Submitted to *Journal of Applied Econometrics*

Treatment effects without multicollinearity? Temporal order and the Gram-Schmidt process in causal inference.

Robin M. Cross and Steven T. Buccola

Department of Applied Economics, Oregon State University, Corvallis, Oregon, USA

Correspondence
Robin M. Cross, 221B Ballard Hall, Corvallis, Oregon 97331.
Email: robin.cross@oregonstate.edu, sbuccola@oregonstate.edu

Summary

This paper incorporates information about the temporal order of regressors to estimate orthogonal and economically interpretable regression coefficients. We establish new finite sample properties for the Gram-Schmidt orthogonalization process. Coefficients are unbiased and stable with lower standard errors than those from Ordinary Least Squares. We provide conditions under which coefficients represent average total treatment effects on the treated and extend the model to groups of ordered and simultaneous regressors. Finally, we reanalyze two studies that controlled for temporally ordered and collinear characteristics, including race, education, and income. The new approach expands Bohren *et al.*'s decomposition of systemic discrimination into channel-specific effects and improves significance levels.

KEYWORDS

collinear regressors, recursive Laplace orthogonalization, QR decomposition, temporal order

JEL codes

C51, D63, I21, I32, J71

1 | INTRODUCTION

Multicollinearity, or correlation among regressors, has posed a problem for statistical analysis since the introduction of Ordinary Least Squares (OLS) in 1805. Inflated standard errors depress statistical significance, and negative coefficient covariance renders models sensitive to small changes in regressor selection or functional form. A growing area of concern has been the emergence of *p*-hacking, the selection of regressors or functional forms to inflate statistical significance (Leamer, 1983, Gelman and Loken, 2014). At the same time, rapid improvements in computing power and the emergence of machine-learning have increased the number of closely related regressors that can be analyzed in a single model (Baiardi and Naghi, 2024). The statistical instability induced by even modest levels of multicollinearity in the data exacerbates its

Data and code for replication of results are available for Stata and R at https://github.com/crossrm/GSLS.

This work was not supported by grants or other outside funding.

We have no conflicts of interest to disclose.

We are grateful to Anthony Nearman, Karen Rennick, Nathalie Steinhaur, and Dennis vanEnglesdorp for the initial motivation to solve the multicollinearity problem, W. Jason Beasley and Aaron Watt for experimental implementation, J. Aislinn Bohren and Peter Hull for comments on discrimination decomposition in survey data, Jennifer Alix-Garcia, David Kling, and David Lewis for input on related research and project scope, and Guido Imbens for encouragement to address temporal order. Juan Carlos López-Morate provided an excellent proof for the lower standard errors under included-irrelevant variables property Theorem 2(G).

influence on reported findings, as strategic or arbitrary inclusion of even a single regressor can radically vary the conclusions. All this raises the stakes for discovering an alternative to OLS that will be both rationally interpretable and robust in the face of multicollinearity.

Various statistical treatments for multicollinearity have been developed, each coming at some cost in terms of either coefficient bias or interpretability. Step-wise methods (Hocking, 1976) and discretionary elimination of correlated regressors introduce omitted-variable bias. The advantage of retaining all regressors via ridge regression (Hoerl and Kennard, 2000) comes at the cost of systematic downward bias.

Orthogonalization is a popular machine-learning solution for computation, forecasting, and signal processing wherever directly interpretable coefficients are not required (Despois and Doz, 2023). Laplace (1816) introduced his orthogonalization approach to compute Legendre's newly popularized (1805) least-squares regression, free of Gauss's (1809) computationally more expensive normal equations (see Stigler, 1981). Laplace's method was independently discovered by Gram (1883), Schmidt (1907), and Iwasawa (1949), and coined the Modified Gram-Schmidt Process by Wong (1935) (see Farebrother, 1988).[1] The Gram-Schmidt is a special case of upper-triangular matrix decomposition, referred to by Francis (1961) as QR decomposition, which preserves all information in the original data set, a so-called lossless process. A long list of orthogonal models have followed Laplace, notably Pearson's (1901) lossless Principal Component Analysis (PCA) for covariate-importance ranking and Golub and Reinsch's (1970) Singular Value Decomposition (SVD) eigenvalue-based method for data compression.

Research interest in multicollinearity has waned in recent decades due in part to a growing consensus that growing data sizes will ameliorate multicollinearity's impact (Greene, 2018) and partly from the unreliability of existing diagnostic approaches like pairwise correlation and the Variance Inflation Factor (VIF) (Kalnins and Praitias Hill, 2023). In addition, multicollinearity is commonly perceived to be a statistical rather than economic problem (Enikolopov *et al.*, 2023). Unfortunately, statistical confidence reported for the incorrect sign (Type I error), so-called sign-switching, increases with sample size (Kalnins, 2018). We will explore how multicollinearity's statistical symptoms may arise from omitting readily available economic information heretofore overlooked – the temporal order of the regressors.

This paper makes four contributions. First, we derive the finite sample properties of the Gram-Schmidt least-squares model. We show how the model preserves all regressor information and that its coefficients are unbiased and stable, with standard errors lower than in OLS. Second, we explore conditions under which the model returns average treatment effects when treatment responses are heterogeneous. We show that for late treatments - those applied after all other personal characteristics have been determined - the estimate is identical to OLS, and the average treatment effect bias can be estimated and corrected with existing approaches (Słoczyński, 2022). For early treatments, applied before other characteristics and associations are formed, the model returns an unbiased estimate of the average total treatment effect on the treated (ATTT). Third, we extend the model to groups of simultaneous regressors common in economic data sets, expanding the model's use beyond strictly temporally distinct regressors.[2] Finally, we apply our model to Słoczyński's (2022) decomposition of Angrist and Pischke's (2009) analysis of job program effectiveness, and Lubotsky and Wittenberg's (2006) assessment of contributors to child reading-and-comprehension outcomes. Both studies controlled for several collinear and temporally-ordered individual characteristics. Our analysis extends earlier findings to the Bohren *et al.* (2023) decomposition, which breaks total discrimination

[1] Langou (2009) provides an English translation of Laplace's original 1816 manuscript.

[2] Here, we refer to regressors that are simultaneous to one another rather than to the regressand or error term as are usually considered in simultaneous and endogenous models.

into its direct and systemic elements. Our model expands this decomposition, dividing systemic discrimination further into such channel-specific mechanisms as education, income, and household formation.

Our interpretability result is motivated by earlier research in interdependent dynamic systems, first explored in the 1920s by economists interested in describing and forecasting simultaneous supply and demand relations. There, system identification was a primary focus. Early strategies included the instantaneous equilibrium condition, lagged endogenous variables, causal chains, and instrumental variables. Schultz (1928) introduced the cobweb model, extended by Wright (1925, 1934) to path analysis and instrumentation. Wold generalized these causal chain and dynamic system approaches to the Linear Systems of Equations Model (LSEM) (1951, 1960), showing in a time-series framework that model coefficients were asymptotically consistent (1963). Goldberger (1972) explored systems with unobservable regressors and early latent factor models. Alwin and Hauser (1975) first calculated indirect effects and their confidence intervals by using a product-of-coefficients method.

Rubin (1974) refocused causal identification away from LSEM's lagged variables and instantaneous equilibria toward the random control trial (RCT) strategies inspired by Neyman (1923), the so-called Rubin causal model (see Holland, 1986), or Potential Outcomes approach (see Imbens, 2020). Imbens and Angrist (1994) and Card and Krueger (1994) extended these strategies greatly, particularly the use of instrumental variables, difference-in-difference, and shape restrictions, frequently considering models with potential endogeneity from omitted variables. The resulting Causal Inference (CI) framework is used widely across academic disciplines.

Parallel to these efforts, Pearl (2000) expanded LSEM to the probabilistic, graph-centric Causal Mediation (CM) framework, of growing popularity in the biological, computer, and social sciences outside economics. Pearl showed a regressor's total effect to be a partial derivative of an independent variable with respect to the regressor's path through the CM system (Hünermund and Bareinboim, 2023), reformulating Wright's (1934) method of path coefficients and extending Wold's chain principle (1963) beyond a time series setting. Pearl recast Wold's conditions in terms of a probabilistic Directed Acyclic Graph (DAG) or Bayesian Network. One-way causality is achieved either when regressors are temporally separated from one another by the natural occurrence of the data or by experimental design, referred to as Sequential Ignorability. Pearl (2001, p. 418) shows identification in this system is equivalent to the Gauss-Markov assumptions under the additional conditions of independent residual vectors and freedom from heterogeneity induced by unobservable regressors.[3]

Simultaneous regressors remain unidentified in both the LSEM and CM frameworks.[4] We utilize this graphical representation to motivate the linkage between temporal order and direct and total effects and provide properties sufficient for equivalence between the two frameworks. That DAGs follow from the LSEM framework is well documented, and we will not provide further equivalence conditions here.

Temporal order is by no means the only path to separability among regressors. It may be achieved either partially or fully by experimental design (Imbens, 2018) or awareness of the underlying economic or other known causal mechanisms (Borusyak and Hull, 2023). Temporal order is readily observable from data-collection frequency and timing. The passage of time precludes feedback (simultaneity) among regressors in a way that other data-collection or generating processes sometimes lack. In certain cases regressors may be simultaneously determined when measured over a given time scale, in monthly average expenditures and prices

[3]We show later how independent residuals and non-heterogeneity from unobservable regressors follow directly from the Gauss-Markov zero-conditional-mean condition and linearity assumptions, respectively.

[4]Wold (1963, Proposition 9) shows that the reduced-form parameters of an interdependent (simultaneous) system are not identified.

for instance, but well-ordered over smaller time scales such as daily transactions with posted prices.[5] Finally, temporal order removes multicollinearity but does not preclude endogeneity resulting from unobserved regressors, a phenomenon explored by Imbens and Angrist (1994). The relationship between temporal and unobserved regressors is being explored in a companion paper.

Linearity will be an important motivating assumption throughout this paper. It is restrictive and no longer explicitly imposed by many relevant CM and CI models (Hünermund and Bareinboim, 2023, Pearl, 2001). Despite this, linearity provides a tractable approach to temporal order and an accessible alternative to OLS, which remains a persistent tool for applied research (Imbens, 2015) with growing interest from machine-learning CM approaches (Kumor *et al*., 2020).

This paper then proceeds as follows. The next section introduces the Gram-Schmidt process for a simplified data set of two recursive regressors. Section 3 reviews the LSEM and recursive DAG frameworks and provides conditions under which Gram-Schmidt coefficients are equivalent in expected value to a recursive LSEM representation. Section 4 derives finite-sample estimation properties for our proposed model. Section 5 relaxes homogeneity and provides the conditions for decomposing the Gram-Schmidt coefficients into average total treatment effects for both early and late-occurring treatments. Section 6 extends the method to a mixed data set containing simultaneous as well as recursive regressors and derives the extended estimation properties. Section 7 illustrates the new model by replicating and expanding the work of Angrist and Pischke (2009) and Lubotsky and Wittenberg (2006), who consider the effects of collinear and temporally-ordered family and individual characteristics on earnings and childhood reading and comprehension. Section 9 concludes.

