}}_D(\alpha)$: a smaller value means the sensitivity bounds are narrower than the range the researcher's own robustness checks span, which is inconsistent with the calibration requirement. Either way the comparison replaces a judgment about an unobservable with a judgment about perturbations the researcher has already run.

In the application of Section 4, centering at specification (3) and using all explicit robustness checks, $b_{RT} = 0.0087$ gives $\hat{\bar{\nu}}_D = 0.060$: the sensitivity allowance implied by the robustness checks is an unobservable about six percent as important for selection into nonstandardized pay as the eight controls of Table 2 together, the fixed effects being partialled out. Centering at specification (4), where the tolerance is zero for practical purposes, so is $\hat{\bar{\nu}}_D$. The map from tolerance to sensitivity depends on $R^2_{D\sim X_0}$, which is 0.091 in specification (3) and 0.007 in specification (4), where the fixed effects absorb almost all of the variation in treatment: the same discrepancy in the estimand corresponds to a far stronger unobservable, relative to the included controls, when the included controls explain little. The weighting checks show this: their tolerances are 0.0023 at (3) and 0.0007 at (4), yet the implied $\hat{\bar{\nu}}_D$ are 0.016 and 0.023, in the opposite order.

A remark on what the bound does and does not claim: $\hat{\bar{\nu}}_D(\alpha)$ is a lower bound, not a recommended value. It says that any sensitivity parameter smaller than $\hat{\bar{\nu}}_D(\alpha)$ is inconsistent with the researcher's own robustness checks under the calibration requirement; it does not say that $\hat{\bar{\nu}}_D(\alpha)$ itself is enough. How much larger than the bound the parameter should be is exactly the judgment the sensitivity literature leaves to the researcher, and nothing in the robustness checks informs it. The exercise is performed here for one scalar sensitivity parameter; it applies in the same way to any sensitivity parameter whose implied bias is increasing in the parameter, with the interpretation of the bound specific to each.

6 Conclusion

This paper introduces the robustness tolerance as a formal measure of what a set of robustness checks establishes. Rather than testing whether the estimands of different specifications are equal, it asks how far apart the data show them to be: the tolerance is the smallest threshold at which the hypothesis that every check lies within that distance of the main specification is not rejected, and it is a lower confidence bound for the maximal

discrepancy. Treating robustness as the null is what makes the measure safe for the author: a check whose estimand lies inside the interval spanned by the other checks leaves the true maximal discrepancy unchanged, and the measure exceeds that discrepancy with probability at most α however noisy the check is, whereas any measure that takes fragility as the null exceeds it with probability at least $1 - \alpha$, and is driven to infinity by a noisy check. At the same time a check whose estimand lies outside the interval moves the tolerance as soon as its departure is large relative to the standard error of its own discrepancy, which is what a referee asking for the check is entitled to.

The test sees only the estimated discrepancies and their covariance, so the behavior of the tolerance is a matter of how many standard errors of a discrepancy the true discrepancies are worth, and of the configuration and correlation of the checks. n the designs simulated, with up to nine checks, the tolerance far from zero sits within about two standard errors of the truth whatever their number, because the test gives little weight to the checks that do not bind. Near zero, configuration and correlation decide how often the tolerance is zero: a joint test of equalities has little power against one deviating check among many uncorrelated ones, and little power against correlated checks that all move the same way. A zero tolerance is therefore a statement that the data do not reject equality, not that equality holds. Adding a check whose discrepancy is inside the interval costs nothing when its discrepancy is precisely measured and and typically lowers the tolerance when it is not.

In the application to Massenkoff and Wilmers (2023), the tolerances across the authors' explicit robustness checks areall consistent with robustness with respect to magnitude, against a reference of fourteen times the main effect across the baseline specifications that omit the fixed effects the identification argument requires. The check the authors single out as an exception moves the point estimate by 79% of the main effect; the relative tolerance is 9%, because the discrepancy is less than three standard errors from zero. Two of the eight tolerances across explicit checks are zero, in cases where the discrepancies are jointly within sampling error of zero. In every case the tolerance is below the point comparison, by an amount that depends on how precisely the discrepancies are measured rather than on how large they are.

The relative and standardized tolerances are alternative normalizations of the same quantity: relative to the magnitude of the main estimand, which is informative when that magnitude is well away from zero and uninformative when it is not, and relative to the variation in the outcome and treatment, which allows comparison across studies with

different units. Which to read depends on the question asked of the robustness checks, and the application shows that the absolute and relative tolerances can order the same cases differently.

Neither tolerance says what counts as robust. The threshold of one for the relative tolerance is a benchmark, and stricter thresholds will often be appropriate; what they should be is a question for each field. The tolerance makes that question answerable: it separates what the data establish about the discrepancies from the judgment of how large a discrepancy a result can bear, and turns the assessment of robustness into a comparison of two numbers, one from the data and one from the field.

Finally, the robustness tolerance is not a measure of omitted variable bias. It can, however, calibrate one: if the researcher requires the sensitivity model to allow at least as much bias as the robustness checks establish, the tolerance gives a lower limit for the sensitivity parameter, and so expresses the scope of a robustness exercise on the scale the sensitivity literature uses for what cannot be explored. In the application that scope is an unobservable about six percent as important for selection as the included controls together. Such comparisons do not remove the need for judgment about which checks or which sensitivity parameters are appropriate, but they replace a judgment about an unobservable with one about perturbations the researcher has already run.

A Proofs

A.1 Asymptotically Linear Estimators

Assumption 2. Let $\theta(F) = (\theta_0(F), \ldots, \theta_m(F))'$. Suppose that $\hat{\theta}$ admits the representation

$$\hat{\theta} - \theta(F) = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i, F) + R_n(F),$$

where $W_1, \ldots, W_n$ are i.i.d. under F, $\psi(W_i, F)$ satisfies $E_F[\psi(W_i, F)] = 0$, and $R_n(F)$ is the remainder in the linear representation. Define

$$\Sigma(F) = E_F[\psi(W_i, F)\psi(W_i, F)'], \qquad V(F) = \operatorname{diag}(\Sigma(F)),$$

and

$$\Omega(F) = V(F)^{-1/2}\Sigma(F)V(F)^{-1/2}.$$

There exist constants $\epsilon > 0$, $\delta > 0$, and $K < \infty$ such that the following conditions hold uniformly over $F \in \mathcal{F}$:

a) The remainder is uniformly negligible after standardization: for every $\eta > 0$,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} P_F\left(\sqrt{n}\left\|V(F)^{-1/2}R_n(F)\right\|_2 > \eta\right) = 0.$$

b) For every $j = 0, \ldots, m$,

$$\operatorname{Var}_F(\psi_j(W_i, F)) > 0,$$

$$E_F\left[\left|\frac{\psi_j(W_i, F)}{\sqrt{\operatorname{Var}_F(\psi_j(W_i, F))}}\right|^{2+\delta}\right] \leq K,$$

and

$$\det(\Omega(F)) \geq \epsilon.$$

c) The covariance estimator $\hat{\Sigma}$ satisfies, for every $\eta > 0$,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} P_F\left(\left\|V(F)^{-1/2}\hat{\Sigma}V(F)^{-1/2} - \Omega(F)\right\| > \eta\right) = 0.$$

Proposition 4 below is stated for a general family of polyhedral null hypotheses, indexed

by a tested value t, so that it can be applied to the robustness tolerance, to each branch of the relative robustness tolerance, and to the standardized robustness tolerance without repeating the argument. Let $\mathcal{T}$ be an index set and, for each $t \in \mathcal{T}$, let $A(t)$ be a deterministic $d_A \times (m+1)$ matrix and $b(t) \in \mathbb{R}^{d_A}$ a deterministic vector, where d_A does not depend on t. For $t \in \mathcal{T}$ consider the null hypothesis

$$H_0(t) : A(t)\theta(F) \leq b(t),$$

and let

$$\mathcal{T}_0(F) = \{t \in \mathcal{T} : A(t)\theta(F) \leq b(t)\}$$

denote the set of tested values for which $H_0(t)$ holds under F. Let $\phi_n^{\mathrm{RCC}}(t, \alpha)$ denote the RCC test of $H_0(t)$ constructed from $\widehat{\theta}$, $\widehat{\Sigma}$, $A(t)$ and $b(t)$; that is, the test defined by (8)–(11) with $A(t)$ and $b(t)$ in place of A and $b\mathbf{1}_{2m}$. For $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$, let

$$J(F, t) = \{\ell \in \{1, \ldots, d_A\} : e_\ell' A(t)\theta(F) = e_\ell' b(t)\}$$

denote the set of inequalities that bind at $\theta(F)$. The robustness tolerance corresponds to $\mathcal{T} = [0, \infty)$, $A(t) \equiv A$ and $b(t) = t\mathbf{1}_{2m}$.

Proposition 4. Suppose Assumption 2 holds, and let $\alpha \in (0, 1/2)$.

(a) The RCC test is asymptotically uniformly valid over $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$:

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{t\in\mathcal{T}_0(F)} E_F\left[\phi_n^{\mathrm{RCC}}(t, \alpha)\right] \leq \alpha.$$

No condition on the family $\{(A(t), b(t)) : t \in \mathcal{T}\}$ is required.

(b) Fix $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$. Suppose that $A(t)$ has no zero rows and that $J(F, t)$ is nonempty. Then

$$\lim_{n\to\infty} E_F\left[\phi_n^{\mathrm{RCC}}(t, \alpha)\right] = \alpha.$$

Consequently, if there exist $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$ satisfying the conditions of part (b), the RCC test has uniformly asymptotically exact size:

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{t\in\mathcal{T}_0(F)} E_F\left[\phi_n^{\mathrm{RCC}}(t, \alpha)\right] = \alpha.$$

Proof. Throughout, we use the notation of Cox and Shi (2023, Supplementary Appendix B.2), and we first record how the objects of this paper map into theirs. In their framework

the tested parameter is t, with parameter space $\mathcal{T}$. The moment vector is $\bar{m}_n \equiv \widehat{\theta}$; the coefficient matrix and intercept are $A(t)$ and $b(t)$ (Cox and Shi (2023) allow both to depend on the tested parameter; see footnote 5 of their main text); and the covariance estimator is $\widehat{\Sigma}$. With these identifications, $T_n(t)$, $\widehat{\mathcal{J}}(t)$ and $\hat{r}(t)$ are equations (17) and (19) and the quantity $\hat{r}$ of Cox and Shi (2023), and $\phi_n^{\mathrm{RCC}}(t,\alpha)$ is the test defined in equation (39) of their Supplementary Appendix, with $\hat{\beta} \in [\alpha, 2\alpha]$ as in their equation (38); the restriction $\alpha < 1/2$ guarantees $\hat{\beta} < 1$.

One point of the mapping requires care. Because their moments are sample means, Cox and Shi (2023) write the centering vector in their Assumption 3(a) and the null set $\Theta_0(F) = \{\theta : A(\theta)E_F\bar{m}_n(\theta) \leq b(\theta)\}$ in terms of $E_F\bar{m}_n$. Here $\widehat{\theta}$ is not a sample mean, and $E_F[\widehat{\theta}]$ neither equals $\theta(F)$ nor is guaranteed to exist. This is immaterial for the application of their Theorem 5. An inspection of its proof (Cox and Shi, 2023, Supplementary Appendix D.3) shows that $E_{F_n}\bar{m}_n(\theta_n)$ enters only (i) as the centering vector in their displays (130)–(131), which restate conditions (53)–(54) of Assumption 3(a), and (ii) through the standardized slack vector $g_q = \sqrt{n_q}\,\Lambda_q\big(b(\theta_{n_q}) - A(\theta_{n_q})E_{F_{n_q}}\bar{m}_{n_q}(\theta_{n_q})\big)$, that is, through the requirement $\theta_n \in \Theta_0(F_n)$ in part (a) and the rate conditions in parts (b) and (c). No property of $E_F\bar{m}_n$ as an expectation is used. The theorem therefore holds verbatim with $E_F\bar{m}_n$ replaced by any deterministic vector $\mu(F)$ for which (53) holds, and with the null set read as $\{\theta : A(\theta)\mu(F) \leq b(\theta)\}$. We apply it with $\mu(F) = \theta(F)$, the centering in Assumption 2a); accordingly, their null set $\Theta_0(F)$ is our $\mathcal{T}_0(F)$, and all references to Assumption 3 and Theorem 5 of Cox and Shi (2023) below are to be read with $\theta(F)$ in place of $E_F\bar{m}_n$.

Two further features of the mapping are used repeatedly: neither $\widehat{\theta}$ nor $\widehat{\Sigma}$ depends on t, and the supremum over $\theta \in \Theta_0(F)$ in Theorem 5(a) of Cox and Shi (2023) is exactly the supremum over $t \in \mathcal{T}_0(F)$ in part (a) of the proposition.

Preliminaries. Define

$$Z_i(F) = V(F)^{-1/2}\psi(W_i, F).$$

By construction,

$$E_F[Z_i(F)] = 0, \qquad E_F[Z_i(F)Z_i(F)'] = \Omega(F),$$

and $\Omega(F)$ is a correlation matrix. Assumption 2b) implies that, for every $j = 0, \ldots, m$,

$$\sup_{F\in\mathcal{F}} E_F|Z_{ji}(F)|^{2+\delta} \leq K,$$

and, since m is fixed,

$$\sup_{F \in \mathcal{F}} E_F \|Z_i(F)\|_2^{2+\delta} < \infty.$$

Moreover, $V(F)$ is positive definite for every $F \in \mathcal{F}$, since its diagonal entries are $\mathrm{Var}_F(\psi_j(W_i, F)) > 0$ by Assumption 2b). Assumption 2a) gives

$$\sqrt{n}\, V(F)^{-1/2} \left(\widehat{\theta} - \theta(F)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(F) + o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$, and Assumption 2c) gives

$$V(F)^{-1/2} \widehat{\Sigma}\, V(F)^{-1/2} = \Omega(F) + o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$.

Part (a). By Theorem 5(a) of Cox and Shi (2023), it suffices to verify their Assumption 3(a) along every sequence $\{(F_n, t_n)\}_{n \geq 1}$ with $F_n \in \mathcal{F}$ and $t_n \in \mathcal{T}_0(F_n)$; by Remark (1) following their Theorem 5, no condition on $A(t_n)$ or $b(t_n)$ is needed. Fix such a sequence. Because $\widehat{\theta}$ and $\widehat{\Sigma}$ do not depend on t_n, only the sequence $\{F_n\}$ is relevant. Let $\{n_m\}$ be an arbitrary subsequence. Since $\Omega(F_{n_m})$ is a correlation matrix of fixed dimension, it ranges over a compact set, so there exists a further subsequence $\{n_q\}$ along which

$$\Omega(F_{n_q}) \longrightarrow \Omega_\infty.$$

By continuity of the determinant and Assumption 2b), $\det(\Omega_\infty) \geq \epsilon > 0$, so Ω_∞ is a positive definite correlation matrix. Take

$$D_q = V(F_{n_q}),$$

which is positive definite as noted above.

We verify condition (53) of Cox and Shi (2023). Fix $a \in \mathbb{R}^{m+1}$ with $\|a\|_2 = 1$. The triangular array $\{a' Z_i(F_{n_q}) : i \leq n_q\}$ has i.i.d. rows with mean zero, variance $a'\Omega(F_{n_q})a \to a'\Omega_\infty a > 0$, and $\sup_q E_{F_{n_q}} |a' Z_i(F_{n_q})|^{2+\delta} < \infty$. Hence Lyapunov's condition holds:

$$\frac{n_q\, E_{F_{n_q}} |a' Z_i(F_{n_q})|^{2+\delta}}{\left\{n_q\, a'\Omega(F_{n_q})a\right\}^{1+\delta/2}} = O\big(n_q^{-\delta/2}\big) \longrightarrow 0,$$

and the Lyapunov central limit theorem together with the Cramér–Wold device yields

$$\frac{1}{\sqrt{n_q}} \sum_{i=1}^{n_q} Z_i(F_{n_q}) \xrightarrow{d} N(0, \Omega_\infty).$$

Combining this with the uniform remainder condition (Assumption 2a)),

$$\sqrt{n_q}\, D_q^{-1/2} \left(\widehat{\theta} - \theta(F_{n_q}) \right) \xrightarrow{d} N(0, \Omega_\infty),$$

which is condition (53). For condition (54), Assumption 2c) and $\Omega(F_{n_q}) \to \Omega_\infty$ give

$$\left\| D_q^{-1/2} \widehat{\Sigma} D_q^{-1/2} - \Omega_\infty \right\| \leq \left\| V(F_{n_q})^{-1/2} \widehat{\Sigma} V(F_{n_q})^{-1/2} - \Omega(F_{n_q}) \right\| + \left\| \Omega(F_{n_q}) - \Omega_\infty \right\| \xrightarrow{p} 0.$$

Thus Assumption 3(a) of Cox and Shi (2023) holds along every sequence $\{(F_n, t_n)\}$ with $t_n \in \mathcal{T}_0(F_n)$, and Theorem 5(a) of Cox and Shi (2023) gives

$$\limsup_{n \to \infty} \sup_{F \in \mathcal{F}} \sup_{t \in \mathcal{T}_0(F)} E_F \left[\phi_n^{\mathrm{RCC}}(t, \alpha) \right] \leq \alpha.$$

This establishes part (a).

