

Organizational Change and Work Design

Organizations, Personnel, and Internal Labor Markets

Week 9

Sai Zhang

The same reform can change different work systems

Central question

Why do similar changes in practices, technology, and work arrangements perform differently across organizations?

- A tool can accelerate drafting while leaving review as the production bottleneck.
- An owner can gain from a practice that imposes uncompensated implementation costs on workers.
- A successful launch can fade when support ends or routines lose maintenance.
- A recorded productivity gain can reflect real output, time savings, selection, or changed logging.

Hoffman–Stanton (2025) supplies the field overview. Week 9 evaluates work-system reforms; aggregate automation and national employment effects belong to the technology course.

Learning goals: specify the reform and verify its consequences

- Separate assignment, access, competence, use, and sustained integration.
- Distinguish component, conditional, marginal, bundle, and interaction effects.
- Measure bottlenecks, implementation burdens, transition costs, and maintenance.
- Design evaluations with independent quality, worker outcomes, spillovers, and attrition.

Core points

A positive bundle effect does not identify every component. More activity is not necessarily more output. A short-run gain does not establish persistence or positive net value.

Research deliverable: a reform evaluation with two competing mechanisms, an explicit horizon, and a component/outcome ledger extending beyond headline productivity.

Four linked records make the intervention concrete

Record	Minimum fields	Main distinction
Components	Initial eligibility, tool version, permissions, support, authority, compensation	Shared reform name versus identical content
Implementation	Announcement, assignment, activation, competence, use, exceptions, maintenance	Delivery and use versus sustained routine
Production	Task arrivals, handoffs, review, queues, completion, rework, resources	Task speed versus verified throughput and quality
Workers	Hours, pay, unpaid effort, burden, control, exits, follow-up response	Firm performance versus worker conditions and composition

Retain the initial population and the control regime. Record updates, suspensions, support withdrawal, and control crossover; these change treatment histories.

Implementation evidence does not automatically identify mediation

$$Z_i \rightarrow A_{it} \rightarrow U_{it} \rightarrow \text{sustained routine}, \quad R_{it} = \text{complementary redesign.}$$

- Z_i : assigned offer; A_{it} : actual access; U_{it} : use. None is interchangeable with competence.
- A login or completed checklist can show delivery without proving integration into productive work.
- Users select tasks and effort after assignment; a users/nonusers comparison is not the offer ITT.
- Training, reassurance, support, or expectations can affect outcomes without logged use, violating a use-IV exclusion restriction.

Preserve descriptive first stages. Identifying how much of an effect runs through use requires additional design or mediation assumptions.

Complementarity is a comparison of returns across configurations

$$\mathcal{C} = \mathbb{E}[v_i(1, 1) - v_i(1, 0) - v_i(0, 1) + v_i(0, 0)].$$

- $v_i(a, r)$: verified output or net value for unit i , with tool a and redesign r .
- $\mathcal{C} > 0$ means one component has a higher return when the other is present.
- It does not imply either standalone component is beneficial.
- The result depends on the outcome scale: output levels, log output, and net surplus are different objects.

Milgrom–Roberts (1990); Athey–Stern (2026), *Journal of Organization Design*, online September 4. The latter framework circulated as NBER 6600 in 1998; it is one intellectual work, not two independent studies.

Four cells separate a bundle gain from component effects

Original output example	No redesign $r = 0$	Redesign $r = 1$
No tool $a = 0$	100	103
Tool $a = 1$	96	112

tool alone = -4 , redesign alone = 3 ,

bundle = $112 - 100 = 12$, $\mathcal{C} = 112 - 96 - 103 + 100 = 13$.

Invented values on one fixed output scale. The bundle works in this example even though the tool alone reduces output. Costs and worker outcomes are not yet included.

A factorial design supports eight distinct contrasts

Contrast	Formula in cell means v_{ar}	Value
Tool without redesign	$v_{10} - v_{00}$	-4
Tool with redesign	$v_{11} - v_{01}$	9
Redesign without tool	$v_{01} - v_{00}$	3
Redesign with tool	$v_{11} - v_{10}$	16
Bundle	$v_{11} - v_{00}$	12
Interaction	$v_{11} - v_{10} - v_{01} + v_{00}$	13
Tool, equal-weighted marginal	$[(-4) + 9]/2$	2.5
Redesign, equal-weighted marginal	$(3 + 16)/2$	9.5

Original arithmetic. An equal-weighted marginal effect averages over the other factor; it is neither the standalone nor the bundle effect. State cell weights before interpreting a regression coefficient.