2 | THE GRAM-SCHMIDT PROCESS

The Gram-Schmidt process begins with a data set X consisting of K real-valued variables arranged in random order. The second variable is regressed on the first and replaced with the resulting residual; then, the regression coefficient is saved. This process is repeated, each variable regressed successively on all prior and then replaced by the resulting residual, saving the coefficients. The outcome is an upper-triangular matrix of estimation coefficients C and a new variable matrix U constituting the orthogonal basis of the original data set, essentially preserving all variable information but in orthogonal form. The final step is to divide each residual by its standard deviation, creating the regressor set's orthonormal basis. The latter step will be excluded below, although it is sometimes useful, and such coefficients can be interpreted in terms of standard deviation units.

The process can be represented by a set of line-by-line OLS regression equations, with each residual replaced by the prior equation's residual. The simple recursive system represented in Figure 1(a) can be represented as follows:

$$x = u_x, \tag{1}$$

$$d = c_{xd}u_x + u_d, \tag{2}$$

$$y = c_{xy}u_x + c_{dy}u_d + u_y, \tag{3}$$

[5] Lewis-Beck and Mohr (1976, p. 37) provide a helpful discussion of the source of simultaneity posed by Strotz and Wold (1960) and developed by Fisher (1970) and Johnston (1972). They suggest that, by nature, regressors tend to arise sporadically rather than all-at-once. Even when feedback occurs between two particular regressors, it is usually by way of a succession of stimuli and responses. In brief, regressor simultaneity arises in naturally occurring data whenever its collection intervals span regressor creation and feedback response activities.

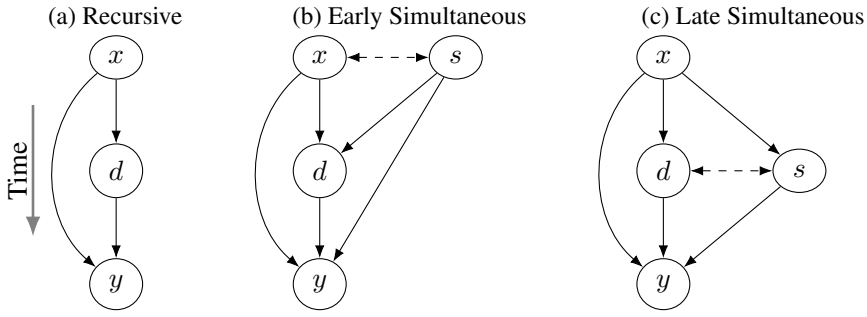

FIGURE 1. Three directed-acyclic graphs representing three causal systems: (a) fully identified and recursive; (b) unidentified on account of an earlier-determined simultaneous regressor; and (c) system (b) but with a later-determined simultaneous regressor instead. Arrows indicate the directions of causality of the recursive regressors x, d, and simultaneous regressor s. The passage of time is shown on the far left.

where c_{xd}, c_{xy}, and c_{dy} are estimated coefficients, x, d, and y are data vectors, and u_x, u_y, and u_d the residuals. The coefficients can be specified as the inner-product ratios $c_{xd} = x'd/x'x$, where $'$ is the transpose operator.[6]

In matrix form, the system including y can be written as $X = U(C + I)$, where the $(K + 1) \times N$ matrix $X = [x\, d\, y]$ is decomposed into three components: (i) the $(K + 1) \times N$ matrix U, containing orthogonal residuals vectors $[u_x\ u_d\ u_y]$; (ii) the identity matrix I; and (iii) the upper-triangular $K \times K$ matrix

$$C = \begin{bmatrix} 0 & c_{xd} & c_{xy} \\ 0 & 0 & c_{dy} \\ 0 & 0 & 0 \end{bmatrix}. \tag{4}$$

This procedure produces a convenient orthogonal data set U. But how can coefficient matrix C be interpreted? To answer this, we next explore the LSEM and recursive DAG models.

3 | ECONOMIC INTERPRETATION AND EQUIVALENCE

DAGs are one approach to formulating a structural system of equations, deriving the reduced form, checking identification, and recovering the parameters. Three types of parameters are recovered in such systems: (i) direct effects, where a regressor acts, as is the case with OLS, directly on the dependent variable, holding other regressors constant; (ii) indirect effects, when a first regressor acts indirectly on the dependent variable by influencing a second regressor; and (iii) total effects, the sum of direct and indirect effects.

To illustrate, consider the linear structural equation representation for the recursive example in Figure 1(a), where x and d are centered and ordered regressors; y is the independent variable; β_{ij}, $i, j = x, d, y$, are the unobservable population parameters; and υ_i is the residual:

$$x = \upsilon_x, \tag{5}$$

$$d = \beta_{xd}x + \upsilon_d, \tag{6}$$

[6]For comparability with OLS, the algorithm is presented here in the inner-product form suggested by Longley (1981), under which Laplace's computational accuracy remains identical to that of the modified Gram-Schmidt. See Farebrother (1988).

$$y = \beta_{xy}x + \beta_{dy}d + \upsilon_y. \tag{7}$$

Here, variables x and d are recursive because they occur one at a time in causal order by way of a temporal or experimentally designed separation. Specifically, x is determined before d and then affects d directly through parameter β_{xd}. Regressor x is also determined before y, acting upon y directly through β_{xy}. Finally, x exerts an indirect effect on y by way of its influence on d, in turn influencing y through β_{dy}. Together, the total effect of x on y is the sum of its direct and indirect effects, i.e., $(\beta_{xy} + \beta_{xd}\beta_{dy})$. By the chain rule, this total effect is also the partial derivative of y with respect to x.

We can now assemble the total effects of the reduced form system, where residuals υ_i, $i = x, d, y$ now serve as regressors:

$$x = \upsilon_x, \tag{8}$$

$$d = \beta_{xd}\upsilon_x + \upsilon_d, \tag{9}$$

$$y = (\beta_{xy} + \beta_{xd}\beta_{dy})\upsilon_x + \beta_{dy}\upsilon_d + \upsilon_y. \tag{10}$$

Note that parameter β_{dy} is both the direct and total effect of d on y, given that d is the last regressor to be determined and influences no other regressors. This can also be shown by the Frisch-Waugh-Lovell decomposition theorem (Frisch and Waugh, 1933, Lovell, 1963), first described by Yule (1907), that β_{dy} is recoverable from the regression of the independent variable y on residual υ_d and regressor d.

The source of multicollinearity in this recursive system is that each regressor influences those determined later. For instance, when regressors are standardized and no spurious or incidental multicollinearity is present, the covariance between x and d is exactly the direct-effect parameter β_{xd}.

For convenience, the entire reduced-form system, including the dependent variable, can also be represented as a matrix decomposition $X = U(A + I)$, with upper-triangular parameter matrix:

$$A = \begin{bmatrix} 0 & \beta_{xd} & \beta_{xy} + \beta_{xd}\beta_{dy} \\ 0 & 0 & \beta_{dy} \\ 0 & 0 & 0 \end{bmatrix} \tag{11}$$

Matrix A also represents the matrix of first-order partial derivatives because $\frac{\partial y}{\partial x} = \beta_{xy} + \beta_{xd}\beta_{dy}$. The parameter matrix A appears similar to the Gram-Schmidt coefficient matrix C, which we now formalize.

Under the following assumptions, an equivalence between Gram-Schmidt coefficients $c_{ij} \in C$ in (4), $i, j = 1, ..., K$, and the reduced-form LSEM parameters $a_{ij} \in A$ in (11) can be shown:

(a) $X \subseteq \mathcal{R}^{[N \times K]}$ is a full-rank, ordered regressor matrix including dependent variable $y = x_K$;
(b) X has positive degrees of freedom $N > K$; and
(c) Error vectors have zero conditional means: $E[\upsilon_i | x_j] = 0$, $j = 1, ..., i - 1$.

We will denote the sample residual vectors as $u_i \subseteq U$, $i = 1, ..., K$.

Theorem 1 – Equivalence. Under assumptions (a) - (c), there exists a unique Gram-Schmidt coefficient matrix $C \subseteq \mathcal{R}^{[K \times K]}$ equivalent in expectation to the recursive LSEM reduced-form matrix $A \subseteq \mathcal{R}^{[K \times K]}$

$$E[C] = A. \tag{12}$$

Proof. See online Appendix.

4 | ESTIMATION PROPERTIES

In the next theorem, we will compare the finite sample properties of Gram-Schmidt coefficients c_{ij} in equations (1) - (3) with those estimated by OLS b_{ij}. Define: (*i*) X_{-j} as the full regressor set X excluding regressor x_j; (*ii*) X_{+k} as the regressor set including the additional variable x_k; (*iii*) $X_{<j}$ as the regressor set including ordered regressors x_i up to, but not including, x_j, $i = 1, ..., j-1$; and (iv) $R^2_{jX_{<j}}$ as the coefficient of determination of the regression of x_j on the set of remaining regressors $X_{<j}$. We will assume the regressor order is $i < k < j$ throughout, in which x_j will be the j^{th} equation's dependent variable. We make two additional assumptions:[7]

(d) The true underlying model is given by structural equations (5) - (7); and

(e) OLS and Gram-Schmidt residuals υ_i are free of autocorrelation and heteroscedasticity, $E[\upsilon_i \upsilon_i' | \mathbf{X}] = \sigma_i^2 \mathbf{I}$.

Theorem 2 – Properties. Under assumptions (a) - (e), the Gram-Schmidt system $X = U(C + I)$ obtains the following estimation properties:

(A) Regressors are orthogonal, $u_i \perp u_j$;

(B) Coefficients are stable, $Cov[c_{ij}, c_{ji}] = 0$;

(C) Coefficients are unbiased, $E[c_{ij}] = a_{ij}$;

(D) All information is preserved, $R^2_{j\,X_{<j}} = R^2_{j\,U_{<j}}$;

(E) Omitted-variable bias is zero, $E[c_{ij}|X] = E[c_{ij}|X_{-k}]$;

(F) Gram-Schmidt variance is lower than OLS, $V[c_{ij}] \leq V[b_{ij}]$; and

(G) Gram-Schmidt variance is lower than OLS under included-irrelevant variables, $V[c_{ij}|X_{+k}] \leq V[b_{ij}|X_{+k}]$.