Part (b). Fix $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$ as in the statement, and write $J = J(F, t)$, $A = A(t)$ and $b = b(t)$; these are fixed throughout the proof of part (b). We apply parts (b) and (c) of Theorem 5 of Cox and Shi (2023) along the constant sequence $(F_n, t_n) = (F, t)$; this is precisely the situation described in Remark (2) following their Theorem 5, in which the two parts combine to deliver exact asymptotic size. We first verify their Assumption 3(a,b) along this sequence.

Assumption 3(a) holds by the argument of part (a) applied to the constant sequence $F_n \equiv F$, with $D_q = V(F)$ for every q and $\Omega_\infty = \Omega(F)$.

For Assumption 3(b), recall that Λ_q is the diagonal matrix whose ℓth entry is $\|e_\ell' A D_q\|_2^{-1}$ when $e_\ell' A \neq 0$ and one otherwise. By assumption no row of A is zero, and A and $D_q = V(F)$ do not vary with q, so $\Lambda_q = \Lambda(F)$ is a fixed positive definite diagonal matrix with ℓth entry $\|e_\ell' A V(F)\|_2^{-1}$. Hence

$$\Lambda_q A D_q = \Lambda(F) A V(F) =: \bar{A}_0$$

for every q, and, for every $L \subseteq \{1, \ldots, d_A\}$,

$$\operatorname{rk}(I_L A D_q) = \operatorname{rk}(I_L A) = \operatorname{rk}(I_L \bar{A}_0),$$

because $V(F)$ and $\Lambda(F)$ are nonsingular. Thus Assumption 3(b) holds for the full system (A, b). The same argument applied to the reduced system (A_J, b_J), where b_J is the subvector of b with indices in J, shows that Assumption 3(b) holds for it as well, with limit $\bar{A}_{0,J} = \Lambda_J(F) A_J V(F) = I_J \bar{A}_0$, which is nonzero because J is nonempty and every row of A is nonzero.

Irrelevance of the slack inequalities. For every $\ell \notin J$, the slack $e_\ell'(A\theta(F) - b) < 0$ is a fixed negative number, and the ℓth diagonal entry of $\Lambda(F)$ is a fixed positive number, so

$$\sqrt{n}\, e_\ell' \Lambda(F)\, (A\theta(F) - b) \longrightarrow -\infty.$$

Theorem 5(c) of Cox and Shi (2023), applied with this J, therefore gives

$$\lim_{n\to\infty} P_F \left(\phi_n^{\mathrm{RCC}}(t, \alpha) \neq \phi_{n,J}^{\mathrm{RCC}}(t, \alpha) \right) = 0,$$

where $\phi_{n,J}^{\mathrm{RCC}}$ denotes the RCC test constructed from (A_J, b_J) in place of (A, b) (Cox and Shi, 2023, Supplementary Appendix B.1).

Exactness for the binding inequalities. For the reduced system, every inequality binds: $A_J \theta(F) = b_J$. Hence, for every $\ell \in J$,

$$\sqrt{n}\, e_\ell' \Lambda_J(F)\, (A_J \theta(F) - b_J) = 0 \longrightarrow 0,$$

and $\bar{A}_{0,J} \neq 0$. Theorem 5(b) of Cox and Shi (2023) therefore gives

$$\lim_{n\to\infty} E_F \left[\phi_{n,J}^{\mathrm{RCC}}(t, \alpha) \right] = \alpha.$$

Combining the two displays,

$$\lim_{n\to\infty} E_F \left[\phi_n^{\mathrm{RCC}}(t, \alpha) \right] = \alpha,$$

which establishes part (b).

We note that A_J may be rank deficient; for instance, in the robustness tolerance at $t = b_F = 0$ all $2m$ inequalities bind and $\operatorname{rk}(A_J) = \operatorname{rk}(A) = m < 2m$, because each

inequality is paired with its exact opposite. Theorem 5(b) of Cox and Shi (2023) requires only $\bar{A}_{0,J} \neq 0$ and the rank-stability condition, both of which hold, so this case is covered. (When J contains a row and its exact opposite, $\hat{\tau} = 0$ and the RCC test coincides with the CC test; see the discussion preceding equation (38) in the Supplementary Appendix of Cox and Shi (2023).)

Exact asymptotic size. Let $F \in \mathcal{F}$ and $t \in \mathcal{T}_0(F)$ satisfy the conditions of part (b). Then part (b) implies

$$\liminf_{n\to\infty} \sup_{F'\in\mathcal{F}} \sup_{t'\in\mathcal{T}_0(F')} E_{F'}\left[\phi_n^{\mathrm{RCC}}(t',\alpha)\right] \geq \lim_{n\to\infty} E_F\left[\phi_n^{\mathrm{RCC}}(t,\alpha)\right] = \alpha.$$

Combining this lower bound with part (a) yields

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{t\in\mathcal{T}_0(F)} E_F\left[\phi_n^{\mathrm{RCC}}(t,\alpha)\right] = \alpha,$$

which is the notion of exact asymptotic size in Remark (2) following Theorem 5 of Cox and Shi (2023). □

A.2 Proof of Proposition 3

Proof. We verify the conditions of Assumption 2. Fix $j \in \{0,\ldots,m\}$ and suppress the j subscript where there is no ambiguity. Let

$$q_j = E_F[d_{ji}u_{ji}^2].$$

By the Frisch–Waugh–Lovell theorem,

$$\widehat{\theta}_j - \theta_j = \frac{a_n - b_n' C_n^{-1} c_n}{s_n - b_n' C_n^{-1} b_n},$$

where

$$a_n = E_n[du\varepsilon], \qquad b_n = E_n[dZu], \qquad c_n = E_n[dZ\varepsilon],$$

and

$$C_n = E_n[dZZ'], \qquad s_n = E_n[du^2].$$

By construction,

$$E_F[a_n] = 0, \qquad E_F[b_n] = 0, \qquad E_F[c_n] = 0,$$

while

$$E_F[C_n] = Q_Z, \qquad E_F[s_n] = q_j.$$

Assumption (c) implies uniformly over $F \in \mathcal{F}$ that

$$b_n = O_{P_F}(n^{-1/2}), \qquad c_n = O_{P_F}(n^{-1/2}),$$

and

$$C_n - Q_Z = O_{P_F}(n^{-1/2}), \qquad s_n - q_j = O_{P_F}(n^{-1/2}).$$

For b_n, by Cauchy–Schwarz,

$$E_F\|dZu\|_2^2 \le \left(E_F\|dZ\|_2^4\right)^{1/2} \left(E_F[d|u|^4]\right)^{1/2} \le K.$$

The corresponding second-moment bounds for $dZ\varepsilon$, dZZ', and du^2 follow directly from Assumption (c). Since the relevant dimensions are fixed, the displayed stochastic orders follow uniformly over $F \in \mathcal{F}$.

By Assumption (b),

$$\lambda_{\min}(Q_Z) \ge \epsilon, \qquad q_j \ge \epsilon.$$

It follows that

$$C_n^{-1} = O_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$. Define

$$D_n = s_n - b_n' C_n^{-1} b_n.$$

Then

$$D_n = q_j + O_{P_F}(n^{-1/2}),$$

and D_n is bounded away from zero with probability approaching one uniformly over $F \in \mathcal{F}$. Moreover,

$$b_n' C_n^{-1} c_n = O_{P_F}(n^{-1}).$$

Define, for the purposes of this proof,

$$\psi_{ji}(F) = \frac{d_{ji} u_{ji} \varepsilon_{ji}}{q_j}.$$

Then

$$\widehat{\theta}_j - \theta_j = \frac{1}{q_j} E_n[d_{ji} u_{ji} \varepsilon_{ji}] + R_{jn} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ji}(F) + R_{jn},$$

where

$$R_{jn} = a_n \left(\frac{1}{D_n} - \frac{1}{q_j} \right) - \frac{b_n' C_n^{-1} c_n}{D_n}.$$

By population OLS orthogonality,

$$E_F[\psi_{ji}(F)] = 0,$$

and, by the definition of σ_j^2,

$$\mathrm{Var}_F(\psi_{ji}(F)) = \sigma_j^2.$$

Therefore,

$$\frac{\sqrt{n}\, a_n}{q_j \sigma_j} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\psi_{ji}(F)}{\sigma_j} = O_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$. Hence

$$\begin{aligned} \frac{\sqrt{n}}{\sigma_j} a_n \left(\frac{1}{D_n} - \frac{1}{q_j} \right) &= \frac{\sqrt{n}\, a_n}{q_j \sigma_j} \frac{q_j - D_n}{D_n} \\ &= o_{P_F}(1) \end{aligned}$$

uniformly over $F \in \mathcal{F}$.

For the second term, Assumption (d) gives $\sigma_j^2 \geq \epsilon$. Therefore,

$$\frac{\sqrt{n}}{\sigma_j} \frac{b_n' C_n^{-1} c_n}{D_n} = O_{P_F}(n^{-1/2}) = o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$. Thus

$$\frac{\sqrt{n}\, R_{jn}}{\sigma_j} = o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$.

Since m is fixed, stacking across $j = 0, \ldots, m$ gives

$$\widehat{\theta} - \theta(F) = \frac{1}{n} \sum_{i=1}^{n} \psi(W_i, F) + R_n(F),$$

where

$$\sqrt{n}\left\|V(F)^{-1/2}R_n(F)\right\|_2 = o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$. This establishes Assumption 2a).

Next, Assumption (d) and $\mathrm{Var}_F(\psi_{ji}(F)) = \sigma_j^2$ imply

$$\mathrm{Var}_F(\psi_{ji}(F)) > 0$$

and

$$E_F\left[\left|\frac{\psi_{ji}(F)}{\sqrt{\mathrm{Var}_F(\psi_{ji}(F))}}\right|^{2+\delta}\right] = E_F\left[\left|\frac{d_{ji}u_{ji}\varepsilon_{ji}}{q_j\sigma_j}\right|^{2+\delta}\right] \leq K.$$

Assumption (d) also gives

$$\det(\Omega(F)) \geq \epsilon.$$

Hence Assumption 2b) holds.

Finally, by Lemma 3 below, the plug-in estimator $\widehat{\Sigma}$ in (13) satisfies, for every $\eta > 0$,

$$\lim_{n\to\infty}\sup_{F\in\mathcal{F}} P_F\left(\left\|V(F)^{-1/2}\widehat{\Sigma}V(F)^{-1/2} - \Omega(F)\right\| > \eta\right) = 0.$$

Thus Assumption 2c) holds.

All parts of Assumption 2 are therefore satisfied. Parts (a) and (b) of the result follow from Proposition 4 applied with $\mathcal{T} = [0,\infty)$, $A(t) \equiv A$ and $b(t) = t\mathbf{1}_{2m}$, so that $\mathcal{T}_0(F) = \{b \geq 0 : A\theta(F) \leq b\mathbf{1}_{2m}\}$. For part (b), A has no zero rows, and at $t = b_F$ the binding set $J(F, b_F)$ is nonempty because the maximum defining b_F is attained.

□

A.3 Uniform Consistency of the Plug-in Covariance Estimator

Lemma 3. Suppose Assumption 1 holds, and let $\hat{\Sigma}$ be the plug-in covariance estimator (13). Then, for every $\eta > 0$,

$$\lim_{n\to\infty}\sup_{F\in\mathcal{F}} P_F\left(\left\|V(F)^{-1/2}\hat{\Sigma}V(F)^{-1/2} - \Omega(F)\right\| > \eta\right) = 0.$$

Proof. Throughout, $O_{P_F}(\cdot)$ and $o_{P_F}(\cdot)$ statements hold uniformly over $F \in \mathcal{F}$, and ϵ, δ, K are the constants of Assumption 1. Since m is fixed, it suffices to show that, for each pair

(j,k),

$$\frac{\hat{\Sigma}_{jk}}{\sigma_j\sigma_k} - \Omega_{jk}(F) = o_{P_F}(1).$$

Step 1: Representation of the estimated influence function. Fix j and use the notation of the proof of Proposition 3 above: $q_j = E_F[d_{ji}u_{ji}^2]$,

$$a_n = E_n[d_{ji}u_{ji}\varepsilon_{ji}], \quad b_n = E_n[d_{ji}Z_{ji}u_{ji}], \quad c_n = E_n[d_{ji}Z_{ji}\varepsilon_{ji}], \quad C_n = E_n[d_{ji}Z_{ji}Z_{ji}'], \quad s_n = E_n[d_{ji}u_{ji}^2],$$

$D_n = s_n - b_n'C_n^{-1}b_n$, and $\psi_{ji} = \psi_{ji}(F) = d_{ji}u_{ji}\varepsilon_{ji}/q_j$, so that $\sigma_j^2 = E_F[\psi_{ji}^2]$ and $\Omega_{jk}(F) = E_F[\psi_{ji}\psi_{ki}]/(\sigma_j\sigma_k)$. That proof established, under Assumption 1,

$$b_n = O_{P_F}(n^{-1/2}), \quad c_n = O_{P_F}(n^{-1/2}), \quad C_n^{-1} = O_{P_F}(1), \quad D_n = q_j + O_{P_F}(n^{-1/2}), \quad \frac{\sqrt{n}(\hat{\theta}_j - \theta_j)}{\sigma_j} = O_{P_F}(1), \tag{P.1}$$

with $q_j \geq \epsilon$ and $\sigma_j^2 \geq \epsilon$.

Write $D_{ji} = u_{ji} + \gamma_{Dj}'Z_{ji}$ and $y_{ji} = \theta_j D_{ji} + \gamma_{yj}'Z_{ji} + \varepsilon_{ji}$. The coefficient from regressing D_{ji} on Z_{ji} over the observations with $d_{ji} = 1$ is $C_n^{-1}E_n[d_{ji}Z_{ji}D_{ji}] = \gamma_{Dj} + C_n^{-1}b_n$, so

$$\hat{u}_{ji} = u_{ji} - Z_{ji}'C_n^{-1}b_n, \qquad E_n[d_{ji}\hat{u}_{ji}^2] = D_n. \tag{P.2}$$

By the Frisch–Waugh–Lovell theorem, the OLS residual satisfies $\hat{\varepsilon}_{ji} = \check{y}_{ji} - \hat{\theta}_j\hat{u}_{ji}$, where $\check{y}_{ji} = y_{ji} - Z_{ji}'C_n^{-1}E_n[d_{ji}Z_{ji}y_{ji}]$ is the residual from regressing y_{ji} on Z_{ji}. Since $E_n[d_{ji}Z_{ji}y_{ji}] = c_n + \theta_j(b_n + C_n\gamma_{Dj}) + C_n\gamma_{yj}$, direct substitution gives $\check{y}_{ji} = \varepsilon_{ji} + \theta_j\hat{u}_{ji} - Z_{ji}'C_n^{-1}c_n$, and therefore

$$\hat{\varepsilon}_{ji} = \varepsilon_{ji} - (\hat{\theta}_j - \theta_j)\hat{u}_{ji} - Z_{ji}'C_n^{-1}c_n. \tag{P.3}$$

Multiplying (P.2) and (P.3),

$$d_{ji}\hat{u}_{ji}\hat{\varepsilon}_{ji} = d_{ji}u_{ji}\varepsilon_{ji} + d_{ji}e_{ji},$$

where

$$\begin{aligned} e_{ji} = & -(\hat{\theta}_j - \theta_j)\,u_{ji}\hat{u}_{ji} - u_{ji}\,Z_{ji}'C_n^{-1}c_n - \varepsilon_{ji}\,Z_{ji}'C_n^{-1}b_n \\ & + (\hat{\theta}_j - \theta_j)\,\hat{u}_{ji}\,Z_{ji}'C_n^{-1}b_n + (Z_{ji}'C_n^{-1}b_n)(Z_{ji}'C_n^{-1}c_n). \end{aligned}$$

Hence, by (12) and (P.2),

$$\frac{\hat{\psi}_{ji}}{\sigma_j} = \frac{q_j}{D_n}\left(\tilde{\psi}_{ji} + r_{ji}\right), \qquad \tilde{\psi}_{ji} = \frac{\psi_{ji}}{\sigma_j}, \qquad r_{ji} = \frac{d_{ji}e_{ji}}{q_j\sigma_j}. \tag{P.4}$$

Step 2: The remainder is negligible in mean square. We show $E_n[r_{ji}^2] = O_{P_F}(n^{-1})$. By Assumption 1(c), Markov's inequality, and the Cauchy–Schwarz inequality $E_F[d_{ji}u_{ji}^2\|Z_{ji}\|_2^2] \le (E_F[d_{ji}u_{ji}^4])^{1/2}(E_F\|d_{ji}Z_{ji}\|_2^4)^{1/2} \le K$,

$$E_n[d_{ji}u_{ji}^4], \quad E_n[d_{ji}\|Z_{ji}\|_2^4], \quad E_n[d_{ji}u_{ji}^2\|Z_{ji}\|_2^2], \quad E_n[d_{ji}\varepsilon_{ji}^2\|Z_{ji}\|_2^2]$$

are all $O_{P_F}(1)$. By (P.2), $E_n[d_{ji}\hat{u}_{ji}^4] \le 8\{E_n[d_{ji}u_{ji}^4] + \|C_n^{-1}b_n\|_2^4 E_n[d_{ji}\|Z_{ji}\|_2^4]\} = O_{P_F}(1)$, so by Cauchy–Schwarz $E_n[d_{ji}u_{ji}^2\hat{u}_{ji}^2]$ and $E_n[d_{ji}\hat{u}_{ji}^2\|Z_{ji}\|_2^2]$ are $O_{P_F}(1)$ as well. Each of the five terms of e_{ji} is the product of a data-dependent coefficient and an observation-level factor. Using $q_j \ge \epsilon$, $\sigma_j \ge \epsilon^{1/2}$, (P.1), and $(\sum_{\ell=1}^5 t_\ell)^2 \le 5\sum_{\ell=1}^5 t_\ell^2$,

$$\begin{aligned} E_n[r_{ji}^2] \le \frac{5}{\epsilon^2}\Bigg\{ &\frac{(\hat{\theta}_j-\theta_j)^2}{\sigma_j^2}E_n[d_{ji}u_{ji}^2\hat{u}_{ji}^2] + \frac{\|C_n^{-1}c_n\|_2^2}{\sigma_j^2}E_n[d_{ji}u_{ji}^2\|Z_{ji}\|_2^2] + \frac{\|C_n^{-1}b_n\|_2^2}{\sigma_j^2}E_n[d_{ji}\varepsilon_{ji}^2\|Z_{ji}\|_2^2] \\ &+ \frac{(\hat{\theta}_j-\theta_j)^2\|C_n^{-1}b_n\|_2^2}{\sigma_j^2}E_n[d_{ji}\hat{u}_{ji}^2\|Z_{ji}\|_2^2] + \frac{\|C_n^{-1}b_n\|_2^2\|C_n^{-1}c_n\|_2^2}{\sigma_j^2}E_n[d_{ji}\|Z_{ji}\|_2^4]\Bigg\}. \end{aligned}$$

The first three coefficients are $O_{P_F}(n^{-1})$ and the last two are $O_{P_F}(n^{-2})$ by (P.1), so $E_n[r_{ji}^2] = O_{P_F}(n^{-1})$.