A two-arm bundle trial leaves its decomposition unidentified

$$\underbrace{-4}_{\text{tool alone}} + \underbrace{3}_{\text{redesign alone}} + \underbrace{13}_{\text{interaction}} = 12 = \underbrace{5}_{\text{tool alone}} + \underbrace{7}_{\text{redesign alone}} + \underbrace{0}_{\text{interaction}} .$$

- Observing only 00 and 11 can identify the bundle under a credible comparison.
- Both decompositions fit that same bundle gain; the mixed cells are missing.
- Co-adoption can arise because a common unobserved capability makes both practices attractive.
- A factorial needs feasible standalone regimes. Removing necessary safety review merely to fill a cell is not a meaningful treatment.

Original observational-equivalence example; Athey–Stern (2026). Better prediction of co-adoption does not by itself identify complementarity.

Systems-of-practices evidence still requires selection restrictions

	Ichniowski–Shaw–Prennushi	Bresnahan–Brynjolfsson–Hitt
Setting	Steel finishing lines	Firm IT, work organization, and skills
Variation	Observed configurations of HR practices	Joint patterns in technology and workplace organization
Target	Workplace systems and line productivity	Empirical complementarity and skill-demand patterns
Limitation	Practice bundles are not randomly assigned	Co-adoption can reflect shared unobserved returns

Ichniowski et al. (1997), AER 87(3):291–313; Bresnahan et al. (2002), QJE. Production detail strengthens interpretation, but does not convert organizational choice into random assignment.

A faster task need not raise the work system's throughput

$$Q = \min\{C_{\text{draft}}, C_{\text{review}}\}.$$

Original regime	Draft capacity	Review capacity	Throughput
Neither	10	10	10
Tool only	20	10	10
Redesign only	10	16	10
Both	20	16	16

Invented two-stage bottleneck with sufficient demand and comparable quality. The bundle gain is six before costs. Saved task time may reduce burden even when output does not rise; measure arrivals, queues, handoffs, rework, and where time goes.

Information and communication can move authority differently

- **Setting:** Bloom–Garicano–Sadun–Van Reenen, technologies and firm organization.
- **Variation:** distinctions between information and communication technologies in organizational data.
- **Target:** local problem-solving, escalation, discretion, and coordination channels.
- **Limitation:** the evidence is not a randomized claim that every digital tool decentralizes authority.

Make the candidate complement observable

Review capacity, training, data permissions, staffing, and authority should imply different queues, handoffs, and decision records.

Bloom et al. (2014), The Distinct Effects of Information Technology and Communication Technology on Firm Organization, Management Science.

Bloom and coauthors assign assistance with management practices

- **Setting:** textile firms and plants in India.
- **Intervention/comparison:** randomized management-consulting assistance with a collection of practices.
- **Implementation:** observe which recommended routines are adopted and how production records change.
- **Outcomes/horizon:** practice adoption and plant performance during implementation and subsequent observed follow-up, including quality, efficiency, and inventories.

Bloom, Eifert, Mahajan, McKenzie, and Roberts (2013), Does Management Matter? Evidence from India, QJE 128(1):1–51. Required reading; use the published study rather than an older working-paper headline.

The management package is assigned; each routine is not

- The trial supports the effect of assistance under its implementation process.
- A correlation between adoption of one practice and better performance does not separately randomize that practice.
- Distinguish firm-level treatment assignment, selected experimental plants, and daily records: transactions do not add independent firms.
- Consulting can improve recordkeeping; independent verification helps distinguish better production from better recording.

Design contract for the required reading

Map assistance → implementation → verified outcomes, including within-firm diffusion, control practices, and the end of external support.

Bloom et al. (2013). Component mechanisms, support persistence, and the effect of universal rollout need additional comparisons beyond the original package contrast.

A follow-up distinguishes retained routines from their causes

- **Setting:** Bloom–Mahajan–McKenzie–Roberts revisit the earlier Indian management intervention.
- **Variation/horizon:** original assignment histories observed after the initial assistance period.
- **Outcomes:** many adopted practices disappear, while a treatment–control practice gap and within-firm diffusion remain.
- **Limitation:** reported manager-turnover and director-time explanations are not randomized mechanism effects.