Proof. See online Appendix. Confidence intervals of the total effects and model inference are obtained directly from the terminal Gram-Schmidt regression, eliminating the need to reconstruct indirect effects by way of either product-of-coefficients (Alwin and Hauser, 1975) or simulation methods.[8]

5 | CAUSALITY

To explore the causal interpretation of the Gram-Schmidt coefficients, we relax the homogeneity assumption implicit in the structural model (5) - (7), allowing individuals to exhibit heterogeneous responses to the treatment. We will examine two cases: the classic (late) treatment case in which a treatment d is administered to a group of individuals with predetermined characteristics x, and an alternative (early) treatment case in which the treatment is one such characteristic x or a treatment administered before the formation of such characteristics. The early and late cases are respectively represented by treatment variables x and d in Figure 1(a).

In the late-treatment case d, we will show that both the convex combination and causal interpretations of the Gram-Schmidt are identical to those of OLS, introduced by Angrist (1998) and

[7] These complete the classic assumptions of the Gauss-Markov theorem, with homogeneity imposed by the scalar-value parameters in the structural equations (5) - (7).

[8] Imai *et al*. (2010a) propose what they refer to as a nonparametric indirect-effect estimator, calculated by forecasting the independent variable with and without the regressor interaction of interest. Their model relies on parametric estimators of the structural equations. They show their estimator is consistent when all underlying estimators are also consistent. When regressors are present, model errors lack an analytic asymptotic distribution, so confidence intervals must be simulated (Imai *et al*., 2010b, p. 59).

extended by Słoczyński (2022). In the early-treatment case x by contrast, the Gram-Schmidt approach provides the unbiased average total treatment effect (ATTE). The average total treatment effect on the treated (ATTT) is identical to that in the untreated (ATTU). This makes intuitive sense because early-treatment total effect c_x includes x's direct effect β_{xy} as well as its intermediate influence on the formation of later personal characteristics $\beta_{xd}\beta_{dy}$ - which themselves may include associations with various programs and institutions. In effect, if a control group member were to be moved to the early treatment group, such as being black, all regressor characteristics would follow as if they had been a member of the original treatment group.

Two existence and uniqueness assumptions are required for the convex combination interpretation:

(f) Expected values $E[y^2]$ and $E[\|z\|^2]$ are finite, $z = x, d$;
(g) The covariance matrix of (x, d) is nonsingular; and
(h) Variances $V[p(z)|m = j]$ are nonzero, where $m = x, d$, $m \neq z$, and $j = 0, 1$.

Here $p(z)$ is the probability of treatment, defined as the projection $\hat{z}$ from a regression of the treatment variable z on the predetermined characteristics m, as shown in equation (6) for variables d on x.

It is also useful to define the projection of y on the probability of treatment $p(z)$, given a treatment group $j = 0, 1$:

$$L[y \mid p(z), m = j] = \alpha_j + \gamma_j p(z). \tag{13}$$

The average partial linear effect (APLE) of the treatment on group j is then

$$c_{zy(APLE,j)} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)E[p(z) \mid m = j]. \tag{14}$$

A causal interpretation also requires ignorability of the mean and linear probability of a binary treatment variable:

(i) $E[y(j) \mid z, m] = E[y(j) \mid z]$; and
(j) $E[y(j) \mid z] = \alpha_j + \gamma_j p(z)$.

We can now state the interpretation of the late-treatment Gram-Schmidt (LTGS).

Lemma 1 – Late-treatment Gram-Schmidt is identical to OLS.
Under assumptions (a) - (h) and a heterogeneous treatment response:

(A) LTGS is a convex combination of partial linear effects,
$c_{dy} = \omega_1 c_{dy(APLE,1)} + \omega_0 c_{dy(APLE,0)}$; and
(B) LTGS is a convex combination of ATTT and ATTU, also assuming (i) - (j),
$c_{dy} = \omega_1 c_{dy(ATTT)} + \omega_0 c_{dy(ATTU)}$.

Here, ω_j are the convex, variance-weighted treatment proportions defined in Słoczyński (2022).

Proof. (A) The late-treatment Gram-Schmidt coefficient is identical to OLS by the Frisch-Waugh-Lovell decomposition theorem (Frisch and Waugh, 1933, Lovell, 1963), allowing us to apply the result from Słoczyński's Theorem 1 and Corollary 1.

Lemma 2 – Early-treatment Gram-Schmidt (ETGS) is the ATTE.
Under assumptions (a) - (h) and a heterogeneous treatment response:

(A) ETGS is the ATTE,
$c_{xy} = E[y \mid x = 1] - E[y \mid x = 0]$; and
(B) ETGS ATTT and ATTU are identical.

Proof. See online Appendix.

Causal identification is also challenged by endogeneity from unobserved regressors, as explored by Imbens and Angrist (1994) and many others. Gram-Schmidt properties under omitted variables, included-irrelevant variables, and late-treatment variables all have interesting implications for unobserved regressors.

6 | EXTENDED GRAM-SCHMIDT LEAST SQUARES

The Gram-Schmidt transformation is equivalent to the recursive DAG and LSEM when every regressor is recursive, that is, strictly separated by time or when the experimental design is such that no feedback can occur between regressors. However, the latter is an unlikely condition in practice because regressor simultaneity is present in many naturally occurring – and even some experimentally designed – data sets, especially when data collection intervals are wide. In the present section, we extend the Gram-Schmidt process to allow a block of simultaneous regressors to replace a single recursive regressor. This preserves temporal ordering but avoids endogeneity bias because the simultaneous regressors within a block are not regressed on one another. Instead, we construct a block-upper-triangular system of coefficients, the final column representing the total effects on the dependent variable from the relevant regressor. Coefficients thus remain consistent with the dependent variable's partial derivatives with respect to the relevant regressor's path-specific effect described by Wright (1934) and Pearl (2000), excluding feedback effects. By omitting only the mutual regressions of the simultaneous regressors, we allow the recovered total effects to exclude only direct feedback effects among simultaneous regressors. Indirect effects of such feedback that may progress through the system are preserved in the total effects of earlier-occurring regressors.

6.1 | Simultaneous Regressors

It is worthwhile first to explore the identification problem raised by these simultaneous regressors. It is well known that simultaneous equations, such as supply and demand systems, are not directly identifiable, so parameters cannot be immediately recovered. The same is true for simultaneous regressors in structural and DAG models. Consider then a simplified system with two simultaneous regressors s and x:

$$s = \beta_{xs}x + \upsilon_s, \tag{15}$$

$$x = \beta_{sx}s + \upsilon_x, \tag{16}$$

$$y = \beta_{sy}s + \beta_{xy}x + \upsilon_y. \tag{17}$$

This system is unidentified because neither the simultaneous coefficients nor the residual vectors are observable.

Partial derivatives can, however, be obtained by way of the Implicit Function Theorem:

$$\frac{\partial y}{\partial x} = \beta_{xy} + \beta_{sy}\frac{1+\beta_{xs}}{1+\beta_{sx}}.$$

Unfortunately, the partial derivatives here are also functions of the unrecoverable parameters β_{sx} and β_{xs}. We will refer to these unrecoverable parameters as feedback effects. Direct effects β_{iy} *are* however fully recoverable because OLS remains unbiased, $E[b_{iy}] = \beta_{iy}$. We will also find a way to recover some feedback information from earlier-occurring regressors.

Multicollinearity is assured by the simultaneity of x and s, illustrated by the off-diagonal elements of the variance-covariance matrix Σ_s of the standardized regressors, shown here without

the influence of additional spurious or incidental multicollinearity:

$$\Sigma_s = \begin{bmatrix} \frac{1+\beta_{xs}^2}{(1-\beta_{xs}\beta_{sx})^2} & \frac{\beta_{xs}+\beta_{sx}}{(1-\beta_{xs}\beta_{sx})^2} \\ \frac{\beta_{xs}+\beta_{sx}}{(1-\beta_{xs}\beta_{sx})^2} & \frac{1+\beta_{sx}^2}{(1-\beta_{xs}\beta_{sx})^2} \end{bmatrix} \tag{18}$$

In the next section, we explore the information that can be recovered when a system contains both simultaneous and recursive regressors.

6.2 | Mixed Simultaneous and Recursive Regressors

Consider next a system containing both an earlier- and later-determined simultaneous regressor block as illustrated in Figures 1(b) and 1(c), respectively. We will motivate the problem for the earlier-determined simultaneous block case, Fig 1(b), although the properties of the extended method shown here apply to both cases.

The structural equations assumed in Fig 1(b) are:

$$s = \beta_{xs}x + \upsilon_s, \tag{19}$$

$$x = \beta_{sx}s + \upsilon_x, \tag{20}$$

$$d = \beta_{sd}s + \beta_{xd}x + \upsilon_d, \tag{21}$$

$$y = \beta_{sy}s + \beta_{xy}x + \beta_{dy}d + \upsilon_y. \tag{22}$$

As in our simultaneity example in subsection 5.1 above, s and x here are simultaneous. But d has replaced y as the third regressor in the system, so dependent variable y is now a function of three regressors. We know direct feedback effects are unrecoverable, while direct effects β_{id} and β_{iy} are recoverable, though OLS will suffer from steep multicollinearity. This can be seen in the covariance between s and d, illustrated in the following equation for standardized regressors with no additional spurious multicollinearity:

$$Cov[s,d] = \frac{\beta_{sd}(1+\beta_{xs}^2) + \beta_{xd}(\beta_{xs}+\beta_{sx})}{(1-\beta_{xs}\beta_{sx})^2}.$$

As explored next, it will be possible to recover more information than direct effects alone and continue to remove the recursive portion of the multicollinearity.

6.3 | The Gram-Schmidt Extension

Our first step is to derive the reduced-form equations in terms of the recursive residual υ_d – along with simultaneous regressors x and s – rather than in terms of the residuals as was the case for x and d in the original Gram-Schmidt model (8)-(10):

$$\begin{aligned} s &= s, \\ x &= x, \\ d &= \beta_{sd}s + \beta_{xd}x + \upsilon_d, \\ y &= (\beta_{sy} + \beta_{sd}\beta_{dy})s + (\beta_{xy} + \beta_{xd}\beta_{dy})x + \beta_{dy}\upsilon_d + \upsilon_y. \end{aligned}$$

This system can be specified as $X = U(A + I)$, where $U = [s\, x\, \upsilon_d \upsilon_y]$ and parameter matrix A – extending (11) – is now the block-upper-triangular partial derivatives matrix:

$$A = \begin{bmatrix} 0 & 0 & \beta_{sd} & \beta_{sy} + \beta_{sd}\beta_{dy} \\ 0 & 0 & \beta_{xd} & \beta_{xy} + \beta_{xd}\beta_{dy} \\ 0 & 0 & 0 & \beta_{dy} \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{23}$$

The extended method is then estimated as:

$$\left.\begin{aligned} s &= s \\ x &= x \end{aligned}\right\} \textit{ Block 1,}$$

$$\left.\begin{aligned} d &= c_{sd}s + c_{xd}x + u_d \end{aligned}\right\} \textit{ Block 2,}$$

$$\left.\begin{aligned} y &= c_{sy}s + c_{xy}x + c_{dy}u_d + u_y \end{aligned}\right\} \textit{ Block 3.}$$

Block 1 consists of the early simultaneous regressors x and s, as in Figure 1(b), while blocks 2 and 3 contain only one dependent variable, similar to the recursive case in Figure 1(a). Block 1 feedback coefficients b_{sx} and b_{xs} are omitted, so x and s are not regressed on one another. However, each dependent variable in blocks 2 and 3 is regressed on all regressors in the blocks previous to it so the indirect effects b_{sd} and b_{xd} are recovered.