Step 3: Uniform law of large numbers for the leading term. By Assumption 1(d), $E_F|\tilde{\psi}_{ji}|^{2+\delta} \le K$, and $E_F[\tilde{\psi}_{ji}\tilde{\psi}_{ki}] = \Omega_{jk}(F)$. By the Cauchy–Schwarz inequality,

$$E_F\left|\tilde{\psi}_{ji}\tilde{\psi}_{ki}\right|^{1+\delta/2} \le \left(E_F|\tilde{\psi}_{ji}|^{2+\delta}\right)^{1/2}\left(E_F|\tilde{\psi}_{ki}|^{2+\delta}\right)^{1/2} \le K.$$

A uniformly bounded moment of order $1+\delta/2$ implies a weak law of large numbers uniform over F: for $X_i = \tilde{\psi}_{ji}\tilde{\psi}_{ki}$ and $M > 0$, write $X_i = X_i\mathbf{1}\{|X_i| \le M\} + X_i\mathbf{1}\{|X_i| > M\}$; the truncated part has variance at most M^2, so Chebyshev's inequality bounds its deviation from its mean by $M^2/(n\eta^2)$ in probability, while the tail part satisfies $E_F[|X_i|\mathbf{1}\{|X_i| > M\}] \le E_F|X_i|^{1+\delta/2}/M^{\delta/2} \le KM^{-\delta/2}$ uniformly over F. Choosing M large and then n large gives

$$E_n[\tilde{\psi}_{ji}\tilde{\psi}_{ki}] - \Omega_{jk}(F) = o_{P_F}(1). \tag{P.5}$$

In particular, $E_n[\tilde{\psi}_{ji}^2] = 1 + o_{P_F}(1) = O_{P_F}(1)$.

Step 4: Conclusion. Let D_{nj} and D_{nk} denote the quantities D_n for specifications j and k. By (P.4),

$$\frac{\hat{\Sigma}_{jk}}{\sigma_j \sigma_k} = \frac{q_j q_k}{D_{nj} D_{nk}} \left\{ E_n[\tilde{\psi}_{ji}\tilde{\psi}_{ki}] + E_n[\tilde{\psi}_{ji} r_{ki}] + E_n[r_{ji}\tilde{\psi}_{ki}] + E_n[r_{ji} r_{ki}] \right\}.$$

By (P.1) and $q_j \geq \epsilon$, $q_j / D_{nj} = 1 + o_{P_F}(1)$. By Cauchy–Schwarz, Step 2 and Step 3,

$$\left| E_n[\tilde{\psi}_{ji} r_{ki}] \right| \leq \left(E_n[\tilde{\psi}_{ji}^2] \right)^{1/2} \left(E_n[r_{ki}^2] \right)^{1/2} = O_{P_F}(n^{-1/2}), \qquad |E_n[r_{ji} r_{ki}]| \leq \left(E_n[r_{ji}^2] E_n[r_{ki}^2] \right)^{1/2} = O_{P_F}(n^{-1}),$$

and similarly for $E_n[r_{ji}\tilde{\psi}_{ki}]$. Combining with (P.5), $\hat{\Sigma}_{jk}/(\sigma_j\sigma_k) - \Omega_{jk}(F) = o_{P_F}(1)$ uniformly over $F \in \mathcal{F}$, which is the claim. □

A.4 Proof of Corollary 1

Proof. By the proof of Proposition 3, Assumption 2 holds under the hypotheses of the corollary, so Proposition 4 is available. Each branch test is an instance of it: $\phi_n^{\pm}(c, \alpha)$ is the RCC test of the family $\mathcal{T} = [0, \infty)$, $A(t) = A^{\pm}(t)$, $b(t) = \mathbf{0}_{2m}$, with null sets

$$\mathcal{T}_0^{\pm}(F) = \left\{ c \geq 0 : A^{\pm}(c)\theta(F) \leq \mathbf{0}_{2m} \right\}.$$

Let

$$\mathcal{C}_0(F) = \left\{ c \geq 0 : \max_{j=1,\ldots,m} |\theta_j(F) - \theta_0(F)| \leq c|\theta_0(F)| \right\}$$

denote the set of values of c for which $H_0^{\mathrm{rel}}(c)$ holds under F. By the representation $H_0^{\mathrm{rel}}(c) = P_+(c) \cup P_-(c)$ in Definition 2,

$$\mathcal{C}_0(F) = \mathcal{T}_0^+(F) \cup \mathcal{T}_0^-(F).$$

Part (a). Fix $F \in \mathcal{F}$ and $c \in \mathcal{C}_0(F)$. If $c \in \mathcal{T}_0^+(F)$, then, since $\phi_n^{\mathrm{rel}} = \phi_n^+ \phi_n^- \leq \phi_n^+$,

$$E_F\left[\phi_n^{\mathrm{rel}}(c, \alpha)\right] \leq E_F\left[\phi_n^+(c, \alpha)\right] \leq \sup_{F' \in \mathcal{F}} \sup_{c' \in \mathcal{T}_0^+(F')} E_{F'}\left[\phi_n^+(c', \alpha)\right] =: S_n^+.$$

If instead $c \in \mathcal{T}_0^-(F)$, the same argument with the roles of the two branches exchanged gives $E_F[\phi_n^{\mathrm{rel}}(c, \alpha)] \leq S_n^-$, defined analogously. Since every $c \in \mathcal{C}_0(F)$ belongs to at least

one of the two sets,

$$\sup_{F\in\mathcal{F}}\sup_{c\in\mathcal{C}_0(F)} E_F\big[\phi_n^{\mathrm{rel}}(c,\alpha)\big] \le \max\{S_n^+, S_n^-\}.$$

Proposition 4(a), applied to each branch family, gives $\limsup_n S_n^{\pm} \le \alpha$. Part (a) follows.

Part (b). Fix $F \in \mathcal{F}$ and write $c_F = c(\theta(F))$. If $c_F = +\infty$, the event $\{c_{RT}(\alpha) > c_F\}$ is empty and there is nothing to prove. Otherwise, the defining inequality of $\mathcal{C}_0(F)$ is weak and continuous in c, so $\mathcal{C}_0(F)$ is closed; being nonempty and an interval unbounded above, it equals $[c_F, \infty)$, and in particular $c_F \in \mathcal{C}_0(F)$. By the definition of $c_{RT}(\alpha)$ as an infimum, $\phi_n^{\mathrm{rel}}(c_F, \alpha) = 0$ implies $c_{RT}(\alpha) \le c_F$; equivalently,

$$\{c_{RT}(\alpha) > c_F\} \subseteq \left\{\phi_n^{\mathrm{rel}}(c_F,\alpha) = 1\right\}.$$

Hence

$$P_F(c_{RT}(\alpha) > c_F) \le E_F\big[\phi_n^{\mathrm{rel}}(c_F,\alpha)\big] \le \sup_{c\in\mathcal{C}_0(F)} E_F\big[\phi_n^{\mathrm{rel}}(c,\alpha)\big].$$

Taking the supremum over $F \in \mathcal{F}$ and then $\limsup_n$, part (a) gives the result.

Part (c). Fix $F \in \mathcal{F}$, write $\theta = \theta(F)$ and $c_F = c(\theta)$, and recall that $A^+(0) = A^-(0)$, so that $\phi_n^+(0,\alpha) = \phi_n^-(0,\alpha) = \phi_n^{\mathrm{rel}}(0,\alpha)$.

Case $\theta = \mathbf{0}$. Then $c_F = 0$ and $\phi_n^{\mathrm{rel}}(0,\alpha) = \phi_n^+(0,\alpha)$. The matrix $A^+(0)$ has no zero rows, and all $2m$ inequalities bind at $\theta = \mathbf{0}$, so $J(F,0) = \{1,\dots,2m\}$ is nonempty. Proposition 4(b), applied to the plus-branch family, gives $\lim_n E_F[\phi_n^+(0,\alpha)] = \alpha$.

Case $\theta_0 > 0$. (The case $\theta_0 < 0$ is identical with the roles of the two branches exchanged.) Then $c_F = \max_j |\theta_j - \theta_0|/\theta_0 < \infty$ and $\theta \in P_+(c_F)$, so $c_F \in \mathcal{T}_0^+(F)$. We show that the plus-branch test is asymptotically exact at c_F and that, when $c_F > 0$, the minus-branch test rejects with probability approaching one.

(i) Plus branch. Every row of $A^+(c_F)$ contains an entry equal to 1 or -1 in one of columns $2,\dots,m+1$, so $A^+(c_F)$ has no zero rows. Let j^* attain $\max_j |\theta_j - \theta_0|$. If $\theta_{j^*} - \theta_0 = c_F\theta_0$, the inequality $\theta_{j^*} - (1+c_F)\theta_0 \le 0$ binds at θ; otherwise $\theta_{j^*} - \theta_0 = -c_F\theta_0$ and the inequality $(1-c_F)\theta_0 - \theta_{j^*} \le 0$ binds. Hence $J(F, c_F)$ is nonempty, and Proposition 4(b) gives

$$\lim_{n\to\infty} E_F\big[\phi_n^+(c_F,\alpha)\big] = \alpha.$$

If $c_F = 0$, then $\phi_n^{\mathrm{rel}}(c_F,\alpha) = \phi_n^+(c_F,\alpha)$ and the proof is complete. Suppose henceforth that

$c_F > 0$.

(ii) Minus branch. Adding the two inequalities that define $P_-(c_F)$ for any j gives $2c_F\mu_0 \le 0$, so $P_-(c_F) \subseteq \{\mu : \mu_0 \le 0\}$. Let $T_n^-(c_F)$ denote the RCC statistic of the minus branch, that is, (9) with the constraint set $\{\mu : A^-(c_F)\mu \le \mathbf{0}_{2m}\}$. For every $\mu \in P_-(c_F)$,

$$n(\widehat{\theta}-\mu)'\widehat{\Sigma}^{-1}(\widehat{\theta}-\mu) \ge \frac{n\|\widehat{\theta}-\mu\|_2^2}{\lambda_{\max}(\widehat{\Sigma})} \ge \frac{n\,|\widehat{\theta}_0-\mu_0|^2}{\lambda_{\max}(\widehat{\Sigma})} \ge \frac{n\,\max\{\widehat{\theta}_0,0\}^2}{\lambda_{\max}(\widehat{\Sigma})},$$

where the last inequality uses $\mu_0 \le 0$. Taking the minimum over $\mu \in P_-(c_F)$,

$$T_n^-(c_F) \ge \frac{n\,\max\{\widehat{\theta}_0,0\}^2}{\lambda_{\max}(\widehat{\Sigma})}.$$

For fixed F, Assumption 2a) and the central limit theorem give $\widehat{\theta}_0 \to \theta_0 > 0$ in P_F-probability, and the uniform consistency condition on $\widehat{\Sigma}$, with $V(F)$ fixed, gives $\lambda_{\max}(\widehat{\Sigma}) = O_{P_F}(1)$. Hence $T_n^-(c_F) \to \infty$ in P_F-probability.

The RCC critical value is $\chi^2_{\hat{r},1-\hat{\beta}}$ with $\hat{\beta} \in [\alpha, 2\alpha]$ and $\hat{r} \le m+1$, hence at most $\chi^2_{m+1,1-\alpha} < \infty$, provided $\hat{r} \ge 1$. On the event $\{\widehat{\theta}_0 > 0\}$, whose probability tends to one, $\widehat{\theta} \notin P_-(c_F)$; the restricted estimator $\hat{\mu}$ is then the projection of $\widehat{\theta}$ onto the closed convex set $P_-(c_F)$ in the metric induced by $\widehat{\Sigma}^{-1}$ and lies on its boundary, so at least one inequality is active at $\hat{\mu}$, and $\hat{r} \ge 1$ because every row of $A^-(c_F)$ is nonzero. Therefore

$$\lim_{n\to\infty} P_F\big(\phi_n^-(c_F,\alpha)=1\big) = 1.$$

(iii) Combination. Since $0 \le \phi_n^+ - \phi_n^+\phi_n^- \le 1-\phi_n^-$,

$$0 \le E_F\big[\phi_n^+(c_F,\alpha)\big] - E_F\big[\phi_n^{\mathrm{rel}}(c_F,\alpha)\big] \le 1 - P_F\big(\phi_n^-(c_F,\alpha)=1\big) \longrightarrow 0,$$

and (i) gives $\lim_n E_F[\phi_n^{\mathrm{rel}}(c_F,\alpha)] = \alpha$. This establishes part (c). □

A.5 Proofs for Section 2.2

Proof of Lemma 1. Upper bound. At $b = \max_{j\le m}|\hat{\Delta}_j|$ the point $\mu = \hat{\theta}$ satisfies $A\mu \le b\mathbf{1}_{2m}$, so $T_n(b) = 0$ in (9) and the test does not reject. By Definition 1, $b_{RT}(\alpha) \le b$.

Lower bound. Fix $j \le m$ and write $a = e_j - e_0$, with e_j the jth unit vector, so that $\hat{\sigma}^2_{\Delta,j} = a'\hat{\Sigma}a$. For every μ with $A\mu \le b\mathbf{1}_{2m}$ one has $|a'\mu| = |\mu_j - \mu_0| \le b$, hence $|a'(\hat{\theta}-\mu)| \ge$

$|\hat{\Delta}_j| - b$. By the Cauchy–Schwarz inequality in the $\hat{\Sigma}^{-1}$ metric, $(a'x)^2 \le (a'\hat{\Sigma}a)(x'\hat{\Sigma}^{-1}x)$ for every x, so

$$T_n(b) = \min_{\mu: A\mu \le b\mathbf{1}_{2m}} n(\hat{\theta}-\mu)'\hat{\Sigma}^{-1}(\hat{\theta}-\mu) \;\ge\; \frac{n\,(|\hat{\Delta}_j| - b)_+^2}{\hat{\sigma}^2_{\Delta,j}}.$$

The matrix A has rank m, so $\hat{r}(b) \le m$ and the critical value of the CC test is at most $\chi^2_{m,1-\alpha}$; the RCC critical value differs from the CC one only when $\hat{r}(b) = 1$, where it equals $\chi^2_{1,1-\hat{\beta}}$ with $\hat{\beta} \in [\alpha, 2\alpha]$, which is also at most $\chi^2_{m,1-\alpha}$. Hence the test rejects at every b with $n(|\hat{\Delta}_j| - b)^2/\hat{\sigma}^2_{\Delta,j} > \chi^2_{m,1-\alpha}$, that is, at every $b < |\hat{\Delta}_j| - \hat{\sigma}_{\Delta,j}\sqrt{\chi^2_{m,1-\alpha}/n}$. Since $b_{RT}(\alpha)$ is the infimum of the thresholds that are not rejected, it is at least this value, for each j. □

Proof of Proposition 1. Throughout, $\sqrt{n}\,V(F)^{-1/2}(\hat{\theta}-\theta) = O_{P_F}(1)$ uniformly over $F \in \mathcal{F}$ by Assumption 2a,b) (see the preliminaries of the proof of Proposition 4), and $V(F)^{-1/2}\hat{\Sigma}V(F)^{-1/2} - \Omega(F) = o_{P_F}(1)$ uniformly by Assumption 2c). Since $\operatorname{tr}\Omega(F) = m+2$ and $\det\Omega(F) \ge \epsilon$, there is $c > 0$ with $\lambda_{\min}(\Omega(F)) \ge c$ for all F. Two consequences are used below. First, let e_j denote the jth unit vector and $a = e_j - e_0$, so that $\sigma^2_{\Delta,j} = a'\Sigma a$ and $\hat{\sigma}^2_{\Delta,j} = a'\hat{\Sigma}a$. Then

$$\sqrt{n}\,|\hat{\Delta}_j - \Delta_j| = \left|(V^{1/2}a)'\sqrt{n}\,V^{-1/2}(\hat{\theta}-\theta)\right| \le \|V^{1/2}a\|_2\, O_{P_F}(1) \le c^{-1/2}\sigma_{\Delta,j}\, O_{P_F}(1),$$

because $\sigma^2_{\Delta,j} = (V^{1/2}a)'\Omega(V^{1/2}a) \ge c\|V^{1/2}a\|_2^2$; so $\sqrt{n}\,|\hat{\Delta}_j - \Delta_j|/\sigma_{\Delta,j} = O_{P_F}(1)$ uniformly over F and j. Second, by the same inequality,

$$\left|\frac{\hat{\sigma}^2_{\Delta,j}}{\sigma^2_{\Delta,j}} - 1\right| = \frac{|(V^{1/2}a)'(V^{-1/2}\hat{\Sigma}V^{-1/2} - \Omega)(V^{1/2}a)|}{(V^{1/2}a)'\Omega(V^{1/2}a)} \le \frac{\|V^{-1/2}\hat{\Sigma}V^{-1/2} - \Omega\|}{c} = o_{P_F}(1)$$

uniformly over F and j. The set $\{1,\dots,m\}$ satisfies Assumption 2 as well: its objects are the leading blocks of those of $\{1,\dots,m+1\}$, and the determinant of a principal submatrix of $\Omega(F)$ is at least $\lambda_{\min}(\Omega(F))^{m+1} \ge c^{m+1}$ by eigenvalue interlacing.