Bloom et al. (2020), Do Management Interventions Last? AEJ: Applied Economics. A retained-practice count is not a retained-welfare fraction; closures and staff changes belong in the follow-up ledger.

Retaining half the practices need not retain half their value

invented additive practice values = (8, 1, 1, 0).

retain the first two: $\frac{2}{4} = 50\%$ of practices, $\frac{8+1}{8+1+1+0} = 90\%$ of value.

- Equal practice counts implicitly give each routine the same weight.
- Losing a critical complement could instead destroy much more than half the gain.
- Follow actual routines, verified production, quality, people, and maintenance resources.

Original additive illustration, not a universal production assumption or a finding from the management follow-up. Practice measurement and economic value require separate justification.

A consulting offer can deliver tailored advice or different modes

	Bruhn–Karlan–Schoar	Iacovone–Maloney–McKenzie
Setting	Mexican small and medium enterprises	Colombian firms
Variation	Randomized consulting offer	Randomized individual and group-based consulting
Target	Program effects with advice tailored to firm needs	Delivery mode, implementation, outcomes, and costs
Limitation	A common offer is not an identical practice bundle	Compare arms directly; significant versus nonsignificant is not a test of their difference

Bruhn et al. (2018), JPE; Iacovone et al. (2022), REStud. The relevant treatment includes delivery and support, not only the program label.

Historical transfer evidence supplies a different counterfactual

- **Setting:** Giorcelli, historical management and technology transfer.
- **Variation:** an unexpected program budget cut affecting the transfer comparison.
- **Target/horizon:** long-run effects of managerial knowledge, equipment, and their combinations under that historical design.
- **Limitation:** this is not a contemporary researcher-randomized factorial intervention.

Giorcelli (2019), The Long-Term Effects of Management and Technology Transfers, AER. State the historical eligibility and counterfactual before transporting a result to a new reform.

Atkin and coauthors make implementation incentives observable

- **Setting:** soccer-ball producers in Pakistan; a material-saving cutting technology.
- **First intervention:** randomized technology provision, followed by observation of adoption and implementation.
- **Implementation problem:** a piece-rated worker can bear slower cutting and lower income while the owner saves material.
- **Target:** whether incentive misalignment helps explain slow adoption of a potentially valuable technique.

David Atkin, Azam Chaudhry, Shamyla Chaudry, Amit K. Khandelwal, and Eric Verhoogen (2017), Organizational Barriers to Technology Adoption, QJE 132(3):1101–1164. Required reading.

The worker-payment experiment follows the technology intervention

Stage	Eligible comparison	What changes
Technology experiment	Randomized technology provision	Access to the material-saving technique
Later payment experiment	Randomized offer among earlier technology recipients	One-time payment conditional on demonstrating competence in front of the owner

- The later intervention combines incentives, demonstration, and communication.
- It is not a fully crossed technology $\times$ permanent compensation reform.
- The sequence supports an implementation-incentive explanation while leaving intermediate beliefs and mechanisms partly unresolved.

Atkin et al. (2017). Track the two eligibility sets and follow-up windows separately; worker objections may reveal real costs or quality concerns rather than irrational resistance.

A transfer can align incentives without creating extra resources

Original monetary-unit example	No compensation	Transfer of 5
Owner: material saving less transfer	12	7
Worker: transfer less real burden	-4	1
Combined gain	8	8

- The transfer cancels in joint accounting; the real implementation burden does not.
- Measure setup time, pay schedules, risk, skill demands, and worker control.
- Resistance can be informative when the implementer does not share the gain.

Invented sufficient-compensation example. Real workers' disutility is not necessarily known, monetized, or transferable without limits. This is not a calibration of the Pakistan experiment.

Organizational capability requires investment and maintenance

$$k_{i,t+1} = (1 - \delta_i)k_{it} + u_{it}, \quad Q_{it} = F(A_{it}, R_{it}, k_{it}, L_{it}, X_{it}).$$

- k : stock of routines or integration; $\delta \in [0, 1]$: depreciation; $u \geq 0$: maintenance or investment.
- A : actual tool access; R : redesign; L : labor input; X : other inputs and operating conditions.
- From $k_0 = 1$ and $\delta = .2$, no maintenance gives .8, .64, .512, .4096.
- Maintenance $u = .2$ each period holds the stock at one, but consumes resources.