Residual regressor matrix U is no longer completely orthogonal. Rather, our extended method removes all covariance between any two blocks, in the present case between the simultaneous and recursive blocks, such that (i) the off-diagonal elements of the mixed variance-covariance partitioned matrix Σ_{sr} are zero, for instance $Cov[s, u_d] = 0$; but (ii) the simultaneous covariance matrix Σ_s from equation (18) remains in the upper left:

$$\Sigma_{sr} = \left[\begin{array}{c|c} \Sigma_s & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right],$$

where $\mathbf{0}$ is a conforming (2×1) zeros vector.

The extended method works similarly when a simultaneous regressor block follows a recursive regressor, as in Figure 1(c):

$$\left.\begin{aligned} x &= x \end{aligned}\right\} \textit{ Block 1,}$$

$$\left.\begin{aligned} s &= c_{xs}x + u_s \\ d &= c_{xd}x + u_d \end{aligned}\right\} \textit{ Block 2,}$$

$$\left.\begin{aligned} y &= c_{xy}x + c_{sy}u_s + c_{dy}u_d + u_y \end{aligned}\right\} \textit{ Block 3.}$$

Block 2 now contains the two simultaneous regressors s and d in Figure 1(c) that will not be regressed on one another, although each will be regressed on x in block 1. Covariances between pairs of blocks are again removed in the mixed variance-covariance partitioned matrix Σ_{rs}. For instance $Cov[u_s, u_d] = 0$ and the simultaneous covariance matrix Σ_s shifts to the lower right:

$$\Sigma_{rs} = \left[\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & \Sigma_s \end{array}\right].$$

6.4 | Extended Estimation Properties

We can now state the properties of our extended Gram-Schmidt least-squares (GSLS) method, which achieves the Theorem 2 properties across blocks rather than across individual regressors. Multicollinearity is eliminated from the recursive regressors, and recursive total effects are recovered. Multicollinearity among simultaneous regressors in a given block will be reduced but not eliminated. Feedback effects recovered, but the direct and intermediate effects are included in the simultaneous regressors' total effects.

To show this, we revise the orthogonality assumption (a) in Theorems 1 and 2. Let $X \subseteq \mathcal{R}^{[N\times K]}$ be a full-rank matrix of *recursively* ordered regressor blocks $m, n, h = 1, ...M$, each block m containing one or more regressors with coefficient $c_{i(m)j(n)}$, and block m determined prior to block n, $m < n$. The true underlying model in assumption (d) is now represented by block structural equations (19)-(22). Throughout, we assume the regressor order to be $i < k < j$ and the block order to be $m < h < n$, with $x_{j(n)}$ the j^{th} equation's dependent variable.

With this revision in mind, we can summarize the properties of the GSLS estimator.

Theorem 3 – Properties. Under our revised assumptions (a) - (e) above, the GSLS system $X = U(C + I)$ obtains the following estimation properties:

(A) Regressors are orthogonal across blocks, $u_{i(m)} \perp u_{j(n)}$;
(B) Coefficients are stable across blocks, $Cov[c_{i(m)j(n)}, c_{k(h)j(n)}] = 0$;
(C) Coefficients are unbiased, $E[c_{i(m)j(n)}] = a_{i(m)j(n)}$;
(D) All information is preserved, $R^2_{j(n)\,X_{<j(n)}} = R^2_{j(n)\,U_{<j(n)}}$;
(E) Omitted-variable bias is zero, $E[c_{i(m)j(n)}|X] = E[c_{i(m)j(n)}|X_{-k(h)}]$;
(F) GSLS variance is lower than OLS, $V[c_{i(m)j(n)}] \leq V[b_{i(m)j(n)}]$; and
(G) GSLS variance is lower than OLS under included-irrelevant variables, $V[c_{i(m)j(n)}|X_{+k(h)}] \leq V[b_{i(m)j(n)}|X_{+k(h)}]$.

Proof. See online Appendix.

7 | EMPIRICAL APPLICATIONS

We now compare the GSLS estimator to OLS using two empirical examples. Readers may replicate the analysis here with data and software packages available for R and Stata (Cross, 2024).

7.1 | National Supported Work Program

First, we extend Słoczyński's replication of LaLond's (1986) and Angrist and Pischke's (2009) analyses (hereafter SLA) of the National Supported Work (NSW) training program's impact on future earnings. Our late-treatment effect matches the impact of the program reported by SLA. We also look at the early-total treatment effect of being black on both inclusion in the work program and future earnings. GSLS extends the discrimination-decomposition framework of Bohren *et al.* (2023) to a regression context by decomposing the total-effect estimate into a direct and a systemic discrimination component. We will find that black individuals were included in the program at higher rates and experienced lower earnings than non-black individuals, a phenomenon linked to both direct and systemic discrimination.

7.1.1 | *Data*

Prior studies controlled for several strongly interrelated individual characteristics drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). NSW

participants were randomly assigned a training program and a control group. Specific job assignments were however assigned locally with candidate response rates varying across demographic groups. This resulted in an over-representation of black and economically disadvantaged individuals in the treatment group (Słoczyński, 2022). To test the extent to which black individuals were treated differently, we aggregate Hispanic and white participants to create a binary black and non-black treatment variable. This will serve as an early treatment because individuals were black before other study characteristics were formed and inclusion in the job program was assigned.

TABLE 1
Summary statistics of NSW data.

	Job Program		Control	
	Mean	Std. dev.	Mean	Std. dev.
Black (proportion)	84%	0.4	7%	0.3
White (proportion)	10%	0.3	85%	0.4
Age	25.8	7.2	33.2	11.0
Education	10.4	2.0	12.0	2.9
No degree (proportion)	71%	0.5	30%	0.5
Married (proportion)	19%	0.4	71%	0.5
Annual earnings 1978 ($1000s)	6.3	7.9	14.8	9.6

Table 1 summarizes the NSW data for the 185 job program participants and the 15,992 control group individuals. Several regressor means and standard deviations differ materially between the treatment and control groups, suggesting job assignment completions were not random. In the Causal decomposition subsection, we test for the influence of these sample weightings.

7.1.2 | *Results*

TABLE 2
Direct and total effects of selected regressors on NSW participants' 1978 earnings ($1000s).

	OLS direct effects		GSLS total effects	
	Coeff.	S.E.	Coeff.	S.E.
Black	−2.23	(0.28)	−3.74	(0.25)
Age	0.13	(0.01)	0.14	(0.01)
Education	0.23	(0.04)	0.35	(0.03)
No degree	−1.09	(0.23)	−1.29	(0.23)
Married	3.21	(0.19)	3.28	(0.19)
Job program	−3.47	(0.71)	−3.47	(0.71)
R^2	0.12		0.12	

Table 2 compares, for selected regressors, the OLS direct effects and GSLS total effects and standard errors. In both interpretation and expected value the GSLS coefficients differ from

OLS. The latter remains unbiased, $E[b_{xy}] = \beta_{xy}$, in the presence of multicollinearity, and its estimates of structural equation (7) represent direct effects. Coefficients are interpreted in the familiar way, namely as an increase or decrease in y resulting from a unit increase in x, *ceteris paribus*, assuming all other regressors are fixed or independent. In linear models, this is the partial derivative of y with respect to x, ignoring the regressor interrelatedness that induces multicollinearity.

GSLS estimates provide the total effect, namely the total unit increase (decrease) in y induced by a one-unit increase in x, including any indirect effects on y induced or caused by changes in the remaining x-dependent regressors. The latter, by the chain rule, is the partial derivative of y with respect to x given all structural relationships in equations (5) - (7), what Wright (1934) refers to as the path coefficient or Pearl as path-switching (2001) or the path-specific effect (2000). GSLS standard errors are lower by the removal of multicollinearity.

The job program's estimated direct effect in the Table 2 OLS model is identical to its total effect in the GSLS model because, as predicted by Theorem 2, this program is the final regressor in the model. The program's coefficient suggests program participation reduced annual earnings by $3.47 thousand, slightly larger than the $3.44 thousand found by Słoczyński (2022, Table 1, Column 1), the difference attributable to our aggregation of white and Hispanic participants.

7.1.3 | *Race*

To understand the rise in GSLS Black coefficient magnitude, we look to the discrimination decomposition framework of Bohren *et al.* (2023). They define the total expected discrimination function $\Delta(y^0)$ as the sum of direct and systemic discrimination:

$$\Delta(y^0) = \bar{\tau}(w, y^0) + \delta(b, y^0),$$

where b and w are groups and y^0 is the unobservable initial qualification. The direct and systemic components here can be expressed in terms of expected values:

$$\bar{\tau}(w, y^0) = E[A(w;\ S_i) - A(b;\ S_i)|G_i = w;\ Y_i^0 = y^0], \tag{24}$$

$$\delta(b, y^0) = E[A(b;\ S_i)|G_i = w;\ Y_i^0 = y^0] - E[A(b;\ S_i)|G_i = b;\ Y_i^0 = y^0]. \tag{25}$$

Here, $A()$ is the action function and in our case the earnings function; S_i is the set of signals for individual i, in our case the individual's race and personal characteristics; G_i is the set of groups w, b; and Y_i^0 is the set of initial qualifications, which we assume to be equal across all participants in the program since, for the early-treatment, it represents their potential for earnings before they are born.

In words, direct discrimination in (24) is the difference in action A, earnings, received by individual i, were they to belong to group w rather than group b, holding the individual's other signals (background characteristics) S_i and initial qualification y^0 constant at the group-w level. This matches the interpretation of OLS results shown in Table 2. The direct discrimination estimate suggests earnings are now lower by a significant $2.23 thousand dollars per year.

Systemic discrimination in (25) is the expected difference between the earnings of an individual from group b but with background characteristics typical of group w, and the same individual's score given group b background characteristics. It represents the cumulative effects of differences in access to such resources as education and employment, which Bohren *et al.* categorizes as *technological* sources. Our total discrimination estimate of $3.74 thousand less per year is significantly greater than the direct effect magnitude, suggesting systemic discrimination is present.