Part (c). Fix F and let $j^* \le m$ attain $b^{(m)}(\theta) = |\Delta_{j^*}|$. By Lemma 1,

$$|\hat{\Delta}_{j^*}| - \hat{\sigma}_{\Delta,j^*}\sqrt{\chi^2_{m,1-\alpha}/n} \;\le\; b^{(m)}_{RT}(\alpha) \;\le\; \max_{j\le m}|\hat{\Delta}_j|.$$

Both ends converge in probability to $b^{(m)}(\theta)$: the right end because $\hat{\Delta}_j \to_p \Delta_j$ for each j, the left end because in addition $\hat{\sigma}_{\Delta,j^*} \to_p \sigma_{\Delta,j^*} < \infty$. The same argument for the set $\{1,\dots,m+1\}$, with j^* attaining $b^{(m+1)}(\theta)$, gives the second limit.

Part (b). If $|\Delta_{m+1}| \le b^{(m)}(\theta)$ then $b^{(m+1)}(\theta) = b^{(m)}(\theta)$, so $t = b^{(m)}(\theta(F))$ belongs to the null set $\mathcal{T}_0(F) = \{b \ge 0 : A\theta(F) \le b\mathbf{1}_{2(m+1)}\}$ of the problem with checks $\{1, \dots, m+1\}$. By Definition 1, $b_{RT}^{(m+1)}(\alpha) > t$ requires the test to reject at t, so $\mathbb{P}_F(b_{RT}^{(m+1)}(\alpha) > t) \le E_F[\phi_n^{\mathrm{RCC}}(t, \alpha)]$, and Proposition 4(a) bounds the lim sup of the right-hand side by α. This is Property 2.

Part (a), Property 1(i). If $|\Delta_{m+1}| > b^{(m)}(\theta)$, part (c) gives $b_{RT}^{(m+1)}(\alpha) - b_{RT}^{(m)}(\alpha) \to_p |\Delta_{m+1}| - b^{(m)}(\theta) > 0$.

Part (a), additional statement. Let $E_n = \{b_{RT}^{(m)}(\alpha) \le b^{(m)}(\theta(F_n))\}$. By the argument of part (b) applied to the set $\{1, \dots, m\}$, $\liminf_n \mathbb{P}_{F_n}(E_n) \ge 1 - \alpha$. On E_n, by the lower bound of Lemma 1 for the set $\{1, \dots, m+1\}$,

$$\frac{\sqrt{n}\left(b_{RT}^{(m+1)}(\alpha) - b_{RT}^{(m)}(\alpha)\right)}{\sigma_{\Delta,m+1}} \ge \frac{\sqrt{n}\left(|\Delta_{m+1}| - b^{(m)}(\theta)\right)}{\sigma_{\Delta,m+1}} - \frac{\sqrt{n}\,|\hat{\Delta}_{m+1} - \Delta_{m+1}|}{\sigma_{\Delta,m+1}} - \frac{\hat{\sigma}_{\Delta,m+1}}{\sigma_{\Delta,m+1}}\sqrt{\chi^2_{m+1,1-\alpha}},$$

and the right-hand side diverges in probability by the two preliminary facts. Hence

$$\mathbb{P}_{F_n}\left(b_{RT}^{(m+1)}(\alpha) > b_{RT}^{(m)}(\alpha)\right) \ge \mathbb{P}_{F_n}(E_n) - o(1),$$

which gives the claim. Only the standard error of check $m+1$ enters; those of the other checks are absorbed by the event E_n.

Part (a), Property 1(ii). By Lemma 1 applied to both sets,

$$b_{RT}^{(m+1)}(\alpha) - b_{RT}^{(m)}(\alpha) \;\ge\; |\hat{\Delta}_{m+1}| - \hat{\sigma}_{\Delta,m+1}\sqrt{\chi^2_{m+1,1-\alpha}/n} - \max_{j \le m} |\hat{\Delta}_j|.$$

Using $|\hat{\Delta}_{m+1}| \ge |\Delta_{m+1}| - |\hat{\Delta}_{m+1} - \Delta_{m+1}|$ and $\max_{j \le m} |\hat{\Delta}_j| \le b^{(m)}(\theta) + \max_{j \le m} |\hat{\Delta}_j - \Delta_j|$, and multiplying by $\sqrt{n}/\bar{\sigma}_\Delta$,

$$\frac{\sqrt{n}\left(b_{RT}^{(m+1)}(\alpha) - b_{RT}^{(m)}(\alpha)\right)}{\bar{\sigma}_\Delta} \ge \frac{\sqrt{n}\left(|\Delta_{m+1}| - b^{(m)}(\theta)\right)}{\bar{\sigma}_\Delta} - 2\max_{j \le m+1} \frac{\sqrt{n}\,|\hat{\Delta}_j - \Delta_j|}{\sigma_{\Delta,j}} - \frac{\hat{\sigma}_{\Delta,m+1}}{\sigma_{\Delta,m+1}}\sqrt{\chi^2_{m+1,1-\alpha}}.$$

Along the sequence $\{F_n\}$ the first term diverges, the second is $O_{P_{F_n}}(1)$ and the third is $\sqrt{\chi^2_{m+1,1-\alpha}}(1 + o_{P_{F_n}}(1))$ by the two preliminary facts. Hence the left-hand side exceeds zero with probability approaching one, which completes part (a). □

Proof of Proposition 2. Part (a). Fix θ and n. Let $j^* \le m+1$ attain $b(\theta) = |\Delta_{j^*}|$, let $s = \operatorname{sign}(\Delta_{j^*})$ with $s = 1$ if $\Delta_{j^*} = 0$, and for $h > 0$ set $\theta_h = \theta + h\, s\, e_{j^*}$, where e_{j^*} is the j^*th unit vector. Then $\Delta_{j^*}(\theta_h) = \Delta_{j^*} + hs$, so $|\Delta_{j^*}(\theta_h)| = b(\theta) + h$ and the other discrepancies are unchanged; hence $b(\theta_h) = b(\theta) + h > b(\theta)$. Validity at θ_h gives $\mathbb{P}_{\theta_h}(U_n > b(\theta)) \ge \mathbb{P}_{\theta_h}(U_n \ge b(\theta_h)) \ge 1-\alpha$. The laws of $\hat\theta$ under θ and θ_h are $N(\theta, \Sigma/n)$ and $N(\theta_h, \Sigma/n)$, whose total-variation distance is $2\Phi(d_h/2) - 1$ with $d_h^2 = nh^2(\Sigma^{-1})_{j^*j^*} < \infty$, which tends to zero as $h \to 0$. Since $\{U_n > b(\theta)\}$ is an event determined by $\hat\theta$,

$$\mathbb{P}_\theta\big(U_n > b(\theta)\big) \ge \mathbb{P}_{\theta_h}\big(U_n > b(\theta)\big) - \big(2\Phi(d_h/2) - 1\big) \ge 1 - \alpha - \big(2\Phi(d_h/2) - 1\big),$$

and letting $h \to 0$ gives the claim. If $|\Delta_{m+1}| \le b^{(m)}(\theta)$ then $b(\theta) = b^{(m+1)}(\theta) = b^{(m)}(\theta)$, so $\mathbb{P}_\theta(U_n > b^{(m)}(\theta)) \ge 1 - \alpha > \alpha$ for every n, and the lim sup in Property 2 is at least $1 - \alpha$.

Part (b). Fix θ and K, and let $\theta' = \theta + s\, e_{m+1}$ with $s = K + |\Delta_{m+1}|$, so that $b(\theta') \ge |\Delta_{m+1} + s| \ge s - |\Delta_{m+1}| = K$. Validity at θ' gives $\mathbb{P}_{\theta'}(U_n \ge K) \ge \mathbb{P}_{\theta'}(U_n \ge b(\theta')) \ge 1 - \alpha$. The total-variation distance between $N(\theta, \Sigma/n)$ and $N(\theta', \Sigma/n)$ is $2\Phi(d/2) - 1$ with

$$d^2 = ns^2(\Sigma^{-1})_{m+1,m+1} = \frac{ns^2}{\Sigma_{m+1,m+1}(1 - R^2_{m+1})},$$

where $R^2_{m+1} < 1$ is the squared multiple correlation of coordinate $m+1$ with the remaining coordinates under Ω, which is fixed and, since Σ is positive definite, strictly below one. As $\Sigma_{m+1,m+1} \to \infty$, $d \to 0$ and the total-variation distance tends to zero, so $\mathbb{P}_\theta(U_n \ge K) \ge \mathbb{P}_{\theta'}(U_n \ge K) - (2\Phi(d/2) - 1) \ge 1 - \alpha - o(1)$. □

A.6 Bootstrap Covariance Estimator

This appendix describes an alternative to the plug-in covariance estimator (13), for which Proposition 5 establishes uniform consistency under Assumption 1 and the additional conditions of Assumption 3. It therefore satisfies the condition of Lemma 3, and Proposition 3 and Corollary 1 hold with it in place of (13). Relative to the plug-in estimator, the bootstrap requires more conditions and has the same asymptotic limit; it is included because the same procedure can be applied to estimators other than least squares, for which an analytic influence function may be less convenient. The estimator is the pairs bootstrap covariance of the stacked coefficient vector, computed with the design-matrix trimming rule of Hansen (2022, Section 10.27, equation (10.34)), which follows Shao and Tu (1995, p. 291): a bootstrap replication is retained only if, for every specification, the

smallest eigenvalue of the resampled design matrix is at least a fraction τ of that of the original design matrix. Some such rule is needed because OLS coefficients have a ratio structure: convergence in distribution of the bootstrap estimator does not by itself imply convergence of its second moments (Hansen, 2022, Sections 10.13 and 10.28), and the conditions imposed below control the population design but do not rule out a nearly singular design in a given resample.[16] Because each $\hat{\theta}_j$ is a coefficient of a linear regression, the trimming rule is the only trimming required: once the resampled design is bounded away from singularity, the bootstrap estimator is a bounded linear transformation of a bootstrap sample mean, and no trimming of the statistic itself, as in Hansen (2022, Section 10.14), is necessary. I set $\tau = 1/2$, following the recommendation reported by Hansen (2022, Section 10.27); any fixed $\tau \in (0, 1)$ gives the same asymptotic result. The joint rule sets the entire bootstrap coefficient vector equal to its original-sample value whenever the design matrix of at least one specification fails the criterion. In the following, B denotes the number of bootstrap replications.

[16] See Hansen (2022, Section 10.15) for an illustration of how unreliable untrimmed bootstrap standard errors can be.

Algorithm 1 Trimmed pairs-bootstrap covariance estimator

Estimate all $m+1$ specifications on the original sample and collect

$$\widehat{\theta} = (\widehat{\theta}_0, \ldots, \widehat{\theta}_m)'.$$

For each $j = 0, \ldots, m$, compute

$$\widehat{Q}_j = E_n[d_{ji} X_{ji} X_{ji}'].$$

for $b = 1, \ldots, B$ **do**

Draw n observations with replacement from the original sample, using the same bootstrap draw for all specifications.

For each $j = 0, \ldots, m$, compute

$$\widehat{Q}_j^{*(b)} = E_n^*[d_{ji}^* X_{ji}^* X_{ji}^{*\prime}]$$

and

$$\lambda_j^{*(b)} = \frac{\lambda_{\min}(\widehat{Q}_j^{*(b)})}{\lambda_{\min}(\widehat{Q}_j)}.$$

if $\min_{j=0,\ldots,m} \lambda_j^{*(b)} \geq \tau$ **then**

Estimate all $m+1$ specifications in the bootstrap sample and collect their coefficients of interest in $\widehat{\theta}^{*(b)}$.

Set

$$\widetilde{\theta}^{*(b)} = \widehat{\theta}^{*(b)}.$$

else

Set

$$\widetilde{\theta}^{*(b)} = \widehat{\theta}.$$

end if

Compute

$$S_n^{*(b)} = \sqrt{n}\left(\widetilde{\theta}^{*(b)} - \widehat{\theta}\right).$$

end for

Compute

$$\overline{S}_n^* = \frac{1}{B}\sum_{b=1}^{B} S_n^{*(b)}$$

and

$$\widehat{\Sigma}_B^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(S_n^{*(b)} - \overline{S}_n^*\right)\left(S_n^{*(b)} - \overline{S}_n^*\right)'.$$

return $\widehat{\Sigma}_B^*$.

Assumption 3 imposes the additional conditions needed for consistency of the bootstrap covariance estimator. They are not required for Proposition 3, which uses the plug-in estimator.

Assumption 3 (Bootstrap covariance estimation). There exist constants $\epsilon_b > 0$, $\delta_b > 0$, and $K_b < \infty$ such that the following conditions hold uniformly over $F \in \mathcal{F}$, for every $j = 0, \ldots, m$:

(a)
$$\lambda_{\min}\left(E_F[d_{ji} X_{ji} X_{ji}']\right) \geq \epsilon_b.$$

(b)
$$E_F[d_{ji}\|X_{ji}\|_2^{4+\delta_b}] \leq K_b, \qquad E_F\left[\left\|\frac{d_{ji} X_{ji} \varepsilon_{ji}}{\sigma_j}\right\|_2^{2+\delta_b}\right] \leq K_b.$$

Assumption 3(a) implies Assumption 1(b). Write $Q_j = E_F[d_{ji} X_{ji} X_{ji}']$. For any v, setting $w = (0, v')'$ gives $v' Q_{Zj} v = w' Q_j w \geq \lambda_{\min}(Q_j)\|v\|_2^2$, so $\lambda_{\min}(Q_{Zj}) \geq \epsilon_b$. Moreover,

$$E_F[d_{ji} u_{ji}^2] = \min_{\gamma} E_F\left[d_{ji}\left(D_{ji} - Z_{ji}'\gamma\right)^2\right] = \min_{\gamma} w_\gamma' Q_j w_\gamma, \qquad w_\gamma = (1, -\gamma')',$$

and $\|w_\gamma\|_2 \geq 1$, so $E_F[d_{ji} u_{ji}^2] \geq \lambda_{\min}(Q_j) \geq \epsilon_b$. Assumption 1(b) is retained separately because Proposition 3 does not require Assumption 3.

Proposition 5 (Uniform consistency of the bootstrap covariance estimator). Suppose Assumptions 1 and 3 hold. Let

$$\widehat{\Sigma}^* = \mathrm{Var}^*\left(\sqrt{n}(\widetilde{\theta}^* - \widehat{\theta})\right),$$

where $\widetilde{\theta}^*$ is the jointly trimmed pairs-bootstrap estimator defined in Algorithm 1. Then, for every $\eta > 0$,

$$\lim_{n\to\infty} \sup_{F \in \mathcal{F}} P_F\left(\left\|V(F)^{-1/2} \widehat{\Sigma}^* V(F)^{-1/2} - \Omega(F)\right\| > \eta\right) = 0.$$

Moreover, for fixed n, as $B \to \infty$,

$$\widehat{\Sigma}_B^* \xrightarrow{p^*} \widehat{\Sigma}^*.$$

Proof. The argument adapts the regression-bootstrap analysis in Hansen (2022, Sections 10.27–10.28) to the joint collection of regressions considered here and strengthens the

convergence statements to hold uniformly over $\mathcal{F}$. The moment argument used below parallels the argument in Hansen (2022, Sections 10.13 and 10.31) for consistency of bootstrap variance estimators.

For each $j = 0, \ldots, m$, define

$$Q_j(F) = E_F[d_{ji} X_{ji} X_{ji}'].$$

The proof proceeds in five steps.

Step 1: Uniform control of the sample and bootstrap design matrices.