Original accounting organizer. The stock is a hypothesis, not an identified residual named organizational capital. Staff changes, documentation, training, and supervision can matter for persistence.

A positive endline effect need not repay the transition costs

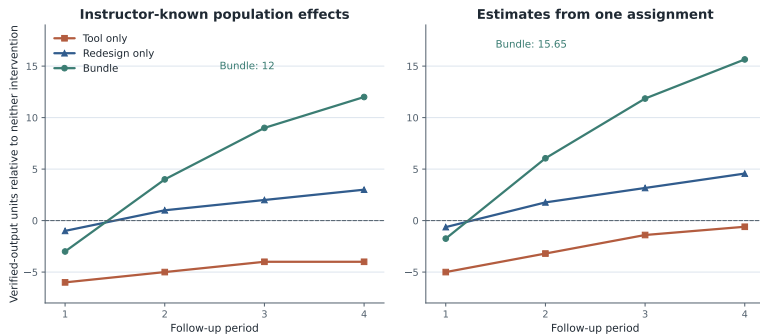
$$\Delta V_H = -C_0 + \sum_{t=1}^H \beta^t [\Delta B_t - C_t] + \beta^H \Delta TV_H.$$

- H : horizon; $0 < \beta \leq 1$: discount factor; C_0 : setup cost; C_t : running costs.
- ΔB_t : benefit change; ΔTV_H : terminal-value change. Specify whose values and which units.
- Invented benefits $(-3, 4, 9, 12)$, $\beta = .95$, $C_0 = 20$, $C_t = 0$, $\Delta TV_H = 0$ give

$$\Delta V_4 = -20 + .95(-3) + .95^2(4) + .95^3(9) + .95^4(12) = -1.74955.$$

Original finite-horizon illustration. Firm profit, worker welfare, and joint value are different accounts. An assumed terminal value can reverse the result and must be defended.

Transition paths distinguish late gains from realized evidence

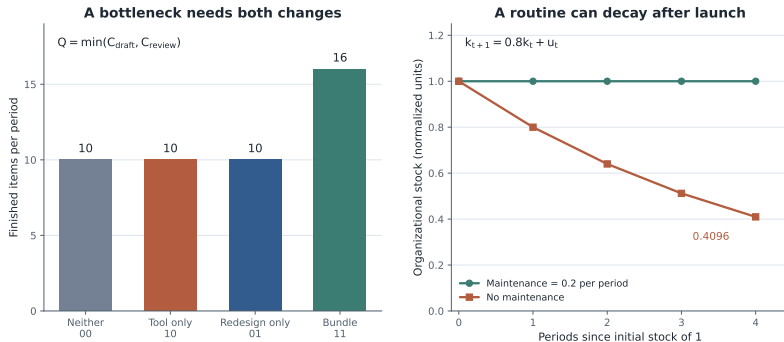


12 establishments • 3 independent blocks • baseline-adjusted estimates • no confidence bands

Original synthetic teaching example • one assignment need not reproduce population effects

Original synthetic teaching example: instructor-known population effects and one realized allocation are separately labeled. Three independent blocks; no normal confidence bands. The productivity J-curve framework (Brynjolfsson–Rock–Syverson, 2021) does not guarantee that early losses reverse.

Bottlenecks and maintenance imply different reform requirements



Separate analytical examples: capacity gains precede costs; maintaining routines consumes resources.

Original synthetic teaching examples • known model quantities, not intervention estimates

Original synthetic teaching examples on separate scales. A workflow complement raises throughput from 10 to 16; ongoing maintenance prevents stock decay from 1 to .4096. Neither pattern identifies the corresponding mechanism in a real organization without additional evidence.

Generative AI at Work studies a defined production workflow

- **Setting:** Brynjolfsson–Li–Raymond, AI assistance in customer support.
- **Comparison:** staggered introduction of the assistant, not an individually randomized trial.
- **Implementation/horizon:** follow access and outcomes around rollout within the observed support workflow.
- **Outcomes:** higher issues resolved per hour, with heterogeneous speed and quality responses across worker groups.

Brynjolfsson, Li, and Raymond (2025), Generative AI at Work, QJE 140(2):889–942. Required reading. Use this published version; do not mix older sample definitions or headline statistics.