7.1.4 | *Causal decomposition*

The top panel of Table 3 reports the regression coefficient and its decomposition into the heterogeneous treatment effect on the treated, untreated, and average participants. The probability of being treated $P(d = 1)$ and the variance-adjusted statistical weight ω_1 are also listed along with the treatment and control sample sizes and standard errors.

TABLE 3

Causal decomposition of race and job program effects on participants' 1978 earnings (1000s).

	OLS early-treatment Black		GSLS early-treatment Black		GSLS late-treatment Job program	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Regression coefficient	−2.23	(0.28)	−3.74	(0.26)	−3.47	(0.71)
Heterogeneous response						
ATE / ATTE	−2.69	(0.28)	−3.74	(0.26)	−6.73	(1.20)
ATT / ATTT	−0.66	(0.42)	−3.74	(0.26)	−3.40	(0.68)
ATU / ATTU	−2.87	(0.29)	−3.74	(0.26)	−6.77	(1.21)
$P(d = 1)$	0.08		0.08		0.01	
ω_1	0.92		0.92		0.98	
Obs. treated	1,332		1,332		185	
Obs. control	14,845		14,845		15,992	
Covariate matching						
ATE / ATTE	−2.45	(0.38)	−3.57	(0.38)	−8.05	(2.05)
ATT / ATTT	−2.11	(0.31)	−3.81	(0.30)	−3.45	(0.85)
Obs. treated	16,177		16,177		16,177	
Obs. control	16,177		16,177		16,177	

For comparison, the bottom panel of Table 3 provides rebalanced estimates from covariate matching along with rebalanced treatment and control sample sizes. Matching methods rebalance data by selecting a second, untreated individual with characteristics similar to those of each treated individual in the sample. This method selects the pair minimizing the Mahalanobis distance between regressor means and variances in the two groups (Dehejia and Wahba, 2002).[9]

In the GSLS late-treatment job program model in Column 3, the heterogeneous response estimates match closely to those in Słoczyński (2022, Table 1, Column 1), a result of the small magnitude differences resulting from aggregating white and Hispanic participants. The regression coefficient is not statistically different from the heterogeneous ATTT estimate, consistent with the potentially lower regression bias following from the lower probability of treatment

[9] As desired in the rebalanced model, matched standardized differences (differences in means) are close to zero; and the matched variance ratios are near unity, suggesting that rebalancing reduces the differences between the probability of moments of race, education, and enrollment in the job program despite significant differences between them in the original data. A complete list of rebalanced probability moments is available for both studies in the replication code repository.

$P(d-1) = 0.01$ and a statistical weight ω_1 close to unity. The smaller ATTE magnitude follows from the larger ATTU estimate along with the convexity restriction from Lemma 1. Finally, the magnitude of the matching ATTE is also larger than the matching ATTT, though not subject to the Lemma 1 restriction.

The total effect of being black is provided in the center column of Table 3. Illustrating Lemma 2, the ATTE, ATTT, and ATTU and their standard errors are identical to one another. This can be seen here from the larger total-effect coefficient magnitude, which includes the systemic impacts of being black on such later-formed personal characteristics as grade level, degree completion, and inclusion in the jobs program. Results from rebalancing are similar to the GSLS regression coefficient but with slightly higher standard errors.

The decomposition of the direct effect of being black - in the leftmost data column of Table 3 - follows a pattern similar to the decomposition in Column 3 but with a much smaller and statistically nonsignificant ATT. This is again partly due to the low probability of treatment $P(d=1) = 0.08$ and a greater ATU. The magnitude of the matching ATT is also smaller than the matching ATE but by less than the heterogeneity estimate since matching is not subject to the convexity restriction.

Finally, the center column of Table 3 provides the heterogeneous decomposition of the GSLS early-treatment effect of being black. As suggested by Lemma 2, the GSLS coefficient and heterogeneity ATTE estimate match, and the ATTE is equal to the ATTT. Covariate matching estimates are both within a standard deviation of the regression coefficient.

7.2 | National Longitudinal Survey of Youth

Our second illustration examines data from the Bureau of Labor Statistics' (2016) National Longitudinal Survey of Youth (NLSY) used by several researchers (Korenman *et al.*, 1995, Blau, 1999, Lubotsky and Wittenberg, 2006) to explore the role of parental income on child reading comprehension test scores while controlling for several strongly interrelated individual and household characteristics. All three studies cite multicollinearity as a leading motivation of their model design, regressor selection, and interpretation of results.

The NLSY began as a survey of 6,283 women and 6,403 men aged 14 to 21 in 1979 and includes a wide range of characteristics, including employment, income, drug use, marriage, education, cognitive assessments, and childbirth. The sample was rebalanced in 1984 to exclude women serving in the military and again in 1990 and 1991 to correct for the survey's original over-weighting of low-income white women. We will test for any remaining sample selection bias at the end of this section. Sampling frequency in the survey was annual through 1994 and biennial thereafter. A second survey, the Child Supplement, was initiated in 1986 to study the children of women included in the 1979 NLSY cohort, biennially recording cognitive achievement and behavior.

Childhood cognitive development is drawn from the Peabody Reading Comprehension Test. From 1986 to 2014, we include all complete observations of children aged 6-14 who were attempting the test for the first time. We exclude children of mothers dropped from the survey during the 1984, 1990, and 1991 revisions. The final sample includes, from the original NLSY cohort, 6,550 children born to 3,181 mothers.

Table 4 summarizes the NLSY data by child, broken out into above- and below-median family income levels with 3,275 observations in each category. Several regressor means and standard deviations differ materially between the high- and low-income groups, suggesting the 1984, 1990, and 1991 NLSY rebalancing efforts were insufficient or did not target income specifically. In the Causal decomposition subsection, we test for the impacts of any remaining influence of sample selection bias.

TABLE 4
SUMMARY STATISTICS OF NLSY DATA.

	Higher income		Lower income	
	Mean	Std. dev.	Mean	Std. dev.
Black (proportion)	17%	0.4	45%	0.5
White (proportion)	64%	0.5	31%	0.5
Mother's age	35.1	5.9	31.8	5.7
Mother's education	13.6	3.2	11.6	2.6
Mother's AFQT (percentile)	48.2	27.0	23.0	21.0
Spouse present (proportion)	91%	1.3	33%	0.5
Spouse's age	34.3	6.8	37.0	6.3
Spouse's education	12.2	7.9	13.7	3.7
Child female (proportion)	49%	0.5	49%	0.5
Child's age	7.6	1.4	8.0	1.6
Family size	4.5	1.2	4.2	1.7
Family income ($10K)	12.3	14.7	2.6	1.3
Child's test score (percentile)	63.9	24.1	51.7	26.4

Consistent with all three prior studies of these data, we include annual fixed effects and recursive regressors recorded on the date of test: child's gender and age; logs of family size and income, and mother's race, age, education, and performance on the 1980 Armed Forces Qualification Test (AFQT), a general knowledge exam administered to all participants in the original NLSY cohort. We also control for spousal presence in the household and spousal age and education to compare results with Lubotsky and Wittenberg (2006). Three simultaneous blocks are specified: (i) annual fixed effects; (ii) mother's race, consisting of a nonwhite indicator variable for black or Hispanic (white the omitted variable); and (iii) spousal characteristics, including spouse's residence in the home. All factors are specified in the natural temporal order in which they are assumed to have been determined. For instance, the child's race is defined in the Child Supplement to be the mother's race originally reported in the NLSY. Race was predetermined at the time of the mother's conception, so precedes the mother's age. In turn, both the mother's race and age were determined before her highest grade level was achieved or her AFQT score was recorded. Spousal characteristics, child age and gender, and family size and income are all recorded when the child attempts the reading and comprehension test for the first time, so they may influence the child's test score, but cannot be influenced by it.

7.2.1 | *Results*

Table 5 compares, for selected regressors, the OLS direct effects, GSLS total effects, and standard errors. The 2.31 family income coefficient implies a one-percent income rise boosts relative reading comprehension by 2.31 percentile points and is comparable to the range reported by Blau (1999) and Lubotsky and Wittenberg (2006).

7.2.2 | *Race*

Consistent with Blau (1999), the OLS estimate suggests no significant direct influence of maternal race on child test scores, indicated in the left column of Table 5.[10] This would be expected if

[10]Race estimates were not reported in Lubotsky and Wittenberg (2006).

TABLE 5
Direct and Total Effects of Selected Regressors on Child's Reading Scores.

	OLS direct effects		GSLS total effects	
	Coeff.	S.E.	Coeff.	S.E.
Nonwhite	−0.40	(0.67)	−10.83	(0.55)
Mother's age	0.06	(0.14)	0.02	(0.13)
Mother's education	0.41	(0.11)	1.80	(0.09)
Mother's AFQT	0.22	(0.01)	0.29	(0.01)
Child female	5.50	(0.55)	6.49	(0.55)
Child's age	−5.68	(0.20)	−5.75	(0.20)
Spouse present	1.92	(2.56)	1.31	(2.53)
Family size (log)	−8.22	(0.92)	−8.30	(0.92)
Family income (log)	2.31	(0.74)	2.31	(0.74)
R^2	0.28		0.28	

test administration and scoring mechanisms did not differ systematically between racial groups. The GSLS model shows race's total effect reduced reading scores by 10.83 percentile points among nonwhite children, relative to white children, significant at the 99.9% confidence level. The GSLS nonwhite coefficient standard error is lower than OLS by 19%, illustrating the impacts of removing multicollinearity.[11]

Systemic discrimination is represented by the negative and significant GSLS total effect for race, significantly greater in magnitude than the direct effect. We can use GSLS to further decompose the systemic effect into its channel-specific mechanisms. Table 6 provides the intermediate-stage GSLS regression results for selected regressors. Race's total effect on education can be seen in the second estimates column showing the regression of maternal school-grade completion on maternal race and age. Here, the total race effect on education was 1.11 fewer grades completed on average by nonwhite mothers than white mothers. In turn, each additional grade completed by the mother raised the child's test score by 1.80 percentile points, as indicated in the last column of Table 6. Thus, we have a 2.0 percentile point (1.11×1.80) differential between nonwhite and white children attributable to maternal education. This indirect education channel constituted approximately 18% of the 10.8-percentile total race effect.

7.2.3 | *Causal decomposition*

Table 7 tests for heterogeneity and sample weighting bias. The GSLS late-treatment income effect in Column 3 shows an ATTE estimate of 1.91, lower than the regression coefficient of 2.31. This downward bias results from the higher probability of treatment $P(d = 1) = 0.50$ and a similar variance-adjusted statistical weight ω_1. The high ATTT is offset by a nonsignificant ATTU close to zero.