By Assumption (a),

$$\inf_{F \in \mathcal{F}} \min_{j=0,\ldots,m} \lambda_{\min}(Q_j(F)) \geq \epsilon_b. \tag{A.1}$$

Moreover,

$$\widehat{Q}_j = E_n[d_{ji} X_{ji} X_{ji}'].$$

Since the dimension of X_{ji} is fixed and Assumption (b) implies uniformly bounded moments of order strictly greater than two for the elements of $d_{ji} X_{ji} X_{ji}'$, a uniform law of large numbers gives

$$\sup_{F \in \mathcal{F}} P_F \left(\max_{j=0,\ldots,m} \|\widehat{Q}_j - Q_j(F)\| > \eta \right) \longrightarrow 0 \tag{A.2}$$

for every $\eta > 0$. Since m is fixed, (A.1), (A.2), and Weyl's inequality imply

$$\sup_{F \in \mathcal{F}} P_F \left(\min_j \lambda_{\min}(\widehat{Q}_j) < \frac{\epsilon_b}{2} \right) \longrightarrow 0. \tag{A.3}$$

Conditional on the original sample,

$$\widehat{Q}_j^* = E_n^*[d_{ji}^* X_{ji}^* X_{ji}^{*\prime}]$$

is the mean of n independent draws from the empirical distribution of $d_{ji} X_{ji} X_{ji}'$. In particular,

$$E^* \left[\|\widehat{Q}_j^* - \widehat{Q}_j\|_F^2 \right] \leq \frac{1}{n} E_n[d_{ji} \|X_{ji}\|_2^4]. \tag{A.4}$$

Assumption (b) and a uniform law of large numbers imply

$$E_n[d_{ji} \|X_{ji}\|_2^4] = O_{P_F}(1)$$

uniformly over F. Hence, by conditional Markov's inequality,

$$\max_j \|\widehat{Q}_j^* - \widehat{Q}_j\| = o_{P^*}(1) \tag{A.5}$$

uniformly over F, in probability.

Recall that

$$\lambda_j^* = \frac{\lambda_{\min}(\widehat{Q}_j^*)}{\lambda_{\min}(\widehat{Q}_j)}.$$

Equations (A.3)–(A.5) and Weyl's inequality therefore imply, for every fixed $\tau \in (0,1)$,

$$\sup_{F \in \mathcal{F}} P_F \left[P^* \left(\min_{j=0,\ldots,m} \lambda_j^* < \tau \right) > \eta \right] \longrightarrow 0 \tag{A.6}$$

for every $\eta > 0$. Thus the joint trimming event is uniformly asymptotically negligible.

This is the joint analogue of the eigenvalue trimming in equation (10.34) of Hansen (2022). Here a bootstrap draw is retained only when the design matrices of all $m+1$ specifications satisfy the eigenvalue criterion.

Step 2: Bootstrap linear representation.

Let N_i^* denote the number of times observation i is selected in a bootstrap replication. Conditional on the original sample,

$$(N_1^*, \ldots, N_n^*) \sim \text{Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

The same vector $(N_1^*, \ldots, N_n^*)$ is used for every specification, thereby preserving the dependence across the coefficient estimators.

For specification j, let

$$\widehat{\varepsilon}_{ji} = y_{ji} - X_{ji}'\widehat{\beta}_j.$$

The original-sample OLS first-order conditions give

$$E_n[d_{ji} X_{ji} \widehat{\varepsilon}_{ji}] = 0. \tag{A.7}$$

Consequently, whenever the bootstrap replication satisfies the eigenvalue criterion,

$$\sqrt{n}(\widehat{\beta}_j^* - \widehat{\beta}_j) = (\widehat{Q}_j^*)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i^* - 1) d_{ji} X_{ji} \widehat{\varepsilon}_{ji}. \tag{A.8}$$

This is the usual pairs-bootstrap representation for OLS; compare Hansen (2022, Section 10.28).

Let e_{1j} denote the vector selecting the coefficient on D_{ji} from β_j, and define

$$\widehat{\psi}_{ji} = e_{1j}'\widehat{Q}_j^{-1} d_{ji} X_{ji}\widehat{\varepsilon}_{ji}, \qquad \widehat{\psi}_i = (\widehat{\psi}_{0i}, \ldots, \widehat{\psi}_{mi})'. \tag{A.9}$$

Using

$$(\widehat{Q}_j^*)^{-1} - \widehat{Q}_j^{-1} = -(\widehat{Q}_j^*)^{-1}(\widehat{Q}_j^* - \widehat{Q}_j)\widehat{Q}_j^{-1}, \tag{A.10}$$

together with (A.3)–(A.5), gives, on retained draws,

$$(\widehat{Q}_j^*)^{-1} = \widehat{Q}_j^{-1} + o_{P^*}(1) \tag{A.11}$$

uniformly over F, in probability.

Let

$$S_j^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i^* - 1) d_{ji} X_{ji}\widehat{\varepsilon}_{ji}$$

denote the centered bootstrap score in (A.8). Conditional on the original sample, $(N_1^*, \ldots, N_n^*)$ is multinomial, so for any vectors $v_1, \ldots, v_n$,

$$\mathrm{Var}^*\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n N_i^* v_i\right) = E_n[v_i v_i'] - E_n[v_i]E_n[v_i]'.$$

Applying this with $v_i = d_{ji} X_{ji}\widehat{\varepsilon}_{ji}$, which has $E_n[v_i] = 0$ by (A.7), gives

$$E^*\left[\left\|\frac{S_j^*}{\sigma_j}\right\|_2^2\right] = E_n\left[\left\|\frac{d_{ji} X_{ji}\widehat{\varepsilon}_{ji}}{\sigma_j}\right\|_2^2\right] \le 2E_n\left[\left\|\frac{d_{ji} X_{ji}\varepsilon_{ji}}{\sigma_j}\right\|_2^2\right] + 2E_n\left[d_{ji}\|X_{ji}\|_2^4\right] \frac{\|\widehat{\beta}_j - \beta_j\|_2^2}{\sigma_j^2} = O_{P_F}(1)$$

uniformly over F, where the inequality uses (A.21) and the final equality uses Assumption (b) and (A.24) below, whose derivation relies only on (A.3) and Assumption (b). Hence $S_j^*/\sigma_j = O_{P^*}(1)$ uniformly over F, in probability.

Substituting (A.11) into (A.8), selecting the coefficient on D_{ji}, and stacking the $m+1$ equations therefore yields

$$V(F)^{-1/2}\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) = V(F)^{-1/2}\frac{1}{\sqrt{n}} \sum_{i=1}^n (N_i^* - 1)\widehat{\psi}_i + o_{P^*}(1) \tag{A.12}$$

uniformly over F, in probability.

Step 3: Conditional Gaussian approximation.

Define the population score

$$\psi_{ji} = e_{1j}' Q_j(F)^{-1} d_{ji} X_{ji} \varepsilon_{ji}, \qquad \psi_i = (\psi_{0i}, \ldots, \psi_{mi})'. \tag{A.13}$$

By the Frisch–Waugh–Lovell theorem,

$$\psi_{ji} = \frac{d_{ji} u_{ji} \varepsilon_{ji}}{E_F[d_{ji} u_{ji}^2]}. \tag{A.14}$$

Thus

$$E_F[\psi_i] = 0, \qquad E_F[\psi_i \psi_i'] = \Sigma(F). \tag{A.15}$$

Let

$$\xi_i(F) = V(F)^{-1/2} \psi_i. \tag{A.16}$$

Then

$$E_F[\xi_i(F)] = 0, \qquad E_F[\xi_i(F)\xi_i(F)'] = \Omega(F). \tag{A.17}$$

Assumption (d) implies

$$\sup_{F \in \mathcal{F}} \max_j E_F \left[\left| \frac{\psi_{ji}}{\sigma_j} \right|^{2+\delta} \right] \leq K. \tag{A.18}$$

Since m is fixed,

$$\sup_{F \in \mathcal{F}} E_F[\|\xi_i(F)\|_2^{2+\delta}] \leq C \tag{A.19}$$

for some finite C.

The diagonal elements of $\Omega(F)$ equal one. Hence

$$\operatorname{tr}(\Omega(F)) = m + 1.$$

Together with $\det(\Omega(F)) \geq \epsilon$ from Assumption (d), fixed dimension implies that there exists $c > 0$ such that

$$\inf_{F \in \mathcal{F}} \lambda_{\min}(\Omega(F)) \geq c. \tag{A.20}$$

We now verify the corresponding properties for the estimated scores. From (A.9) and

$$\widehat{\varepsilon}_{ji} = \varepsilon_{ji} - X_{ji}'(\widehat{\beta}_j - \beta_j), \tag{A.21}$$

we have

$$\widehat{\psi}_{ji} = e_{1j}'\widehat{Q}_j^{-1} d_{ji} X_{ji}\varepsilon_{ji} - e_{1j}'\widehat{Q}_j^{-1} d_{ji} X_{ji} X_{ji}'(\widehat{\beta}_j - \beta_j). \tag{A.22}$$

First,

$$\frac{\widehat{\beta}_j - \beta_j}{\sigma_j} = \widehat{Q}_j^{-1} E_n\left[\frac{d_{ji} X_{ji}\varepsilon_{ji}}{\sigma_j}\right]. \tag{A.23}$$

The second moment of the vector inside the empirical average is uniformly bounded by Assumption (b). Together with (A.3), this gives

$$\frac{\|\widehat{\beta}_j - \beta_j\|_2}{\sigma_j} = O_{P_F}(n^{-1/2}) \tag{A.24}$$

uniformly over F.

Set $\delta_0 = \delta_b/2$. Dividing (A.22) by σ_j and using the triangle inequality,

$$\frac{|\widehat{\psi}_{ji}|}{\sigma_j} \leq \|\widehat{Q}_j^{-1}\| \left\{ \left\| \frac{d_{ji} X_{ji}\varepsilon_{ji}}{\sigma_j} \right\|_2 + d_{ji}\|X_{ji}\|_2^2 \frac{\|\widehat{\beta}_j - \beta_j\|_2}{\sigma_j} \right\}. \tag{A.25}$$

For the first term in braces, Assumption (b) directly gives

$$E_n\left[\left\| \frac{d_{ji} X_{ji}\varepsilon_{ji}}{\sigma_j} \right\|_2^{2+\delta_0}\right] = O_{P_F}(1) \tag{A.26}$$

uniformly over F.

For the second term, since

$$2(2+\delta_0) = 4 + 2\delta_0 \leq 4 + \delta_b,$$

Assumption (b) implies

$$E_n[d_{ji}\|X_{ji}\|_2^{4+2\delta_0}] = O_{P_F}(1) \tag{A.27}$$

uniformly over F. Combining (A.24) and (A.27),

$$E_n\left[\left\{ d_{ji}\|X_{ji}\|_2^2 \frac{\|\widehat{\beta}_j - \beta_j\|_2}{\sigma_j} \right\}^{2+\delta_0}\right] = o_{P_F}(1) \tag{A.28}$$

uniformly over F.

Equations (A.3) and (A.25)–(A.28) therefore imply

$$E_n\left[\left|\frac{\widehat{\psi}_{ji}}{\sigma_j}\right|^{2+\delta_0}\right] = O_{P_F}(1) \tag{A.29}$$

uniformly over F. Since m is fixed,

$$E_n\left[\left\|V(F)^{-1/2}\widehat{\psi}_i\right\|_2^{2+\delta_0}\right] = O_{P_F}(1) \tag{A.30}$$

uniformly over F.

Next, (A.2), (A.21)–(A.24), and the same moment bounds imply

$$V(F)^{-1/2}E_n[(\widehat{\psi}_i - \psi_i)(\widehat{\psi}_i - \psi_i)']V(F)^{-1/2} = o_{P_F}(1) \tag{A.31}$$

uniformly over F. Assumption (d) and a uniform law of large numbers also give

$$V(F)^{-1/2}E_n[\psi_i\psi_i']V(F)^{-1/2} - \Omega(F) = o_{P_F}(1) \tag{A.32}$$

uniformly over F. Applying Cauchy–Schwarz to the cross terms in

$$\widehat{\psi}_i\widehat{\psi}_i' - \psi_i\psi_i'$$

and combining (A.31)–(A.32) yields

$$V(F)^{-1/2}E_n[\widehat{\psi}_i\widehat{\psi}_i']V(F)^{-1/2} - \Omega(F) = o_{P_F}(1) \tag{A.33}$$

uniformly over F.

Conditional on the data, the bootstrap observations are i.i.d. draws from the empirical distribution. Moreover, by (A.7),

$$E_n[\widehat{\psi}_i] = 0.$$

For any $a \in \mathbb{R}^{m+1}$ with $\|a\|_2 = 1$, define

$$\widehat{\xi}_i = V(F)^{-1/2}\widehat{\psi}_i.$$

By (A.30),

$$E_n[|a'\widehat{\xi}_i|^{2+\delta_0}] = O_{P_F}(1) \tag{A.34}$$

uniformly over F. By (A.20) and (A.33),

$$a' E_n[\widehat{\xi}_i \widehat{\xi}_i'] a \geq \frac{c}{2} \tag{A.35}$$

with probability approaching one uniformly over F.

Therefore the conditional Lyapunov ratio is bounded by

$$\frac{n E_n[|a'\widehat{\xi}_i|^{2+\delta_0}]}{\left\{ n a' E_n[\widehat{\xi}_i \widehat{\xi}_i'] a \right\}^{1+\delta_0/2}} \leq \frac{C_n}{n^{\delta_0/2} (c/2)^{1+\delta_0/2}}, \tag{A.36}$$

where $C_n = O_{P_F}(1)$ uniformly over F. The right-hand side converges to zero uniformly. The conditional Lyapunov central limit theorem and the Cramér–Wold device therefore imply

$$V(F)^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (N_i^* - 1) \widehat{\psi}_i \rightsquigarrow^* N(0, \Omega(F)) \tag{A.37}$$

uniformly over $F \in \mathcal{F}$, in probability.

Combining (A.12) and (A.37) gives

$$V(F)^{-1/2} \sqrt{n} (\widehat{\theta}^* - \widehat{\theta}) \rightsquigarrow^* N(0, \Omega(F)) \tag{A.38}$$

uniformly over F. By (A.6), the joint trimming event is uniformly asymptotically negligible, and hence

$$V(F)^{-1/2} \sqrt{n} (\widetilde{\theta}^* - \widehat{\theta}) \rightsquigarrow^* N(0, \Omega(F)) \tag{A.39}$$

uniformly over F.

Step 4: Uniform integrability and convergence of the bootstrap covariance.

Weak convergence alone does not imply convergence of second moments. As emphasized by Hansen (2022, Sections 10.13 and 10.28), consistency of bootstrap variance estimation therefore requires control of the moments of the bootstrap estimator. The argument here parallels Hansen's: the eigenvalue trimming bounds the inverse design matrix, reducing the required moment control to that of a bootstrap sample mean.

On every retained bootstrap replication,

$$\|(\widehat{Q}_j^*)^{-1}\| \leq \frac{1}{\tau \lambda_{\min}(\widehat{Q}_j)}. \tag{A.40}$$

By (A.3), with probability approaching one uniformly over F,

$$\|(\widehat{Q}_j^*)^{-1}\| \leq \frac{2}{\tau\epsilon_b}. \tag{A.41}$$

On a trimmed replication, $\widetilde{\theta}^* = \widehat{\theta}$, so the corresponding bootstrap difference is zero.

Let $p = 2 + \delta_0$ and

$$v_i = \frac{d_{ji} X_{ji} \widehat{\varepsilon}_{ji}}{\sigma_j}, \qquad i = 1, \ldots, n.$$

By (A.21), the c_r inequality, (A.26) and (A.28),

$$E_n[\|v_i\|_2^p] = O_{P_F}(1) \tag{A.42}$$

uniformly over F; this is (A.29) without the factor $\|\widehat{Q}_j^{-1}\|$. Conditional on the original sample,

$$\sum_{i=1}^n (N_i^* - 1) v_i = \sum_{b=1}^n v_b^*,$$

where $v_1^*, \ldots, v_n^*$ are i.i.d. draws from the empirical distribution of $\{v_i\}_{i=1}^n$, which has mean zero by (A.7). Rosenthal's inequality for sums of independent mean-zero random variables, applied to each coordinate and combined with the c_r inequality, therefore gives

$$E^*\left[\left\|\frac{S_j^*}{\sigma_j}\right\|_2^p\right] = E^*\left[\left\|\frac{1}{\sqrt{n}}\sum_{b=1}^n v_b^*\right\|_2^p\right] \leq C_p\left\{\left(E_n\|v_i\|_2^2\right)^{p/2} + n^{1-p/2} E_n\|v_i\|_2^p\right\} = O_{P_F}(1) \tag{A.43}$$

uniformly over F, by (A.42).