A workflow gain does not identify the return to redesigning the firm

- Rollout comparisons need a credible untreated path, including timing and co-changing practices.
- Keep resolution counts, handling speed, quality, working time, and worker heterogeneity separate.
- Tool effects may depend on review, task selection, training, or discretion that is not independently assigned.
- A support-task productivity gain is not a measured effect on total firm value or economy-wide labor demand.

Design contract for the required reading

Specify eligibility, access timing, control workers, task mix, independent quality, and the horizon over which the comparison is supported.

Brynjolfsson–Li–Raymond (2025). Observed heterogeneity can suggest mechanisms; it does not independently randomize every proposed complement.

Individual AI access can change activity before the work system changes

- **Setting:** Dillon and coauthors, knowledge workers across multiple firms.
- **Variation:** randomized access to an integrated generative-AI assistant.
- **Targets:** offer effects and separately qualified use estimates for email and other work patterns.
- **Limitation:** application time is not verified output; organizational complements are not separately randomized.

Dillon, Jaffe, Immorlica, and Stanton, Shifting Work Patterns with Generative AI, NBER 33795, November 2025 revision; arXiv 2504.11436v4. Earlier abstract/sample/usage numbers are superseded.

Stable earnings can coexist with substantial task reorganization

- **Setting:** Humlum–Vestergaard, Danish adoption surveys linked to administrative labor records.
- **Variation:** difference-in-differences comparisons; adoption is not randomly assigned.
- **Targets/horizon:** early task reorganization alongside earnings and recorded hours.
- **Limitation:** precise null estimates for those administrative outcomes do not establish no task change or no future effects.

Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI, NBER 33777, initially issued 2025, revised March 2026. Earlier title is provenance for the same study, not another independent result.

Time accounting separates reallocation, saving, and logging changes

Original example	Before	After	Implication
Email hours, with core hours rising	$8 + 30 = 38$	$6 + 32 = 38$	Reallocation
Email hours, core hours unchanged	$8 + 30 = 38$	$6 + 30 = 36$	Less recorded time
Output / recorded hours	$200/100 = 2$	$200/80 = 2.5$	Rate +25%

- Neither email example establishes output without a production measure.
- The rate gain occurs with unchanged output; distinguish real saved hours from changed logging.
- Missing app time is not necessarily leisure: track off-platform and unpaid work where feasible.

Invented numerical diagnostics. Keep raw numerators and denominators, overlapping activity windows, and coverage changes visible.

Hybrid work is an offered arrangement with specified days

- **Setting:** Bloom–Han–Liang, Trip.com employees eligible for a hybrid arrangement.
- **Assignment:** an option to work from home on specified days, not identical forced attendance.
- **Outcomes:** retention, satisfaction, performance grades, promotions, and code measures.
- **Horizon/limitation:** the control group later receives access; later comparisons have different treatment histories.

Bloom, Han, and Liang (2024), Nature 630:920–925. Focused companion discussion on offer, compliance, and horizon; this is distinct from redoing Week 1's measurement lab.

Control crossover changes the long-run comparison

Period	Assignment histories	Interpretation
Original trial	Hybrid offer versus control regime	ITT of the specified offer over that trial window
After controls gain access	Earlier versus later access	Original-assignment history effect, not permanently treated versus never treated
Performance analysis	Prespecified endpoints and margins	Equivalence is specific to those outcomes and scales
Promotion analysis	Estimated differences with uncertainty	Nonsignificance alone does not establish equivalence

Bloom–Han–Liang (2024). Retain announcement, access, attendance, and crossover dates rather than labeling every later observation with one timeless hybrid indicator.

Remote-work experiments and closure comparisons differ

	Bloom–Liang–Roberts–Ying	Emanuel–Harrington
Setting	Eligible call-center volunteers	Pre-existing remote and onsite workers around office closure
Variation	Random assignment of a particular home-work arrangement	Closure-based change under an explicit counterfactual
Target	Arrangement effects for that selected workforce	Remote-work treatment versus selection
Limitation	Later reselection is not the original random contrast	Closure does not randomize all concurrent organizational changes

Bloom et al. (2015), QJE; Emanuel–Harrington (2024), AEJ: Applied Economics. Compare eligibility, task interdependence, managerial practices, counterfactual, and horizon before comparing estimates.