The OLS early-treatment nonwhite effect estimates are nonsignificant for every heterogeneous response and covariate matching estimate, suggesting there is no evidence of direct discrimination in test administrations.

[11]For a given GSLS regressor, the standard error reduction falls below that indicated by VIF, as the former identifies and removes each regressor's contribution to total model multicollinearity, while the VIF attributes all model multicollinearity to each regressor successively.

TABLE 6
Intermediate GSLS step-wise regressions for selected regressors.

Dependent variable	Mother's age	Mother's education	Mother's AFQT	Spouse present	Family size	Family income	Child's score
Nonwhite	−1.82	−1.11	−29.78	−0.32	−0.01	−0.33	−10.83
	(0.05)	(0.07)	(0.49)	(0.01)	(0.01)	(0.01)	(0.55)
Mother's age		0.10	2.05	0.01	0.003	0.02	0.02
		(0.03)	(0.13)	(0.002)	(0.002)	(0.002)	(0.13)
Mother's education			3.44	0.01	−0.01	0.04	1.80
			(0.67)	(0.002)	(0.002)	(0.003)	(0.09)
Mother's AFQT				0.004	0.001	0.005	0.29
				(0.00)	(0.00)	(0.00)	(0.01)
Spouse present					0.14	0.40	1.31
					(0.04)	(0.05)	(2.53)
Family size (log)						−0.03	−8.30
						(0.01)	(0.92)
Income (log)							2.31
							(0.74)
R^2	0.88	0.11	0.48	0.15	0.17	0.46	0.28

The GSLS early-treatment coefficient in the center column again matches the heterogeneous responses exactly. The rebalanced sample matching estimates are - though slightly greater in magnitude - not statistically significantly different from the regression coefficient.

8 | CONCLUSION

Multicollinearity has posed a statistical challenge for over 200 years. We have shown that when coefficients are linearly separable through temporal order or experimental design, the Gram-Schmidt process removes multicollinearity and produces economically interpretable estimates, equivalent in expected value to total effects from a recursive Linear System of Equations Model or Directed Acyclic Graph. Total effects and statistical inference follow directly from the regression, eliminating the need for ex-post simulation or product-of-coefficients methods. The coefficients are unbiased estimates of the partial derivatives, invariant to the omission of later-determined regressors, with standard errors lower than in OLS. For early treatments, Gram-Schmidt returns the average total treatment effect on the treated even when the treatment response is heterogeneous.

We have exploited these properties to extend the Gram-Schmidt approach to recover causal effects from a previously unidentified mixed system of recursive and simultaneous regressors. The extended approach removes multicollinearity between recursive and simultaneous regressors, allowing unbiased estimates of the total effects that are free from multicollinearity.

Using a mix of temporally ordered and simultaneous individual characteristics, we have illustrated extended Gram-Schmidt least squares by applying it to previous studies using the National Supported Work Program and National Longitudinal Survey of Youth. GSLS total effects tended to be greater in magnitude than OLS direct effects, especially among early-determined or inter-related regressors. The approach expanded our interpretation of discrimination impacts reported in earlier studies and lowered standard errors by removing multicollinearity.

TABLE 7

Causal decomposition of race and income effects on children's reading scores (test percentile points).

	OLS early-treatment Nonwhite		GSLS early-treatment Nonwhite		GSLS late-treatment Higher income	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Regression coefficient	−0.40	(0.67)	−10.83	(0.55)	2.31	(0.74)
Heterogeneous response						
ATE / ATTE	−0.37	(0.67)	−10.83	(0.55)	1.91	(0.75)
ATT / ATTT	−1.02	(0.82)	−10.83	(0.55)	4.06	(0.86)
ATU / ATTU	0.34	(0.81)	−10.83	(0.55)	−0.23	(1.03)
$P(d=1)$	0.52		0.52		0.50	
ω_1	0.55		0.55		0.59	
Obs. treated	3,427		3,427		3,275	
Obs. control	3,123		3,123		3,275	
Covariate matching						
ATE / ATTE	−1.32	(0.81)	−11.74	(0.74)	1.67	(1.03)
ATT / ATTT	−0.12	(0.98)	−11.49	(0.86)	1.49	(1.24)
Obs. treated	6,550		6,550		6,550	
Obs. control	6,550		6,550		6,550	

REFERENCES

ALWIN, D.F., AND R.M. HAUSER (1975): "The Decomposition of Effects in Path Analysis," *American Sociological Review* 40 (1), 37-47. DOI:10.2307/2094445. [3, 7]

ANGRIST, J.D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66 (2), 249-288. DOI:10.2307/2998558. [7]

ANGRIST, J.D., AND J.S. PISCHKE (2009): "*Mostly Harmless Econometrics: An Empiricist's Companion*," (Princeton: Princeton University Press). DOI:10.1515/9781400829828. [2, 4, 12]

BAIARDI, A. AND A.A. NAGHI (2024): "The value added of machine learning to causal inference: evidence from revisited studies," *Econometrics Journal* 27, 213-34. DOI:10.1093/ectj/utae004. [1]

BLAU, D.M. (1999): "The Effect of Income on Child Development," *The Review of Economics and Statistics* 81(2), 261-76. DOI:10.1162/003465399558067. [16, 17]

BOHREN, J.A., P. HULL, AND A. IMAS (2023): "Systemic Discrimination: Theory and Measurement," NBER Working Paper 29820. http://www.nber.org/papers/w29820. Accessed 1.11.2024. [2, 12, 14]

BORUSYAK, K. AND P. HULL (2023): 'Nonrandom Exposure to Exogenous Shocks"," *Econometrica* 91 (6), 2155–2185. DOI:10.3982/ECTA19367. [3]

BUREAU OF LABOR STATISTICS, U.S. DEPARTMENT OF LABOR (2016): "*National Longitudinal Survey of Youth 1979 cohort, 1979-2016 (rounds 1-27)*," Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH. https://www.nlsinfo.org/content/access-data-investigator/investigator-user-guide. Accessed 04.27.2022. [16]

CARD, D., AND A. KRUEGER (1994): "Minimum Wages and Employment: A Case-Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review* 84 (4), 772–93. DOI:10.1257/aer.90.5.1397. [3]

CROSS, R.M. (2024): "GSLS Software in R and Stata and NLSY Replication Data," https://github.com/crossrm/GSLS. Accessed 01.11.2024. [12]

DEHEJIA, R.H., AND S. WAHBA (2002): "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics* 84 (1), 151-61. DOI:10.1162/003465302317331982. [15]

DESPOIS, T, AND C. DOZ (2023): "Identifying and interpreting the factors in factor models via sparsity: Different approaches," *Journal of Applied Econometrics* 38 (4), 533–55. DOI:10.1002/jae.2967. [2]

ENIKOLOPOV, E., A. MAKARIN, AND M. PETROVA (2023): 'Online Corrigendum to "Social Media and Protest Participation: Evidence From Russia"," *Econometrica* 91 (3), 1–24. DOI:10.3982/ECTA20824. [2]

FAREBROTHER, R.W. (1988): "*Linear Least Squares Computations*," 1^{st} ed. (Boca Raton: Routledge). DOI:10.1201/9780203748923. [2, 5]

FISHER, F.M. (1970): "A Correspondence Principle for Simultaneous Equation Models," *Econometrica* 38(1), 73-92. DOI:10.2307/1909242. [4]

FRANCIS, J.G.F. (1961): "The QR Transformation A Unitary Analogue to the LR Transformation–Part 1," *Computer Journal* 4 (3), 265-71. DOI:10.1093/comjnl/4.3.265. [2]

FRISCH, R., AND F. WAUGH (1933): "Partial Time Regressions as Compared with Individual Trends," *Econometrica* 1 (4), 387–401. DOI:10.2307/1907330. [6, 8]

GAUSS, C.F. (1809): "*Theoria Motus Corporum Coelestium*," in *Sectionibus Conicis Solem Ambientium*. F. Perthes and I. H. Besser, Hamburg. English translation by C. H. Davis (1857), Little, Brown, Boston. [2]

GELMAN, A., AND E. LOKEN (2014): "The Statistical Crisis in Science," *American Scientist* 102 (6), 460-465. DOI:10.1511/2014.111.460. [1]

GOLDBERGER, A.S. (1972): "Structural Equation Methods in the Social Sciences," *Econometrica* 40 (6), 979-1001. DOI:10.2307/1913851. [3]

GOLUB, G.H., AND C. REINSCH (1970): "Singular Value Decomposition and Least Squares Solutions," *Numerische Mathematik* 14 (5), 403-420. DOI:10.1007/BF02163027. [2]

GRAM, J.P. (1883): "Ueber die Entwickelung reeller Functionen in Reihen mittels der Methode der kleinsten Quadrate," *Journal für die reine und angewandte Mathematik* 94, 41–78. DOI:10.1515/crll.1883.94.41 [2]

GREENE, W.H. (2018): "Structural Equation Methods in the Social Sciences," *Econometric Analysis,* 8^{th} ed. (New York, NY: Pearson, 2018). [2]

HOCKING, R.R. (1976): "The Analysis and Selection of Variables in Linear Regression," *Biometrics* 32 (1), 1–49. https://www.jstor.org/stable/2529336. Accessed 04.27.2022. [2]

HOERL, A.E., AND R.W KENNARD (2000): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics* 42 (1), 80–86. DOI:10.1080/00401706.2000.10485983. [2]

HOLLAND, P.W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association* 81 (396), 945-60. DOI:10.2307/2289064. [3]

HÜNERMUND, P., AND E. BAREINBOIM (2023): "Causal Inference and Data Fusion in Econometrics," *The Econometrics Journal*, 1–44. DOI:10.1093/ectj/utae008. [3, 4]

IMAI, K.L., L. KEELE, AND D. TINGLEY (2010): "A General Approach to Causal Mediation Analysis," *Psychological Methods* 15 (4), 309-334. DOI:10.1037/a0020761. [7]

IMAI, K.L., L. KEELE, AND T. YAMAMOTO (2010b): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science* 25 (1), 51-71. DOI:10.1214/10-STS321. [7]

IMBENS, G.W. (2015): "Matching Methods in Practice: Three Examples," *The Journal of Human Resources,* 50 (2), 373-419. https://www.jstor.org/stable/24735990. Accessed 6.9.24. [4]

IMBENS, G.W. (2018): "Understanding and misunderstanding randomized controlled trials: A commentary on Deaton and Cartwright," *Social Science & Medicine* 210, 50–52. DOI:10.1016/j.socscimed.2018.04.028. [3]

IMBENS, G.W. (2020): "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics," *Journal of Economic Literature* 58 (4), 1129-1179. DOI:10.1257/jel.20191597. [3]

IMBENS, G.W. AND J.D. AGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62 (2), 467-475. DOI:10.2307/2951620. [3, 4, 9]