On a retained bootstrap replication, (A.8) and (A.41) give

$$\frac{\left|\sqrt{n}(\widehat{\theta}_j^* - \widehat{\theta}_j)\right|}{\sigma_j} = \frac{\left|e_{1j}'(\widehat{Q}_j^*)^{-1} S_j^*\right|}{\sigma_j} \leq \|(\widehat{Q}_j^*)^{-1}\| \left\|\frac{S_j^*}{\sigma_j}\right\|_2 \leq \frac{2}{\tau\epsilon_b}\left\|\frac{S_j^*}{\sigma_j}\right\|_2$$

with probability approaching one uniformly over F; on a rejected replication, $\widetilde{\theta}^* = \widehat{\theta}$ and the left-hand side is zero. Hence

$$E^*\left[\left|\frac{\sqrt{n}(\widetilde{\theta}_j^* - \widehat{\theta}_j)}{\sigma_j}\right|^p\right] \leq \left(\frac{2}{\tau\epsilon_b}\right)^p E^*\left[\left\|\frac{S_j^*}{\sigma_j}\right\|_2^p\right] = O_{P_F}(1)$$

uniformly over F, and since m is fixed,

$$E^*\left[\left\|V(F)^{-1/2}\sqrt{n}(\widetilde{\theta}^* - \widehat{\theta})\right\|_2^{2+\delta_0}\right] = O_{P_F}(1) \tag{A.44}$$

uniformly over F.

This is the analogue in the present setting of Hansen's argument following equation (10.34): the design trimming bounds the moments of the bootstrap OLS estimator by the moments of a bootstrap sample mean. Hansen then uses uniform square integrability to obtain consistency of the bootstrap variance.

Since $2 + \delta_0 > 2$, (A.44) implies uniform square integrability. Combining this with the conditional weak convergence in (A.39) therefore yields convergence of the first two conditional moments:

$$E^*\left[V(F)^{-1/2}\sqrt{n}(\widetilde{\theta}^* - \widehat{\theta})\right] = o_{P_F}(1) \tag{A.45}$$

and

$$E^*\left[V(F)^{-1/2}n(\widetilde{\theta}^* - \widehat{\theta})(\widetilde{\theta}^* - \widehat{\theta})'V(F)^{-1/2}\right] = \Omega(F) + o_{P_F}(1) \tag{A.46}$$

uniformly over F.

Since

$$\widehat{\Sigma}^* = \mathrm{Var}^*\left(\sqrt{n}(\widetilde{\theta}^* - \widehat{\theta})\right),$$

(A.45)–(A.46) imply

$$V(F)^{-1/2}\widehat{\Sigma}^*V(F)^{-1/2} - \Omega(F) = o_{P_F}(1) \tag{A.47}$$

uniformly over F. Equivalently, for every $\eta > 0$,

$$\lim_{n\to\infty}\sup_{F\in\mathcal{F}} P_F\left(\left\|V(F)^{-1/2}\widehat{\Sigma}^*V(F)^{-1/2} - \Omega(F)\right\| > \eta\right) = 0.$$

Step 5: Finite number of bootstrap replications.

For fixed n, conditional on the original sample, the vectors

$$S_n^{*(b)} = \sqrt{n}(\widetilde{\theta}^{*(b)} - \widehat{\theta}), \qquad b = 1, \ldots, B,$$

are i.i.d. bootstrap draws. On retained draws, the eigenvalue criterion bounds each inverse bootstrap design matrix; on rejected draws, $S_n^{*(b)} = 0$. Since the empirical distribution has finite support, the conditional second moments are therefore finite. Consequently, the

conditional law of large numbers gives

$$\frac{1}{B}\sum_{b=1}^{B} S_n^{*(b)} \xrightarrow{p^*} E^*[S_n^*]$$

and

$$\frac{1}{B}\sum_{b=1}^{B} S_n^{*(b)} S_n^{*(b)\prime} \xrightarrow{p^*} E^*[S_n^* S_n^{*\prime}].$$

It follows that

$$\widehat{\Sigma}_B^* \xrightarrow{p^*} \mathrm{Var}^*(S_n^*) = \widehat{\Sigma}^*.$$

This completes the proof. □

A.7 Standardized Robustness Tolerance

Let

$$\hat{p}_0 = E_n[d_{0i}], \qquad \hat{\mu}_D = \frac{E_n[d_{0i}D_{0i}]}{\hat{p}_0}, \qquad \hat{\mu}_y = \frac{E_n[d_{0i}y_{0i}]}{\hat{p}_0},$$

$$\hat{s}_D^2 = \frac{E_n[d_{0i}D_{0i}^2]}{\hat{p}_0} - \hat{\mu}_D^2, \qquad \hat{s}_y^2 = \frac{E_n[d_{0i}y_{0i}^2]}{\hat{p}_0} - \hat{\mu}_y^2,$$

the sample means and variances of D_{0i} and y_{0i} over the observations used in the main specification, and let $\hat{a} = \hat{s}_D/\hat{s}_y$ and $\hat{\theta}^s = \hat{a}\,\hat{\theta}$.

Recall from Section 5.1 that $s_D(F)$ and $s_y(F)$ are the standard deviations of D_{0i} and y_{0i} conditional on $d_{0i} = 1$, and that $\theta^s(F) = \{s_D(F)/s_y(F)\}\theta(F)$. Write $p_0(F) = E_F[d_{0i}]$, $\mu_D(F) = E_F[d_{0i}D_{0i}]/p_0(F)$ and $\mu_y(F) = E_F[d_{0i}y_{0i}]/p_0(F)$, and define the influence functions

$$\psi_{D,i}(F) = \frac{d_{0i}}{2p_0(F)s_D(F)}\left\{(D_{0i} - \mu_D(F))^2 - s_D(F)^2\right\}, \qquad \psi_{y,i}(F) = \frac{d_{0i}}{2p_0(F)s_y(F)}\left\{(y_{0i} - \mu_y(F))^2 - s_y(F\right.$$

$$\psi_{a,i}(F) = \frac{1}{s_y(F)}\psi_{D,i}(F) - \frac{s_D(F)}{s_y(F)^2}\psi_{y,i}(F),$$

of $\hat{s}_D$, $\hat{s}_y$ and $\hat{a}$, respectively. Let $\psi_i(F)$ be the influence function of $\hat{\theta}$ from the proof of Proposition 3, with jth element $d_{ji}u_{ji}\varepsilon_{ji}/E_F[d_{ji}u_{ji}^2]$. Lemma 5(a) below shows that the influence function of $\hat{\theta}^s$ is $\psi_i^s(F) = a(F)\psi_i(F) + \theta(F)\psi_{a,i}(F)$, with $a(F) = s_D(F)/s_y(F)$.

The covariance estimator for the standardized tolerance is the plug-in estimator built from the estimated influence function of $\hat{\theta}^s$, in the same way as (13). Let $\hat{\psi}_{D,i}$, $\hat{\psi}_{y,i}$ and $\hat{\psi}_{a,i}$

denote $\psi_{D,i}(F)$, $\psi_{y,i}(F)$ and $\psi_{a,i}(F)$ with $(p_0, \mu_D, \mu_y, s_D, s_y)$ replaced by $(\hat{p}_0, \hat{\mu}_D, \hat{\mu}_y, \hat{s}_D, \hat{s}_y)$, let $\hat{\psi}_i$ be the estimated influence function of $\hat{\theta}$ in (12), and define

$$\hat{\psi}_i^s = \hat{a}\,\hat{\psi}_i + \hat{\theta}\,\hat{\psi}_{a,i}, \qquad \widehat{\Sigma}^s = E_n\left[\hat{\psi}_i^s \hat{\psi}_i^{s\prime}\right]. \tag{23}$$

Assumption 4 (Standardized tolerance). Assumption 1 holds, and there exist constants $\epsilon > 0$ and $K < \infty$ such that, uniformly over $F \in \mathcal{F}$:

(a) $s_D(F) \geq \epsilon$, $s_y(F) \geq \epsilon$, $\|\theta(F)\|_2 \leq K$, and $\sigma_j^2 \leq K$ for every $j = 0, \ldots, m$.

(b) $E_F[d_{0i}|D_{0i}|^{4+\delta}] \leq K$ and $E_F[d_{0i}|y_{0i}|^{4+\delta}] \leq K$, with δ the constant of Assumption 1.

(c) $\lambda_{\min}\big(\mathrm{Var}_F\left((\psi_i(F)', \psi_{a,i}(F))'\right)\big) \geq \epsilon$.

The upper bound on σ_j^2 in (a) is needed because the standardized estimator is analyzed on its original scale rather than after standardization by $V(F)$.

Corollary 2 (Standardized robustness tolerance). Suppose Assumption 4 holds, and let $\phi_n^s(\kappa, \alpha)$ be the RCC test of $H_0^s(\kappa) : A\theta^s(F) \leq \kappa \mathbf{1}_{2m}$ constructed from $\hat{\theta}^s$ and the plug-in covariance estimator $\widehat{\Sigma}^s$ in (23). Then, for $\alpha \in (0, 1/2)$:

(a)

$$\limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \sup_{\substack{\kappa\geq 0: \\ \max_{j=1,\ldots,m} |\theta_j^s(F) - \theta_0^s(F)| \leq \kappa}} E_F[\phi_n^s(\kappa, \alpha)] \leq \alpha.$$

(b) For every fixed $F \in \mathcal{F}$, with $\kappa_F = \max_{j=1,\ldots,m} |\theta_j^s(F) - \theta_0^s(F)|$,

$$\lim_{n\to\infty} E_F[\phi_n^s(\kappa_F, \alpha)] = \alpha.$$

The bootstrap of Appendix A.6 can also be applied to the standardized estimator. In each bootstrap replication, $\hat{p}_0^{*(b)}$, $\hat{s}_D^{*(b)}$ and $\hat{s}_y^{*(b)}$ are recomputed from the same bootstrap sample and $\hat{\theta}^{s,*(b)} = (\hat{s}_D^{*(b)}/\hat{s}_y^{*(b)})\hat{\theta}^{*(b)}$. The standardized estimator introduces one new denominator, $\hat{s}_y^{*(b)}$, and the trimming rule is extended accordingly: a replication is retained only if, in addition to the design criterion of Algorithm 1,

$$\hat{p}_0^{*(b)} \geq \tau\,\hat{p}_0 \qquad \text{and} \qquad \hat{s}_y^{*(b)2} \geq \tau\,\hat{s}_y^2;$$

otherwise $\widetilde{\theta}^{s,*(b)} = \hat{\theta}^s$. The covariance estimator $\widehat{\Sigma}_B^{s*}$ is then obtained by replacing $S_n^{*(b)}$ in Algorithm 1 with $S_n^{s,*(b)} = \sqrt{n}(\widetilde{\theta}^{s,*(b)} - \hat{\theta}^s)$. Parts (c) and (d) of Lemma 5 show that,

under the additional Assumption 3, this estimator is also uniformly consistent, so that Corollary 2 holds with it in place of (23).

The following elementary fact is used to control higher-order bootstrap moments; it is a uniform version of the Marcinkiewicz weak law of large numbers in Hansen (2022, Theorem 10.20).

Lemma 4. Let $u_1, \dots, u_n$ be i.i.d. under F with $\sup_{F \in \mathcal{F}} E_F|u_i|^{1+\eta} \leq K$ for some $\eta > 0$ and $K < \infty$. Then, for every $s > 1$ and $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup_{F \in \mathcal{F}} P_F \left(n^{-s} \sum_{i=1}^{n} |u_i|^s > \varepsilon \right) = 0.$$

Proof. Since $|u_i|^s \leq (\max_{k \leq n} |u_k|)^{s-1} |u_i|$,

$$n^{-s} \sum_{i=1}^{n} |u_i|^s \leq \left(n^{-1} \max_{k \leq n} |u_k| \right)^{s-1} \cdot n^{-1} \sum_{i=1}^{n} |u_i|.$$

By Markov's inequality and Jensen's inequality, $P_F(n^{-1} \sum_i |u_i| > M) \leq K^{1/(1+\eta)}/M$ for every $M > 0$, uniformly over F. By the union bound and Markov's inequality,

$$P_F \left(\max_{k \leq n} |u_k| > \varepsilon n \right) \leq n \, P_F \left(|u_1| > \varepsilon n \right) \leq \frac{n \, E_F|u_1|^{1+\eta}}{(\varepsilon n)^{1+\eta}} \leq \frac{K}{\varepsilon^{1+\eta}} \, n^{-\eta} \longrightarrow 0$$

uniformly over F. The product of a uniformly $o_{P_F}(1)$ factor and a uniformly $O_{P_F}(1)$ factor is uniformly $o_{P_F}(1)$. □

Lemma 5. Suppose Assumption 4 holds, and set $\delta_s = \delta/2$. Then:

(a) $\hat{\theta}^s$ admits a uniform asymptotically linear representation,

$$\sqrt{n}\{\hat{\theta}^s - \theta^s(F)\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^s(F) + o_{P_F}(1)$$

uniformly over $F \in \mathcal{F}$, where $\psi_i^s(F) = a(F)\psi_i(F) + \theta(F)\psi_{a,i}(F)$ with $a(F) = s_D(F)/s_y(F)$, $E_F[\psi_i^s(F)] = 0$, and $\sup_{F \in \mathcal{F}} E_F\|\psi_i^s(F)\|_2^{2+\delta_s} < \infty$. Moreover, $\Sigma^s(F) = E_F[\psi_i^s(F)\psi_i^s(F)']$ satisfies $\lambda_{\min}(\Sigma^s(F)) \geq \epsilon \, a(F)^2$ and $\sup_F \lambda_{\max}(\Sigma^s(F)) < \infty$, and $a(F)$ is bounded away from zero and infinity uniformly over F.

(b) The plug-in covariance estimator (23) satisfies, for every $\eta > 0$,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} P_F\left(\|\widehat{\Sigma}^s - \Sigma^s(F)\| > \eta\right) = 0.$$

If, in addition, Assumption 3 holds:

(c) Let $\widetilde{\theta}^{s*}$ be the trimmed bootstrap estimator defined above and $\widehat{\Sigma}^{s*} = \mathrm{Var}^*(\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s))$. Then, for every $\eta > 0$,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} P_F\left(\|\widehat{\Sigma}^{s*} - \Sigma^s(F)\| > \eta\right) = 0.$$

(d) For fixed n, as $B \to \infty$, $\widehat{\Sigma}^{s*}_B \xrightarrow{p^*} \widehat{\Sigma}^{s*}$.

Proof. Throughout, $O_{P_F}(\cdot)$ and $o_{P_F}(\cdot)$ statements hold uniformly over $F \in \mathcal{F}$, and bootstrap statements hold in P_F-probability uniformly over F, as in the proof of Proposition 5. We suppress the argument F of population quantities where no confusion arises.

Step 1: Preliminaries. Write $p_0 = E_F[d_{0i}]$. Since u_{0i} is nonzero only when $d_{0i} = 1$, Cauchy–Schwarz and Assumption 1 give

$$\epsilon \le E_F[d_{0i}u_{0i}^2] \le p_0^{1/2}\left(E_F[d_{0i}u_{0i}^4]\right)^{1/2} \le p_0^{1/2}K^{1/2},$$

so $p_0 \ge \underline{p} := \epsilon^2/K$ uniformly over F. Let

$$R_i = \left(d_{0i},\, d_{0i}D_{0i},\, d_{0i}D_{0i}^2,\, d_{0i}y_{0i},\, d_{0i}y_{0i}^2\right)', \qquad \rho(F) = E_F[R_i], \qquad \bar{R} = E_n[R_i].$$

By Assumption 4(b), the components of R_i have uniformly bounded moments of order $2 + \delta_s$ (the highest-order components are $d_{0i}D_{0i}^2$ and $d_{0i}y_{0i}^2$, whose $(2+\delta/2)$th moments are bounded by the $(4+\delta)$th moments of D_{0i} and y_{0i}). Hence $\sup_F E_F\|R_i\|_2^{2+\delta_s} < \infty$, and by Chebyshev's inequality

$$\bar{R} - \rho(F) = O_{P_F}(n^{-1/2}). \tag{B.1}$$

The quantities p_0, s_D^2, s_y^2 and $a = s_D/s_y$ are the values at $\rho(F)$ of the maps

$$h_p(r) = r_1, \qquad h_D(r) = \frac{r_3}{r_1} - \left(\frac{r_2}{r_1}\right)^2, \qquad h_y(r) = \frac{r_5}{r_1} - \left(\frac{r_4}{r_1}\right)^2, \qquad h_a(r) = \left(\frac{h_D(r)}{h_y(r)}\right)^{1/2},$$

and $\hat{p}_0$, $\hat{s}_D^2$, $\hat{s}_y^2$, $\hat{a}$ are their values at $\bar{R}$. Let

$$\mathcal{N}(F) = \{r : \|r - \rho(F)\|_2 \le \varrho\},$$

where $\varrho > 0$ is chosen, using $p_0 \ge \underline{p}$, $s_D \ge \epsilon$, $s_y \ge \epsilon$ and the uniform upper bound on $\|\rho(F)\|_2$ implied by (b), so that on $\mathcal{N}(F)$ one has $h_p \ge \underline{p}/2$, $h_D \ge \epsilon^2/2$, $h_y \ge \epsilon^2/2$, and the first and second derivatives of h_D, h_y and h_a are bounded by a constant M that does not depend on F. By (B.1), $P_F(\bar{R} \in \mathcal{N}(F)) \to 1$ uniformly over F.

Step 2: Part (a). On the event $\bar{R} \in \mathcal{N}(F)$, a second-order Taylor expansion of h_a around $\rho(F)$ and (B.1) give

$$\sqrt{n}(\hat{a} - a) = \nabla h_a(\rho)' \sqrt{n}(\bar{R} - \rho) + O_{P_F}(n^{-1/2}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{a,i}(F) + o_{P_F}(1), \tag{B.2}$$

where $\psi_{a,i}(F) = \nabla h_a(\rho)'(R_i - \rho)$. A direct computation of ∇h_D and ∇h_y shows that $\nabla h_D(\rho)'(R_i - \rho) = 2 s_D \psi_{D,i}$ and $\nabla h_y(\rho)'(R_i - \rho) = 2 s_y \psi_{y,i}$, with $\psi_{D,i}$ and $\psi_{y,i}$ as defined before Corollary 2, and the chain rule then gives $\psi_{a,i} = \psi_{D,i}/s_y - s_D \psi_{y,i}/s_y^2$, as stated there.