Geographic flexibility is distinct from pandemic home working

	Choudhury–Foroughi–Larson	Gibbs–Mengel–Siemroth
Setting	Patent examiners already working from home	IT professionals with personnel and activity records
Variation	Work-from-home to work-from-anywhere	Before/during pandemic work-from-home patterns
Target	Additional geographic flexibility and productivity	Working time, coordination, and output measurement
Limitation	Not an office-to-home treatment	Concurrent pandemic shocks remain; code availability is not unrestricted employer data

Choudhury et al. (2021), SMJ; Gibbs et al. (2023), JPE Microeconomics. Different treatment contracts need not deliver the same result.

Scheduling bundles bring worker conditions into the evaluation

	Kesavan and coauthors	Kelly and coauthors
Setting	Gap retail stores	Work, Family, and Health Network workplace intervention
Variation	Randomized responsible-scheduling package	Group-randomized supervisor support and work-time control redesign
Targets	Store performance and staffing practices	Perceived control and work-family outcomes
Limitation	Components are not separately identified; journal means differ from early-report medians	A bundle, not isolated WFH permission or supervisor support

Kesavan, Lambert, Williams, and Pendem (2022), Management Science; Kelly et al. (2014), American Sociological Review. Worker conditions should be measured directly rather than inferred from sales or retention.

Independent outcome checks protect against recording-only gains

Original logging example	Before	After	Change
Recorded completed actions	100	150	+50%
Independently audited deliverables	100	100	0%

- Prespecify verified production plus independent quality and worker conditions.
- Maintain old/new definitions on an overlap sample when software changes recording.
- An index can manage multiplicity; it cannot make incompatible welfare units commensurate.

Invented recording-only diagnostic. A primary endpoint should not hide quality, time, cost, or worker outcomes that move differently.

Selective exit can raise mean productivity while total output falls

Original workforce	Initial	Remaining workers
Individual outputs	(80, 100, 120, 140)	(120, 140)
Mean	110	130
Total	440	260

- Nobody improves; the workforce denominator changes.
- Track initial units, exits, closures, and response. Missing destination outcomes are not zero.
- Lee bounds need monotone selection and relevant assignment restrictions; reform-induced exits can violate them.

Original composition example; Lee (2009). A closed establishment can have zero site output while former workers' income and welfare remain unobserved.

Factorial coefficients must refer to assigned components

$$\Delta Y_i = \lambda_{b(i)} + \tau_A Z_i^A + \tau_R Z_i^R + \tau_{AR} Z_i^A Z_i^R + \varepsilon_i.$$

- $\Delta Y_i = Y_{i,H} - Y_{i,0}$; Z_i^A, Z_i^R : assigned tool and redesign; $\lambda_{b(i)}$: baseline-block effect.
- Under balanced randomized cells: τ_A is the tool effect without redesign; $\tau_A + \tau_{AR}$ is its effect with redesign.
- The bundle is $\tau_A + \tau_R + \tau_{AR}$; the equal-weighted tool marginal effect is $\tau_A + \tau_{AR}/2$.
- Publish cell means and weights. Actual access or use must not silently replace the assigned factors.

The balanced teaching allocation has one unit per cell per block. Unequal probabilities, sizes, or population weights require explicit standardization; a software coefficient is not automatically the intended average.

A modern event-study estimator does not validate rollout timing

- Define announcement, delivery, first login, intensive support, stable routine use, suspension, and crossover dates.
- Backdating adoption to first successful use selects on implementation outcomes.
- Staggered comparisons require a credible untreated trend, supported controls and horizons, and explicit anticipation assumptions.
- Absorbing-adoption methods do not automatically handle reversals, changing versions, endogenous timing, or spillovers.

Callaway–Sant’Anna (2021); Roth (2022). A nonsignificant pretrend test is not a certificate, and selecting specifications because they pass can distort inference. State the institutional counterfactual.

An individual pilot and full rollout are different policy objects

$$Y_i(z, e) = b_i + 2z + 4e.$$

$$Y_i(1, e) - Y_i(0, e) = 2, \quad Y_i(1, 1) - Y_i(0, 0) = 6.$$

- z : own assignment; e : a defined peer-exposure share; b_i : baseline outcome.
- Shared collaborators, review capacity, supervisors, and infrastructure can transmit effects.
- Randomize feasible rollout intensities or organizational units and track untreated collaborators.
- Larger clusters may limit contamination while reducing the number of independent assignments.