IWASAWA, K. (1949): "On Some Types of Topological Groups," *Annals of Mathematics*, Second Series, 50 (3), 507-558. DOI:10.2307/1969548. [2]

JOHNSTON, J. (1972): "*Econometric Methods*," 2^{nd} ed. (New York: McGraw-Hill) [4]

KALNINS A. (2018): "Multicollinearity: How common factors cause Type 1 errors in multivariate regression," *Strategic Management Journal* 39 (8), 2362–2383. DOI:10.1002/smj.2783. [2]

KALNINS A. AND K. PRAITIAS HILL (2023): "The VIF Score. What is it good for? Absolutely nothing," *Organizational Research Methods* 0 (0), 1–18. DOI:10.1177/1094428123121. [2]

KORENMAN, S., J.E. MILLER, AND J.E. SJAASTAD (1995): "Long-Term Poverty and Child Development in the United States: Results from the NLSY," *Children and Youth Services Review* 17 (1), 127-55. DOI:10.1016/0190-7409(95)00006-X. [16]

LALOND, R.J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review* 76, 604-620. https://www.jstor.org/stable/1806062. Accessed 11.10.24. [12]

KUMOR, D., C. CINELLI, AND E. BAREINBOIM (2020): "Efficient Identification in Linear Structural Causal Models with Auxiliary Cutsets," *Proceedings of the 37th International Conference on Machine Learning*. https://proceedings.mlr.press/v119/kumor20a.html. Accessed 1.11.2024. [4]

LAPLACE, P.S. (1816): "Premier Supplément," in *Théorie Analytique des Probabilités,* 3^{rd} ed. with an introduction and three supplements (1816, 1818, 1820), (Mme. Courcier, Paris, 1820). [2]

LANGOU, J. (2009): "*Translation and modern interpretation of Laplace's Théorie Analytique des Probabilités*, pages 505-512, 516-520," UC Denver CCM Technical Report no. 280. arXiv:0907.4695v1 [math.NA]. [2]

LEAMER, E.E. (1983): "Let's Take the Con Out of Econometrics," *The American Economic Review* 73 (1), 31–43. https://www.jstor.org/stable/1803924. [1]

LEGENDRE, A.M. (1805): "*Nouvelles Méthodes pour la Détermination des Orbites des Comètes*," Firmin Didot, Paris; second edition Courcier, Paris, 1806. Pages 72-75 of the appendix are printed in Stigler (1986, p.56). English translation of these pages by H.A. Ruger and H.M. Walker in D.E. Smith, A Source Book of Mathematics, McGraw-Hill Book Company, New York, 1929, pp.576-579. [1, 2]

LEWIS-BECK, M.S., AND L.B. MOHR (1976): "Evaluating Effects of Independent Variables," *Political Methodology* 3 (1), 27-47. http://www.jstor.org/stable/25791441. [4]

LONGLEY, J.W. (1981): "Least squares computations and the condition of the matrix," *Communications in Statistics - Simulation and Computation* 10, 593–615. DOI:10.1080/03610918108812237. [5]

LOVELL, M. (1963): "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association.* 58 (304), 993–1010. DOI:10.1080/01621459.1963.10480682. [6, 8]

LUBOTSKY, D., AND M. WITTENBERG (2006): "Interpretation of Regressions with Multiple Proxies," *The Review of Economics and Statistics* 88 (3), 549-62. DOI:10.1162/rest.88.3.549. [2, 4, 16, 17]

NEYMAN, J. (1923, 1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science* 5 (4), 465-480. DOI:10.1214/ss/1177012031. [3]

PEARL, J. (2000): "*Causality: Models, Reasoning, and Inference*," , 2^{nd} ed. (Cambridge: Cambridge University Press) [3, 9, 14]

PEARL, J. (2001): "*Direct and Indirect Effects*," UAI, arXiv:1301.2300 [cs.AI]. [3, 4, 14]

PEARSON, K.F.R.S. (1901): "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11), 559–572. DOI:10.1080/14786440109462720. [2]

R CORE TEAM (2021): "*R: A language and environment for statistical computing*," Version 4.1.2. (Vienna, Austria: R Foundation for Statistical Computing, 2021) https://www.R-project.org/. [12]

RUBIN, D.B (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (5), 688-701. DOI:10.1037/h0037350. [3]

SCHMIDT, E (1907): "Zur Theorie der linearen und nichtlinearen Integralgleichungen," *Mathematische Annalen* 63, 433-476. DOI:10.1007/BF01449770. [2]

SCHULTZ, H. (1928): "*Statistical Laws of Demand and Supply, with Special Application to Sugar*," (Illinois: The University of Chicago Press), OCLC 1936159. [3]

SŁOCZYŃSKI, T. (2022): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economic Statistics* 104 (3), 501-09. DOI:10.1162/rest_a_00953. [2, 8, 12, 13, 14, 15]

STATACORP (2020): "*Stata Statistical Software*," Release 17 (College Station, TX: StataCorp, LP) [12]

STIGLER, S.M. (1981): "Gauss and the Invention of Least Squares," *The Annals of Statistics* 9 (3), 465-74. http://www.jstor.org/stable/2240811. Accessed 04.27.2022. [2]

STROTZ, R.H, AND H.O.A. WOLD (1960): "Recursive vs. Nonrecursive Systems: An Attempt at Synthesis (Part I of a Triptych on Causal Chain Systems)," Econometrica 28 (2), 417-27. DOI:10.2307/1907731. [4]

WOLD, H.O.A. (1951): "Dynamic Systems of the Recursive Type: Economic and Statistical Aspects," *Sankhyā: The Indian Journal of Statistics (1933-1960)* 11 (3/4), 205–216. https://www.jstor.org/stable/25048090. Accessed 04.27.2022. [3]

WOLD, H.O.A. (1960): "A Generalization of Causal Chain Models (Part III of a Triptych on Causal Chain Systems)," *Econometrica* 28 (2), 443–63. DOI:10.2307/1907733. [3]

WOLD, H.O.A. (1963): "*Forecasting by the chain principal*," (pp. 471-497), in M. Rosenblatt (ed.) *Time Series Analysis* (New York: Wiley). [3]

WONG, Y.K. (1935): "An Application of Orthogonalization Process to the Theory of Least Squares," *The Annals of Mathematical Statistics* 6 (2), 53-75. http://www.jstor.org/stable/2957660. Accessed 04.27.2022. [2, 23]

WRIGHT, S. (1925): "*Corn and Hog Correlations*," Washington: U.S. Dept. of Agriculture Bulletin 1300. DOI:10.5962/bhl.title.108042 [3]

WRIGHT, S. (1934): "The Method of Path Coefficients," *Annals of Mathematical Statistics* 5 (3), 161–215. DOI:10.1214/aoms/1177732676. [3, 9, 14]

YULE, G.U. (1907): "On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation," *Proceedings of the Royal Society A.* 79 (529), 182–193. DOI:10.1098/rspa.1907.0028 [6]

ZHANG, F. (ED.) (2010): "*The Schur Complement and Its Applications. Numerical Methods and Algorithms*," 4. (New York City: Springer). DOI:10.1007/b105056. ISBN 0-387-24271-6. [24]

APPENDIX

Proof of Theorem 1: Existence, uniqueness, and orthogonality are shown by Wong (1935, pp. 57-59).

Coefficients a_{ij} of reduced-form parameter matrix A in (11) can be expressed as the recursive sequence

$$a_{ij} = \beta_{ij} + \sum_{n=i+1}^{j-1} a_{in}\beta_{nj},$$

for $i < j$, and zero otherwise.

Define vector $k_i = u_i'/(u_i'u_i)$ and note that $k_i'x_i = k_i'u_i = 1$.

Now, equivalence in expectation between Gram-Schmidt coefficient matrix C in (4) and the reduced-form true, underlying matrix A in (11), $E[c_{ij}] = a_{ij}$, can be shown recursively, since by assumptions (a)-(c) and Theorem 2(A) for each coefficient:

$$E[c_{ij}] = E[k_i' x_j], \tag{26}$$

$$= E[k_i'(\sum_{n=1}^{j-1} x_n \beta_{nj} + \upsilon_j)], \tag{27}$$

$$= E[\sum_{n=i}^{j-1} k_i' x_n \beta_{nj} + k_i' \upsilon_j], \tag{28}$$

$$= \beta_{ij} + \sum_{n=i+1}^{j-1} E[c_{in}]\beta_{nj}, \tag{29}$$

$$= \beta_{ij} + \sum_{n=i+1}^{j-1} a_{in}\beta_{nj}, \tag{30}$$

for $i < j$, and zero otherwise. Line (30) holds because $E[c_{ij}] = a_{ij}$ follows directly from (29) for $j = i + 1$, proving the $j = i + 2$ case, and so forth through $j = K$.

Proof of Theorem 2: (A) The orthogonality of the regressors is shown by Wong (1935, pp. 57-59).

(B) Stability of coefficients follows directly from their orthogonality (A) since coefficient covariance

$$Cov[c_{ik}, c_{jk}] = \sigma^2_{\upsilon_k} \frac{-r_{u_i,u_j}}{1 - r^2_{u_i,u_j}}$$

is linear in regressor correlation $r_{u_i,u_j} = u_i'u_j/\sqrt{(u_i'u_i\ u_j'u_j)}$ and zero for any recursive regressor pair.

(C) Unbiasedness follows from the equivalence of conditional expectations in Theorem 1.

(D) Information preservation holds since the R^2 is a linear function of squared residual u_j, which in turn is preserved in the reduced form

$$R^2_{jX_{<j}} = 1 - \tfrac{u_j'u_j}{x_j'x_j} = R^2_{jU_{<j}}.$$

(E) Zero omitted-variable bias follows directly for the later-occurring regressors $k > i$ since regressors are orthogonal by 2(A) and coefficients are unbiased by 2(C).

(F) Variances of Gram-Schmidt coefficients are lower than in OLS. Define (i) X_{-i} as the regressor set X excluding regressor x_i; (ii) $X_{<i}$ as the regressor set excluding later-determined regressors $x_i, ..., x_K$; (iii) $B \subseteq \mathcal{R}^{[j-1\times 1]}$ as the OLS coefficient vector; and (iv) $C_i \subseteq \mathcal{R}^{[K\times 1]}$ as the i^{th} Gram-Schmidt coefficient vector. Consider the non-trivial case when $\beta_{ij} \neq 0$.

Lemma 3. The coefficient of determination weakly declines as regressors are excluded:

$$R^2_{x_i\, X_{<i}} \leq R^2_{x_i\, X_{-i}}.$$

This lemma is true for any regressor order, though for convenience, we specify order-of-exclusion in terms of regressors determined after x_i. The proof is standard and omitted here.