By the proof of Proposition 3, $\hat{\theta} - \theta = n^{-1} \sum_i \psi_i(F) + R_n(F)$ with $\sqrt{n}\, V(F)^{-1/2} R_n(F) = o_{P_F}(1)$; since the diagonal entries of $V(F)$ are $\sigma_j^2 \le K$ by (a), also $\sqrt{n}\, R_n(F) = o_{P_F}(1)$. Hence

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(F) + o_{P_F}(1), \qquad \hat{\theta} - \theta = O_{P_F}(n^{-1/2}). \tag{B.3}$$

Now $\hat{\theta}^s - \theta^s = \hat{a}\hat{\theta} - a\theta = a(\hat{\theta} - \theta) + (\hat{a} - a)\theta + (\hat{a} - a)(\hat{\theta} - \theta)$, and the last term is $O_{P_F}(n^{-1})$ by (B.2)–(B.3). Combining (B.2) and (B.3),

$$\sqrt{n}(\hat{\theta}^s - \theta^s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{a\, \psi_i(F) + \theta\, \psi_{a,i}(F)\} + o_{P_F}(1),$$

which is the representation in part (a) with $\psi_i^s = a\psi_i + \theta\psi_{a,i}$; $E_F[\psi_i^s] = 0$ because $E_F[\psi_i] = 0$ and $E_F[R_i - \rho] = 0$.

Moments and scale. Since $s_D \le (E_F[d_{0i} D_{0i}^2]/p_0)^{1/2}$ and similarly for s_y, (b) and $p_0 \ge \underline{p}$ give $s_D, s_y \le \bar{s}$ for a constant $\bar{s}$; with the lower bounds in (a), $a(F) \in [\epsilon/\bar{s}, \bar{s}/\epsilon]$ uniformly over F, and $\|\nabla h_a(\rho)\|_2 \le M$. Hence $E_F|\psi_{a,i}|^{2+\delta_s} \le M^{2+\delta_s} E_F\|R_i - \rho\|_2^{2+\delta_s}$ is uniformly bounded. By Assumption 1(d) and $\sigma_j^2 \le K$, $E_F|\psi_{ji}|^{2+\delta} = \sigma_j^{2+\delta} E_F|\psi_{ji}/\sigma_j|^{2+\delta} \le K^{1+\delta/2} K$ is uniformly bounded. Since $\|\theta\|_2 \le K$ and m is fixed, the c_r inequality gives

$\sup_F E_F\|\psi_i^s\|_2^{2+\delta_s} < \infty$, and in particular $\sup_F \lambda_{\max}(\Sigma^s(F)) < \infty$. For the lower bound, write $\psi_i^s = \begin{bmatrix} aI_{m+1} & \theta \end{bmatrix}(\psi_i', \psi_{a,i})'$, so that for every $x \in \mathbb{R}^{m+1}$, by (c),

$$x'\Sigma^s(F)x \geq \epsilon \left\| \begin{pmatrix} ax \\ \theta' x \end{pmatrix} \right\|_2^2 = \epsilon \left\{ a^2\|x\|_2^2 + (\theta' x)^2 \right\} \geq \epsilon\, a^2 \|x\|_2^2.$$

This proves part (a).

Step 3: Part (b). Write $\hat{\psi}_i^s = [\hat{a} I_{m+1} \;\; \hat{\theta}]\,(\hat{\psi}_i', \hat{\psi}_{a,i})'$ and $\psi_i^s = [aI_{m+1} \;\; \theta]\,(\psi_i', \psi_{a,i})'$. Since $\hat{a} \to a$ and $\hat{\theta} \to \theta$ in P_F-probability uniformly over F by (B.2)–(B.3), with a and $\|\theta\|_2$ uniformly bounded, it suffices to show that

$$E_n\left[\begin{pmatrix} \hat{\psi}_i \\ \hat{\psi}_{a,i} \end{pmatrix} \begin{pmatrix} \hat{\psi}_i \\ \hat{\psi}_{a,i} \end{pmatrix}' \right] - E_F\left[\begin{pmatrix} \psi_i \\ \psi_{a,i} \end{pmatrix} \begin{pmatrix} \psi_i \\ \psi_{a,i} \end{pmatrix}' \right] = o_{P_F}(1). \tag{B.4}$$

Three facts are used. First, by Lemma 3 and $\sigma_j^2 \leq K$,

$$E_n[\hat{\psi}_i \hat{\psi}_i'] - \Sigma(F) = V(F)^{1/2} \left\{ V(F)^{-1/2} E_n[\hat{\psi}_i \hat{\psi}_i'] V(F)^{-1/2} - \Omega(F) \right\} V(F)^{1/2} = o_{P_F}(1),$$

and, by (P.4) and Step 2 of the proof of that lemma, $E_n\|\hat{\psi}_i - \psi_i\|_2^2 = o_{P_F}(1)$. Second, the identity $\nabla h_D(r)'(R_i - r) = d_{0i}\{(D_{0i} - r_2/r_1)^2 - h_D(r)\}/r_1$ holds for every r with $r_1 > 0$ (and likewise for h_y), so that $\hat{\psi}_{a,i} = \nabla h_a(\bar{R})'(R_i - \bar{R})$ and $\psi_{a,i} = \nabla h_a(\rho)'(R_i - \rho)$. On the event $\bar{R} \in \mathcal{N}(F)$, the bound M on the second derivatives of h_a gives $\|\nabla h_a(\bar{R}) - \nabla h_a(\rho)\|_2 \leq M\|\bar{R} - \rho\|_2$, hence

$$E_n[(\hat{\psi}_{a,i} - \psi_{a,i})^2] \leq 2M^2\|\bar{R} - \rho\|_2^2 E_n\|R_i - \rho\|_2^2 + 2M^2\|\bar{R} - \rho\|_2^2 = O_{P_F}(n^{-1}),$$

by (B.1) and $E_n\|R_i - \rho\|_2^2 = O_{P_F}(1)$. Third, the entries of $(\psi_i', \psi_{a,i})'(\psi_i', \psi_{a,i})$ have uniformly bounded moments of order $1 + \delta_s/2$, by the Cauchy–Schwarz inequality and the $(2+\delta_s)$th moment bounds of Step 2, so the uniform weak law of large numbers used in Step 3 of the proof of Lemma 3 gives $E_n[(\psi_i', \psi_{a,i})'(\psi_i', \psi_{a,i})] - E_F[(\psi_i', \psi_{a,i})'(\psi_i', \psi_{a,i})] = o_{P_F}(1)$. Expanding each block of the left-hand side of (B.4) as $E_n[\hat{x}_i\hat{y}_i'] - E_F[x_i y_i'] = E_n[(\hat{x}_i - x_i)\hat{y}_i'] + E_n[x_i(\hat{y}_i - y_i)'] + (E_n[x_i y_i'] - E_F[x_i y_i'])$ and applying the Cauchy–Schwarz inequality to the first two terms, with $E_n\|\hat{\psi}_i\|_2^2$, $E_n[\hat{\psi}_{a,i}^2]$, $E_n\|\psi_i\|_2^2$ and $E_n[\psi_{a,i}^2]$ all $O_{P_F}(1)$, gives (B.4). This proves part (b).

Step 4: The retention event is asymptotically negligible. Let $\bar{R}^* = E_n^*[R_i^*]$ denote the

bootstrap sample mean, so that $\hat{p}_0^* = h_p(\bar{R}^*)$, $\hat{s}_D^{*2} = h_D(\bar{R}^*)$ and $\hat{s}_y^{*2} = h_y(\bar{R}^*)$. Conditional on the data, $\bar{R}^* - \bar{R}$ is the mean of n i.i.d. draws from the empirical distribution of $R_i - \bar{R}$, so

$$E^*\|\bar{R}^* - \bar{R}\|_2^2 = \frac{1}{n} E_n\|R_i - \bar{R}\|_2^2 = O_{P_F}(n^{-1}), \tag{B.5}$$

and $\bar{R}^* - \bar{R} = o_{P^*}(1)$. On the event $\bar{R} \in \mathcal{N}(F)$, continuity of h_p and h_y at $\bar{R}$ gives $\hat{p}_0^*/\hat{p}_0 \to 1$ and $\hat{s}_y^{*2}/\hat{s}_y^2 \to 1$ in P^*-probability. Together with (A.6), for every fixed $\tau \in (0,1)$ and $\eta > 0$,

$$\sup_{F \in \mathcal{F}} P_F\left[P^*\left(\text{replication rejected}\right) > \eta\right] \longrightarrow 0. \tag{B.6}$$

Step 5: Decomposition on retained replications. Write $\hat{a}^* = \hat{s}_D^*/\hat{s}_y^*$. On a retained replication,

$$S_n^{s*} := \sqrt{n}(\hat{\theta}^{s*} - \hat{\theta}^s) = \hat{a}^*\sqrt{n}(\hat{\theta}^* - \hat{\theta}) + \sqrt{n}(\hat{a}^* - \hat{a})\,\hat{\theta} =: T_1^* + T_2^*; \tag{B.7}$$

on a rejected replication, $\widetilde{\theta}^{s*} = \hat{\theta}^s$ and $\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s) = 0$. We bound the bootstrap moments of T_1^* and T_2^* on retained replications. From here on Assumption 3 is in force, and without loss of generality $\delta \le \delta_b$, since Assumptions 1(d) and 4(b) remain valid when δ is decreased. Let $p = 2 + \delta_s$. All bounds below are stated on the event $\{\bar{R} \in \mathcal{N}(F)\} \cap \{\min_j \lambda_{\min}(\widehat{Q}_j) \ge \epsilon_b/2\}$, whose P_F-probability tends to one uniformly over F by (B.1) and (A.3).

Two moment facts. First, for any $r \ge 2$, the normalized bootstrap mean of a centered sample satisfies, by Rosenthal's inequality applied coordinatewise as in (A.43),

$$E^*\left\|\sqrt{n}(\bar{R}^* - \bar{R})\right\|_2^r \le C_r\left\{\left(E_n\|R_i - \bar{R}\|_2^2\right)^{r/2} + n^{-r/2}\sum_{i=1}^n \|R_i - \bar{R}\|_2^r\right\} = O_{P_F}(1). \tag{B.8}$$

Here $E_n\|R_i - \bar{R}\|_2^2 = O_{P_F}(1)$, and for the second term the c_r inequality gives $n^{-r/2}\sum_i \|R_i - \bar{R}\|_2^r \le 2^{r-1}\{n^{-r/2}\sum_i \|R_i - \rho\|_2^r + n^{1-r/2}\|\bar{R} - \rho\|_2^r\}$; the first term is $o_{P_F}(1)$ by Lemma 4 applied to $u_i = \|R_i - \rho\|_2^2$ with $\eta = \delta_s/2$ and $s = r/2 > 1$, and the second is $n^{1-r/2}O_{P_F}(n^{-r/2}) = o_{P_F}(1)$ by (B.1). Second, the same argument applied to the OLS bootstrap score S_j^* of (A.8), with $v_i = d_{ji}X_{ji}\hat{\varepsilon}_{ji}/\sigma_j$ as in Step 4 of the proof of Proposition 5, gives for any $r \ge 2$

$$E^*\|S_j^*/\sigma_j\|_2^r \le C_r\left\{\left(E_n\|v_i\|_2^2\right)^{r/2} + n^{-r/2}\sum_{i=1}^n \|v_i\|_2^r\right\} = O_{P_F}(1). \tag{B.9}$$

Indeed, by (A.21) and the c_r inequality,

$$n^{-r/2}\sum_{i=1}^{n}\|v_i\|_2^r \le 2^{r-1}\left\{n^{-r/2}\sum_{i=1}^{n}\left\|\frac{d_{ji}X_{ji}\varepsilon_{ji}}{\sigma_j}\right\|_2^r + \left(\frac{\sqrt{n}\,\|\hat\beta_j-\beta_j\|_2}{\sigma_j}\right)^r n^{-r}\sum_{i=1}^{n}\left(d_{ji}\|X_{ji}\|_2^2\right)^r\right\}.$$

The first term is $o_{P_F}(1)$ by Lemma 4 with $u_i = \|d_{ji}X_{ji}\varepsilon_{ji}/\sigma_j\|_2^2$, $\eta = \delta_b/2$ and $s = r/2$ (Assumption 3(b)); in the second, the first factor is $O_{P_F}(1)$ by (A.24) and the second is $o_{P_F}(1)$ by Lemma 4 with $u_i = d_{ji}\|X_{ji}\|_2^2$, $\eta = 1+\delta_b/2$ and $s = r$. Consequently, by (A.8) and (A.41), on retained replications

$$E^*\left[\|\sqrt{n}(\hat\theta^*-\hat\theta)\|_2^r\,\mathbf{1}\{\text{retained}\}\right] \le (m+1)^{r/2}\left(\frac{2}{\tau\epsilon_b}\right)^r \max_j \sigma_j^r\, E^*\|S_j^*/\sigma_j\|_2^r = O_{P_F}(1) \tag{B.10}$$

for any $r \ge 2$, using $\|x\|_2^r \le (m+1)^{r/2}\max_j|x_j|^r$ for $x \in \mathbb{R}^{m+1}$ and $\sigma_j^2 \le K$.

Term T_1^*. On a retained replication, $\hat s_y^* \ge \sqrt{\tau}\,\hat s_y \ge \sqrt{\tau}\,\epsilon/\sqrt{2}$ and $\hat p_0^* \ge \tau\hat p_0 \ge \tau\underline{p}/2$, so

$$0 \le \hat a^* = \frac{\hat s_D^*}{\hat s_y^*} \le \frac{\sqrt{2}}{\sqrt{\tau}\,\epsilon}\left(\frac{E_n^*[d_{0i}D_{0i}^2]}{\hat p_0^*}\right)^{1/2} \le C_\tau\left(E_n^*[d_{0i}D_{0i}^2]\right)^{1/2}, \qquad C_\tau = \frac{2}{\tau\,\epsilon\,\underline{p}^{1/2}},$$

where we used $\hat s_D^{*2} \le E_n^*[d_{0i}D_{0i}^2]/\hat p_0^*$. By the Cauchy–Schwarz inequality under E^*,

$$E^*\left[\|T_1^*\|_2^p\,\mathbf{1}\{\text{retained}\}\right] \le C_\tau^p\left(E^*\left[\left(E_n^*[d_{0i}D_{0i}^2]\right)^p\right]\right)^{1/2}\left(E^*\left[\|\sqrt{n}(\hat\theta^*-\hat\theta)\|_2^{2p}\,\mathbf{1}\{\text{retained}\}\right]\right)^{1/2}.$$

By Jensen's inequality, $E^*[(E_n^*[d_{0i}D_{0i}^2])^p] \le E^*[E_n^*[(d_{0i}D_{0i}^2)^p]] = E_n[d_{0i}|D_{0i}|^{2p}]$, and $2p = 4+\delta$, so this is $O_{P_F}(1)$ by (b). The second factor is $O_{P_F}(1)$ by (B.10) with $r = 2p$. Hence

$$E^*\left[\|T_1^*\|_2^p\,\mathbf{1}\{\text{retained}\}\right] = O_{P_F}(1). \tag{B.11}$$

Term T_2^*. Write $\Delta^* = \bar R^* - \bar R$. On a retained replication,

$$\hat a^* - \hat a = \frac{(\hat s_D^* - \hat s_D)\hat s_y - \hat s_D(\hat s_y^* - \hat s_y)}{\hat s_y^*\hat s_y}, \qquad |\hat s_D^* - \hat s_D| = \frac{|h_D(\bar R^*) - h_D(\bar R)|}{\hat s_D^* + \hat s_D} \le \frac{|h_D(\bar R^*) - h_D(\bar R)|}{\hat s_D},$$

and likewise $|\hat s_y^* - \hat s_y| \le |h_y(\bar R^*) - h_y(\bar R)|/\hat s_y$, where we used that $h_D(\bar R^*) = \hat s_D^{*2} \ge 0$ is a sample variance; on the data event, $\hat s_D, \hat s_y \ge \epsilon/\sqrt{2}$, $\hat s_D, \hat s_y \le \bar s$, and on retained replications $\hat s_y^*\hat s_y \ge \sqrt{\tau}\,\hat s_y^2 \ge \sqrt{\tau}\,\epsilon^2/2$. To bound $h_D(\bar R^*) - h_D(\bar R)$, write $r = \bar R$, $r^* = \bar R^*$, so that on retained replications $r_1^* \ge \tau r_1$, and on the data event $r_1 \ge \underline{p}/2$ and $\|r\|_2 \le \bar\rho$, where $\bar\rho$ is a

uniform bound on $\|\rho(F)\|_2 + \varrho$. Then

$$\frac{r_3^*}{r_1^*} - \frac{r_3}{r_1} = \frac{\Delta_3^*}{r_1^*} - \frac{r_3\,\Delta_1^*}{r_1^* r_1}, \qquad \left|\frac{r_3^*}{r_1^*} - \frac{r_3}{r_1}\right| \le c_1\|\Delta^*\|_2, \qquad c_1 = \frac{2}{\tau\underline{p}} + \frac{4\bar{\rho}}{\tau\underline{p}^2},$$

and, with $q = r_2/r_1$ and $q^* = r_2^*/r_1^*$, the same identity gives $|q^* - q| \le c_1\|\Delta^*\|_2$, while $(q^*)^2 - q^2 = (q^* - q)\{2q + (q^* - q)\}$ with $|2q| \le c_2 := 4\bar{\rho}/\underline{p}$. Hence

$$|h_D(\bar{R}^*) - h_D(\bar{R})| \le c_1(1 + c_2)\|\Delta^*\|_2 + c_1^2\|\Delta^*\|_2^2,$$

and identically for h_y with (r_4, r_5) in place of (r_2, r_3). Collecting the constants, there is $C'_\tau < \infty$ depending only on τ, $\underline{p}$, ϵ, $\bar{\rho}$ and $\bar{s}$ such that, on retained replications,

$$|\hat{a}^* - \hat{a}| \le C'_\tau \left(\|\Delta^*\|_2 + \|\Delta^*\|_2^2\right).$$

Since $\|\hat{\theta}\|_2 = O_{P_F}(1)$ by (B.3) and $\|\theta\|_2 \le K$,

$$\|T_2^*\|_2 \le C''_\tau \left(\|\sqrt{n}\Delta^*\|_2 + n^{-1/2}\|\sqrt{n}\Delta^*\|_2^2\right)$$

on retained replications, with $C''_\tau = O_{P_F}(1)$. By the c_r inequality and (B.8) with $r = p$ and $r = 2p$,

$$E^*\left[\|T_2^*\|_2^p\,\mathbf{1}\{\text{retained}\}\right] \le C\left\{E^*\|\sqrt{n}\Delta^*\|_2^p + n^{-p/2}E^*\|\sqrt{n}\Delta^*\|_2^{2p}\right\} = O_{P_F}(1). \tag{B.12}$$

Conclusion of Step 5. By (B.7), (B.11), (B.12) and the c_r inequality,

$$E^*\left\|\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s)\right\|_2^{2+\delta_s} = O_{P_F}(1) \tag{B.13}$$

uniformly over F, so $\|\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s)\|_2^2$ is uniformly integrable under P^*, in P_F-probability.