Original spillover model. The two-unit direct effect and six-unit full-rollout effect differ because the interventions differ, not because one estimator must be biased.

Nonsignificance does not establish an acceptably small effect

Normal illustration	Estimate	SE	90% interval
Precise	.1	.2	$[-.229, .429]$
Imprecise	.1	2	$[-3.190, 3.390]$

- With a prespecified equivalence margin $[-1, 1]$, only the first interval lies wholly inside.
- Both zero-effect tests can be nonsignificant.
- Choose a meaningful margin for a specified endpoint, scale, population, and horizon before seeing outcomes.

Invented normal-approximation example, using $1.645 \times \text{SE}$. This interval procedure is not justified automatically for the three-block laboratory.

Research opportunities turn implementation stories into comparisons

Proposed variation	Competing explanations	Discriminating outcome
Tool and feasible review redesign	Faster drafting versus review bottleneck	Queues, throughput, independent errors, burden
Continued support at planned withdrawal	Persistent routines versus ongoing assistance	Routine audits, costs, staff changes, quality
Implementation compensation	Lack of knowledge versus uncompensated burden	Competence, pay/time costs, use, quality
Coordinated rollout intensity	Individual gains versus shared availability	Untreated collaborators and joint deliverables
Parallel audit during system migration	Real improvement versus changed logging	Old/new definitions and audited production

Proposed designs, not completed studies or promises of data access. Each needs feasible treatment arms, two testable mechanisms, spillover boundaries, and initial-cohort follow-up.

Research Lab: assign all four regimes within each baseline block

12 establishments, 3 blocks of 4, $4! = 24$ allocations per block, $24^3 = 13,824$.

- Each block receives one establishment in each regime: 00, 10, 01, 11.
- Fix all potential outcomes before drawing assignments with a separate random stream.
- Follow the same establishments through a baseline and four post periods.
- Complete follow-up and no cross-establishment interference are explicit synthetic assumptions.

Original synthetic factorial rollout, motivated by Atkin et al. and the management studies. It does not reproduce their data, protocol, or estimates. The 13,824 legal allocations do not create more than three independent blocks.

The lab fixes heterogeneous potential outcomes before assignment

Period	Tool alone	Redesign alone	Interaction	Bundle
1	-6	-1	4	-3
2	-5	1	8	4
3	-4	2	11	9
4	-4	3	13	12

- These are instructor-known population means; centered readiness heterogeneity preserves them.
- Common and block-specific untreated changes remain in the observed panel.
- Utilization depends on readiness, time, and redesign after assignment. The worker-conditions index is a separate, nonmonetized endpoint.

Original synthetic teaching parameters. The realized sample estimates need not equal these means. Latent readiness and potential-outcome schedules stay in the instructor directory.

Reconstruct contrasts from observed cells before comparing to truth

$$\Delta Y_{ib,t} = Y_{ib,t} - Y_{ib,0}, \quad \hat{\delta}_{b,t}(c) = \sum_{ar} c_{ar} \Delta Y_{b,ar,t}, \quad \hat{\delta}_t = \frac{1}{3} \sum_{b=1}^3 \hat{\delta}_{b,t}.$$

- Each block has one observed establishment per cell at every horizon.
- For cell order (00, 10, 01, 11): bundle weights are $(-1, 0, 0, 1)$; interaction weights are $(1, -1, -1, 1)$.
- Reconstruct all eight contrasts and four periods from the initial roster and observed panel.
- Baseline adjustment can improve precision; random assignment supplies identification.

Student inputs: `data/teaching/establishments.csv` and `data/teaching/establishment_period.csv`. Do not select successful users or force a realized estimate to equal a known DGP target.

Three block contrasts provide only three independent draws

$$\bar{\delta} = \frac{1}{B} \sum_{b=1}^B \hat{\delta}_b, \quad \widehat{\text{Var}}(\bar{\delta}) = \frac{1}{B(B-1)} \sum_{b=1}^B (\hat{\delta}_b - \bar{\delta})^2, \quad B = 3.$$

- This variance estimate uses independent baseline-block contrasts.
- It is conservative in expectation under the invented design when true block contrasts differ.
- That expectation does not guarantee an accurate realized standard error or a reliable normal approximation with three blocks.
- Panel rows, transactions, and assignment enumerations do not add independent experimental units.