The variance of OLS coefficient b_{ij} is a function of the i^{th} diagonal element of the variance-covariance matrix, which can be expressed, by virtue of the Schur Complement, in terms of the j^{th} regression coefficient of determination:

$$\begin{aligned} Var[b_{ij}] &= \sigma_{\upsilon_{u_j}} (X'X)^{-1}_{ii}, \\ &= \sigma_{\upsilon_{u_j}} (x_i'x_i - B'_{-i}(X'_{-i}X_{-i})B_{-i})^{-1}, \\ &= \sigma_{\upsilon_{u_j}} (x_i'x_i(1 - R^2_{x_i\, X_{-i}}))^{-1}. \end{aligned}$$

The variance of Gram-Schmidt coefficient c_{ij} is by 2(A) a function of the residual u_i, which by 2(A) and Lemma 3 is:

$$\begin{aligned} Var[c_{ij}] &= \sigma_{\upsilon_{u_j}} (u_i'u_i)^{-1}, \\ &= \sigma_{\upsilon_{u_j}} (x_i'x_i - C_i'(X'X)C_i)^{-1}, \\ &= \sigma_{\upsilon_{u_j}} (x_i'x_i(1 - R^2_{x_i\, X_{<i}}))^{-1}, \\ &\leq \sigma_{\upsilon_{u_j}} (x_i'x_i(1 - R^2_{x_i\, X_{-i}}))^{-1}. \end{aligned}$$

The result holds with strict inequality for $i = 1, ...j-2$ and with strict equality for $i = j-1$ because the excluded regressor set is identical between the terminal Gram-Schmidt and the OLS coefficient.

(G) Gram-Schmidt coefficients have lower variance than OLS when a later-determined irrelevant variable x_k is included because the recursive efficiency gain (F) from Lemma 3 is preserved:

$$R^2_{x_i\, X_{<i+k}} \leq R^2_{x_i\, X_{-i+k}}.$$

Proof of Lemma 2: (A) Early-treatment Gram-Schmidt is the ATTE. Recall from Lemma 1(A) the GSLS coefficient c_{xy} is a convex combination of partial linear effects,

$$c_{xy} = \omega_1 c_{xy(APLE,1)} + \omega_0 c_{xy(APLE,0)}. \tag{31}$$

The early-treatment model is given by reduced form equation (10) expressed below in terms of x:

$$y = (\beta_{xy} + \beta_{xd}\beta_{dy})x + \beta_{dy}\upsilon_d + \upsilon_y, \tag{32}$$

where c_{xy} is the coefficient estimate of $\beta_{xy} + \beta_{xd}\beta_{dy}$. The linear probability of treatment $p(\upsilon_d)$ is defined as the projection

$$L[x \mid \upsilon_d] = \alpha_p + \beta_p \upsilon_d. \tag{33}$$

Because x and υ_d are orthogonal by Theorem 2(A), $\beta_p = 0$ and the projection is just a regression on the ones vector. The remaining intercept term α_p is then the unconditional sample mean $E[x]$.

Incorporating this result, the projection of y on the probability of treatment from (13),

$$\begin{aligned} L[y \mid p(\upsilon_d), x = 1] &= \alpha_1 + \gamma_1 p(\upsilon_d), &(34)\\ &= \alpha_1 + \gamma_1 E[x], &(35)\\ &= E[y \mid x = 1], &(36) \end{aligned}$$

is now the regression of y on a constant and equal to the mean outcome when x is positive.

The average partial linear effect of x on y from (14) can now be simplified

$$\begin{aligned} c_{xy(APLE,1)} &= (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)E[p(\upsilon_d) \mid x = 1], &(37)\\ &= (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)E[x], &(38)\\ &= (\alpha_1 + \gamma_1 E[x]) - (\alpha_0 + \gamma_0 E[x]), &(39)\\ &= E[y \mid x = 1] - E[y \mid x = 0]. &(40) \end{aligned}$$

The same is true for $c_{xy(APLE,0)}$ and thus for the convex combination of the two, completing the proof.

(B) Early-treatment Gram-Schmidt ATTT equals ATTU.

From Lemma 1(B) and the identity of the Graham-Schmidt coefficient c_{xy} as total effect $(\beta_{xy} + \beta_{xd}\beta_{dy})$, late-treatment Graham-Schmidt is the convex combination of the average total treatment effect on the treated and untreated,

$$c_{xy} = \omega_1 c_{xy(ATTT)} + \omega_0 c_{xy(ATTU)}. \tag{41}$$

From Lemma 2(A) we have $c_{xy(APLE,0)} = c_{xy(APLE,1)} =$ ATTE. Assumptions (i) and (j) and Lemma 1(B) imply $c_{xy(ATTT)} = c_{xy(APLE,1)}$ and $c_{xy(ATTU)} = c_{xy(APLE,0)}$, completing the proof.

Proof of Theorem 3: (A) GSLS regressors are orthogonal across blocks. Define $U_{(m)}$ as the residual block $m < n$ and $U_{<j(-m)}$ the set of all residuals excluding later-determined residuals $u_j, ..., u_K$ as well as residual block $U_{(m)}$. Theorem 3(A) then follows from Theorem 2(A) by the inclusion of residual block $U_{(m)}$ in the OLS regression of the first regressor, $x_{j(n)}$, in block n:

$$x_{j(n)} = U_{<j(-m)}C_{<j(-m)} + U_{(m)}C_{(m)} + u_{j(n)}.$$

This holds for all regressors in block n, since we may reorder any simultaneous regressor arbitrarily to be the first regressor in the block.

(B) Stability of coefficients across blocks follows directly from their orthogonality across blocks (A) and their stability from Theorem 2(B) *mutatis mutandis*.

(C) Unbiasedness. Coefficients $a_{i(m)j(n)}$ of reduced-form parameter matrix A in (23) can be expressed as the recursive sequence

$$a_{i(m)j(n)} = \beta_{i(m)j(n)} + \sum_{h=i+1}^{j-1} a_{i(m)h(k)}\beta_{h(k)j(n)},$$

for $m < n$, and zero otherwise.

Define: (i) regressor sub-matrix $X_{(m)}$ to include all s simultaneous regressors in block m, $s = 1, ..., K$; (ii) coefficient sub-vector $c_{(m)j(n)}$ to be the corresponding coefficients in the j^{th} GSLS regression; and (iii) matrix $k_{(m)} = (U'_{(m)}U_{(m)})^{-1}U_{(m)}$. Note $k'_{(m)}X_{(m)} = k'_{(m)}U_{(m)} = I$, the $[s \times s]$ identity matrix.

Equivalence in expectation between GSLS coefficient matrix C and the reduced-form, true, underlying matrix A in (23), namely $E[c_{(m)j(n)}] = a_{(m)j(n)}$, can now be shown recursively since by the revised assumptions (a)-(c) and Theorem 3(A), we have for each coefficient:

$$E[c_{(m)j(n)}] = E[k'_{(m)}\, x_{j(n)}], \tag{42}$$

$$= E[k'_{(m)}(\sum_{h=1}^{j-1} x_{h(k)}\, \beta_{h(k)j(n)} + \upsilon_{j(n)})], \tag{43}$$

$$= E[k'_{(m)}\, X_{(m)}\, \beta_{(m)j(n)} \tag{44}$$

$$+ \sum_{h=i+s}^{j-1} k'_{(m)}\, x_{h(k)}\, \beta_{h(k)j(n)} + k'_{(m)}\upsilon_{j(n)}], \tag{45}$$

$$= \beta_{(m)j(n)} + \sum_{h=i+s}^{j-1} E[c_{(m)h(k)}]\beta_{h(k)j(n)}, \tag{46}$$

$$= \beta_{(m)j(n)} + \sum_{h=i+s}^{j-1} a_{(m)h(k)}\beta_{h(k)j(n)}, \tag{47}$$

for $m < n$, and zero otherwise. Line (47) holds because $E[c_{(m)h(k)}] = a_{(m)h(k)}$ is shown by equation (46) for $j = i + 1$, in turn proving the $j = i + 2$ case and so forth until $j = K$.

(D) Information preservation follows from the preservation of reduced-form residuals in Theorem 2(D) because we may arbitrarily order regressors in the simultaneous block (n) such that x_j is the first regressor in the block.

(E) Zero omitted-variable bias follows directly for regressors in later-occurring blocks $k(h) > i(m)$ since regressors are orthogonal across blocks 3(A) and coefficients are unbiased by 3(C).

(F) GSLS coefficients have lower variance than OLS. Define (i) $X_{-i(m)}$ as the regressor set X excluding regressor $x_{i(m)}$; (ii) $X_{<i(m)}$ as the regressor set excluding regressors in later-determined blocks $m + 1, ..., M$; (iii) $B \subseteq \mathcal{R}^{[j-1\times 1]}$ as the OLS coefficient vector; and (iv) $C_{i(m)} \subseteq \mathcal{R}^{[K\times 1]}$ the i^{th} as the GSLS coefficient vector. Consider the non-trivial case when $\beta_{i(m)j(n)} \neq 0$ and order block m so that $x_{i(m)}$ is the last regressor in the block.

The variance of OLS coefficient $b_{i(m)j(n)}$ can now be expressed in terms of the coefficient of determination by the Schur Complement:

$$Var[b_{i(m)j(n)}] = \sigma_{\upsilon_{u_{j(n)}}}(x'_{i(m)}x_{i(m)}(1 - R^2_{x_{i(m)}\, X_{-i(m)}}))^{-1}.$$

The variance of GSLS coefficient $c_{i(m)j(n)}$ is a function of the residual u_i by virtue of 3(A), in which by 3(A) and Lemma 3,

$$\begin{aligned} Var[c_{i(m)j(n)}] &= \sigma_{v_{u_{j(n)}}} (u'_{i(m)} u_{i(m)})^{-1}, \\ &= \sigma_{v_{u_{j(n)}}} (x'_{i(m)} x_{i(m)} (1 - R^2_{x_{i(m)}\, X_{<i(m)}}))^{-1}, \\ &\leq \sigma_{v_{u_{j(n)}}} (x'_{i(m)} x_{i(m)} (1 - R^2_{x_{i(m)}\, X_{-i(m)}}))^{-1}. \end{aligned}$$

The result holds with strict inequality for $m = 1, ...M - 1$ and strict equality for $m = M$ because, in the terminal block, the excluded regressor set in GSLS is identical to that in OLS.

(G) GSLS coefficients $c_{i(m)j(n)}$ have lower variance than OLS coefficients $b_{i(m)j(n)}$ when irrelevant variable $x_{k(h)}$ is included in a later-occurring block $m < h$ because recursive efficiency gain 2(F) from Lemma 3 is preserved:

$$R^2_{x_{i(m)}\, X_{<i(m)+k(h)}} \leq R^2_{x_{i(m)}\, X_{-i(m)+k(h)}}.$$