Step 6: Conditional Gaussian approximation. On retained replications, by (A.12) and $\sigma_j^2 \le K$, $\sqrt{n}(\hat{\theta}^* - \hat{\theta}) = n^{-1/2}\sum_i (N_i^* - 1)\hat{\psi}_i + o_{P^*}(1)$ with $\hat{\psi}_i$ as in (A.9), and $\hat{a}^* = \hat{a} + o_{P^*}(1)$ by Step 4, so $T_1^* = \hat{a}\,n^{-1/2}\sum_i (N_i^* - 1)\hat{\psi}_i + o_{P^*}(1)$, because $n^{-1/2}\sum_i (N_i^* - 1)\hat{\psi}_i = O_{P^*}(1)$ by (A.37). For T_2^*, on the event $\bar{R} \in \mathcal{N}(F)$ a mean-value expansion of h_a between $\bar{R}$ and $\bar{R}^*$, with $\bar{R}^* - \bar{R} = O_{P^*}(n^{-1/2})$ by (B.5) and the bound M on the second derivatives, gives

$$\sqrt{n}(\hat{a}^* - \hat{a}) = \nabla h_a(\bar{R})'\sqrt{n}(\bar{R}^* - \bar{R}) + o_{P^*}(1) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (N_i^* - 1)\hat{\psi}_{a,i} + o_{P^*}(1), \qquad \hat{\psi}_{a,i} = \nabla h_a(\bar{R})'(R_i - \bar{R}).$$

Hence, on retained replications,

$$S_n^{s*} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (N_i^* - 1)\hat{\psi}_i^s + o_{P^*}(1), \qquad \hat{\psi}_i^s = \hat{a}\,\hat{\psi}_i + \hat{\theta}\,\hat{\psi}_{a,i}, \tag{B.14}$$

and $E_n[\hat{\psi}_i^s] = 0$ by (A.7) and $E_n[R_i - \bar{R}] = 0$. By the Frisch–Waugh–Lovell theorem, $\hat{\psi}_i$ in (A.9) coincides with (12), and by Step 3, $\hat{\psi}_{a,i}$ here is the estimated influence function of (23); hence $\hat{\psi}_i^s$ is the vector in (23) and part (b) gives the first statement in (B.15). For the second, $\hat{a}$, $\|\hat{\theta}\|_2$ and $\|\nabla h_a(\bar{R})\|_2$ are $O_{P_F}(1)$, $E_n\|\hat{\psi}_i\|_2^{2+\delta_s} = O_{P_F}(1)$ by (A.30), Jensen's inequality ($\delta_s \le \delta_0$) and $\sigma_j^2 \le K$, and $E_n\|R_i - \bar{R}\|_2^{2+\delta_s} = O_{P_F}(1)$; the c_r inequality then gives

$$E_n[\hat{\psi}_i^s \hat{\psi}_i^{s\prime}] = \Sigma^s(F) + o_{P_F}(1), \qquad E_n\|\hat{\psi}_i^s\|_2^{2+\delta_s} = O_{P_F}(1), \tag{B.15}$$

uniformly over F. Since $\lambda_{\min}(\Sigma^s(F)) \ge \epsilon a(F)^2$ is bounded away from zero by part (a), the conditional Lyapunov condition holds exactly as in (A.34)–(A.36), and the conditional Lyapunov central limit theorem with the Cramér–Wold device gives

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (N_i^* - 1)\hat{\psi}_i^s \rightsquigarrow^* N(0, \Sigma^s(F))$$

uniformly over F, in probability. By (B.14) and (B.6), the same holds for $\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s)$.

Step 7: Parts (c) and (d). Combining the conditional weak convergence of Step 5 with the uniform integrability (B.13), exactly as in (A.45)–(A.47),

$$E^*\left[\sqrt{n}(\widetilde{\theta}^{s*} - \hat{\theta}^s)\right] = o_{P_F}(1), \qquad E^*\left[n(\widetilde{\theta}^{s*} - \hat{\theta}^s)(\widetilde{\theta}^{s*} - \hat{\theta}^s)'\right] = \Sigma^s(F) + o_{P_F}(1),$$

uniformly over F, which is part (c). Part (d) follows as in Step 5 of the proof of Proposition 5: for fixed n the bootstrap draws are i.i.d. with finite conditional second moments (the empirical distribution has finite support, and on retained replications all denominators are bounded below by the trimming rule), so the conditional law of large numbers applies as $B \to \infty$. □

Proof of Corollary 2. We verify Assumption 2 for $\hat{\theta}^s$, $\theta^s(F)$ and $\widehat{\Sigma}^s$ and then apply Proposition 4.

By Lemma 5(a), $\hat{\theta}^s - \theta^s(F) = n^{-1}\sum_i \psi_i^s(F) + R_n^s(F)$ with $\sqrt{n}\,R_n^s(F) = o_{P_F}(1)$ uniformly over F. Let $V^s(F) = \mathrm{diag}(\Sigma^s(F))$ and $\Omega^s(F) = V^s(F)^{-1/2}\Sigma^s(F)V^s(F)^{-1/2}$. By the lemma, $\lambda_{\min}(\Sigma^s(F)) \ge \epsilon a(F)^2 \ge \underline{c}$ and $\lambda_{\max}(\Sigma^s(F)) \le \bar{c}$ for constants $0 < \underline{c} \le \bar{c} < \infty$

not depending on F, so the diagonal entries of $V^s(F)$ lie in $[\underline{c}, \bar{c}]$. Hence: Assumption 2a) holds, since $\sqrt{n}\|V^s(F)^{-1/2}R_n^s(F)\|_2 \leq \underline{c}^{-1/2}\sqrt{n}\|R_n^s(F)\|_2 = o_{P_F}(1)$; Assumption 2b) holds, since $\mathrm{Var}_F(\psi_{ji}^s) \geq \underline{c} > 0$, the $(2+\delta_s)$th moments of $\psi_{ji}^s/\sqrt{\mathrm{Var}_F(\psi_{ji}^s)}$ are bounded by $\underline{c}^{-1-\delta_s/2}\sup_F E_F\|\psi_i^s\|_2^{2+\delta_s} < \infty$, and $\det(\Omega^s(F)) \geq \lambda_{\min}(\Omega^s(F))^{m+1} \geq (\underline{c}/\bar{c})^{m+1} > 0$; and Assumption 2c) holds, since by Lemma 5(b)

$$\left\|V^s(F)^{-1/2}\widehat{\Sigma}^s V^s(F)^{-1/2} - \Omega^s(F)\right\| \leq \underline{c}^{-1}\left\|\widehat{\Sigma}^s - \Sigma^s(F)\right\| = o_{P_F}(1)$$

uniformly over F.

Proposition 4 therefore applies to $\hat{\theta}^s$ with $\mathcal{T} = [0, \infty)$, $A(t) \equiv A$ and $b(t) = t\mathbf{1}_{2m}$, so that $\mathcal{T}_0(F) = \{\kappa \geq 0 : A\theta^s(F) \leq \kappa\mathbf{1}_{2m}\} = \{\kappa \geq 0 : \max_j |\theta_j^s(F) - \theta_0^s(F)| \leq \kappa\}$. Part (a) of the proposition is part (a) of the corollary. For part (b), A has no zero rows, and at $t = \kappa_F$ the binding set is nonempty because the maximum defining κ_F is attained; part (b) of the proposition gives the result. □

A.8 Replication of Massenkoff and Wilmers (2023)

Table 2 replicates Table 2 of Massenkoff and Wilmers (2023), the baseline specifications. Tables 3 and 4 replicate their Tables A.2 and A.3, the alternative weighting schemes and the alternative definitions of flexible pay for specifications (3) and (4). In all three, the number of observations differs slightly from the original because singleton fixed effects are removed in one pass rather than recursively; the estimates show no visible difference.

	(1)	(2)	(3)	(4)	(5)
Nonstandardized Pay	-0.145	-0.108	-0.077	-0.008	-0.010
	(0.003)	(0.003)	(0.002)	(0.001)	(0.004)
log(Workers at Pay Level)	0.044	0.040	0.034	0.013	0.012
	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)
log(Workers in Establishment)	0.050	0.048	0.040	0.014	0.012
	(0.001)	(0.001)	(0.001)	(0.002)	(0.004)
log(Workers in Job)	-0.016	-0.023	-0.012	-0.016	-0.013
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
log(Minimum Wage)	0.085	0.075	0.001	0.010	-0.001
	(0.011)	(0.010)	(0.010)	(0.004)	(0.005)
Collective Bargaining		0.034	0.021	0.002	-0.003
		(0.004)	(0.003)	(0.002)	(0.003)
Share Managerial, Clerical in Establishment		0.018	-0.008	-0.002	-0.018
		(0.006)	(0.005)	(0.004)	(0.011)
Coworkers' Occupational Level		0.556	0.262	0.006	0.002
		(0.008)	(0.006)	(0.004)	(0.009)
Union Density in Industry-Wage Area		0.049	0.030	-0.002	0.008
		(0.005)	(0.004)	(0.002)	(0.006)
Observations	900,359	852,024	829,626	772,809	548,889
R^2	0.12977	0.21547	0.72946	0.96546	0.96128
Within R^2	0.12239	0.20878	0.19237	0.01064	0.00885
Year fixed effects	✓	✓			
Occupation-Year-City-Ind. fixed effects			✓	✓	✓
Occupation-Establishment fixed effects				✓	✓
Firm-Occupation-Year fixed effects					✓

Table 2: Wage Effects of Nonstandardized Pay Rates

	Inv. Row Wt		Survey Wt.		CPS Wt.		N. Employees			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Nonstandardized Pay	-0.077	-0.008	-0.071	-0.011	-0.076	-0.008	-0.070	-0.002	-0.073	-0.006
	(0.002)	(0.001)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Excl. largest 5% of job brackets									✓	✓
Observations	829,626	772,809	829,626	772,809	779,649	726,170	829,626	772,809	785,896	731,742
R^2	0.72946	0.96546	0.71875	0.93325	0.75053	0.96622	0.83179	0.97303	0.77752	0.96441
Within R^2	0.19237	0.01064	0.13896	0.00364	0.19487	0.00897	0.20574	0.01784	0.17861	0.00831
Occupation-Year-City-Ind. fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Occupation-Establishment fixed effects		✓		✓		✓		✓		✓

Table 3: Standardized pay rates effects robustness to alternative weights

	(1)	(2)	(3)	(4)	(5)	(6)
Merit, Combination, or Other	-0.077	-0.008				
	(0.002)	(0.001)				
Merit or Combination			-0.075	-0.009		
			(0.002)	(0.001)		
Merit (Narrow)					-0.064	-0.008
					(0.002)	(0.001)
Controls	✓	✓	✓	✓	✓	✓
Observations	829,626	772,809	829,626	772,809	829,626	772,809
R^2	0.72946	0.96546	0.72851	0.96546	0.72482	0.96546
Within R^2	0.19237	0.01064	0.18951	0.01076	0.17851	0.01051
Occupation-Year-City-Ind. fixed effects	✓	✓	✓	✓	✓	✓
Occupation-Establishment fixed effects		✓		✓		✓

Table 4: Effects robustness to alternative operationalizations of flexible pay-setting

References

Al Mohamad, D., E. W. Van Zwet, E. Cator, and J. J. Goeman (2020): "Adaptive critical value for constrained likelihood ratio testing," *Biometrika*, 107, 677–688.

Altonji, J., T. Elder, and C. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Andrews, I., J. Chen, and O. Tecchio (2026): "The Purpose of an Estimator Is What It Does: Misspecification, Estimands, and Over-Identification," in *Econometric Society 2025 World Congress Volume*, arXiv:2508.13076, https://doi.org/10.48550/arXiv.2508.13076.

Andrews, I., J. Roth, and A. Pakes (2023): "Inference for Linear Conditional Moment Inequalities," *The Review of Economic Studies*, 90, 2763–2791.

Armstrong, T. B. (2025): "Misspecification in Econometrics: A Selective Review," in *Proceedings of the 2025 Econometric Society World Congress*, forthcoming.

Bei, X. (2024): "Inference on Union Bounds," Available at https://xinyuebei-econ.github.io/files/Bei_Xinyue_JMP.pdf.

Breznau, N. et al. (2022): "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty," *Proceedings of the National Academy of Sciences*, 119, e2203150119.

Campbell, D., A. Brodeur, A. Dreber, M. Johannesson, J. Kopecky, L. Lusher, and N. Tsoy (2024): "The Robustness Reproducibility of the American Economic Review," I4R Discussion Paper 124, Institute for Replication (I4R).

Canay, I. A., G. Illanes, and A. Velez (2026): "A User's Guide for Inference in Models Defined by Moment Inequalities," *Journal of Econometrics*, forthcoming.

Canay, I. A. and A. M. Shaikh (2017): "Practical and theoretical advances in inference for partially identified models," *Advances in Economics and Econometrics*, 2, 271–306.

Chernozhukov, V., C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis (2026): "Long Story Short: Omitted Variable Bias in Causal Machine Learning," *The Review of Economics and Statistics*.

Cinelli, C. and C. Hazlett (2020): "Making sense of sensitivity: Extending omitted variable bias," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 82, 39–67.

Cox, G. and X. Shi (2023): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *The Review of Economic Studies*, 90, 201–228.

Diegert, P., M. A. Masten, and A. Poirier (2026a): "Assessing Omitted Variable Bias when the Controls are Endogenous," ArXiv:2206.02303, https://doi.org/10.48550/arXiv.2206.02303.

——— (2026b): "An Axiomatic Approach to Comparing Sensitivity Parameters," ArXiv:2504.21106, https://doi.org/10.48550/arXiv.2504.21106.

Giordano, R., R. Meager, and T. Broderick (2026): "An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Change Conclusions? Part I: Definitions and Experiments," *Philosophical Transactions of the Royal Society A*, 384, 20250001.

Hansen, B. (2022): *Econometrics*, Princeton University Press.

Imbens, G. W. (2003): "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review*, 93, 126–132.

Jaeger, D. A. (2026): "Robustness? Range Tests for Equality and Equivalence Across Specifications," Tech. Rep. 21802, CEPR Press, Paris and London.

Kudô, A. (1963): "A Multivariate Analogue of the One-Sided Test," *Biometrika*, 50, 403–418.

Lu, X. and H. White (2014): "Robustness checks and robustness tests in applied economics," *Journal of Econometrics*, 178, 194–206.

Massenkoff, M. and N. Wilmers (2023): "Wage Stagnation and the Decline of Standardized Pay Rates, 1974–1991," *American Economic Journal: Applied Economics*, 15, 474–507.

Miller, D. L., F. Molinari, and J. Stoye (2025): "Testing Sign Congruence Between Two Parameters," ArXiv:2405.11759, https://doi.org/10.48550/arXiv.2405.11759; also cemmap working paper CWP14/24.

Molinari, F. (2020): "Chapter 5 - Microeconometrics with partial identification," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 355–486.

Oster, E. (2019): "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*, New York: Springer.

Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

Wolak, F. A. (1987): "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793.