Cameron–Miller (2015) is a general inference resource. Do not bootstrap individual panel rows as if they were independent assignments; the lab reports block SEs without automatic normal confidence intervals.

An interaction statistic does not change the global sharp null

$$H_0 : Y_i(00) = Y_i(10) = Y_i(01) = Y_i(11) \quad \text{for every } i.$$

- Hold realized baseline-adjusted outcomes fixed and reassign the four cells within each original block.
- The primary statistic is the mean bundle contrast; enumerate all 13,824 legal allocations.
- Count absolute statistics at least as extreme as observed, including ties and the actual allocation.
- An interaction statistic evaluated under this same null still tests the global no-effect hypothesis.

What is not being tested

Zero interaction while permitting arbitrary nonzero main effects does not generally impute all missing potential outcomes. This lab does not supply that test.

The oracle audit recomputes outcomes; the Fisher test holds them fixed

Calculation	Operation over legal allocations	Interpretation
Fisher reference	Keep observed outcomes fixed under the stated global sharp null	Exact test of that sharp null for the chosen statistic
Instructor oracle audit	Recompute outcomes from every potential-outcome schedule	Check unbiasedness of eight estimators and their true design variances
Observed block SE	Use three realized block contrasts	Limited estimate of uncertainty, not known oracle variance

Synthetic truth permits a design audit unavailable to an empirical researcher. Neither oracle variance nor a sharp-null p-value is a confidence interval for average complementarity. Significance is not a lab acceptance criterion.

Use documented resources without assuming access to employer records

Resource	Useful starting point	Access or inference boundary
WMS and Census MOPS	Practice questions, measurement, questionnaires, public statistics	Documentation is not unrestricted identified establishment data
Hybrid author repository	Article-linked data/code route	Verify version, files, endpoints, and crossover histories
WFHS, ICPSR 36158 v2	Work-family intervention documentation and public-use route	Verify employer, waves, randomization unit, terms, and file coverage
Activity-data code archive	Published analysis-program pointer	Code does not imply public underlying proprietary records

Bloom–Van Reenen (2007) motivates management measurement. Original article/resource pointers were checked; original microdata and estimation programs were not downloaded or executed for this week.

Reproduce, diagnose, and test the workflow without requiring significance

- **Reproduce:** observed roster and panel → eight contrasts → four-period figure and block uncertainty.
- **Diagnose:** missing mixed cells, bottlenecks, transition value, maintenance, time use, recording, composition, equivalence, and spillovers.
- **Verify:** all 82 supplied assertions; two fresh default runs with identical manifests; seed 88771; Python compilation; overwrite refusal with exit 2 and unchanged files.
- **Discuss:** name one additional outcome that separates two explanations of a favorable productivity statistic.

Guaranteed route: original offline, standard-library Python simulation. Observed, reproduced, instructor, diagnose, and transfer artifacts remain separate; tests concern the invented design, not the validity of a published paper.

Three required papers connect assistance, incentives, and AI rollout

Required reading	Prepare a design contract
Bloom et al. (2013), Does Management Matter?	Separate randomized assistance, adopted routines, firm/plant/record units, implementation, and verified outcomes
Atkin et al. (2017), Organizational Barriers to Technology Adoption	Draw both experiments, their eligible groups, implementation costs, and the payment/demonstration bundle
Brynjolfsson–Li–Raymond (2025), Generative AI at Work	State rollout comparison, outcome denominator, independent quality, heterogeneity, and supported horizon

Focused companions: the management follow-up and Bloom–Han–Liang hybrid trial. Week 9 has 28 scholarly readings plus four resources, not 28 compulsory articles; other items are excerpts and rotating presentations.

The proposal connects reform components to decision-relevant outcomes

Ledger row	What the two-page proposal and one-page ledger must specify
Treatment and counterfactual	Initial eligibility, feasible arms, assigner, implementers, dates, versions, control history
Primary comparison	Component or bundle contrast, population weights, independent assignments, horizon
Verified consequences	Production, independent quality and worker conditions, setup and recurring costs
Exposure and observation	Spillover map, untreated collaborators, exits, closures, survey response, destinations
Decision and mechanism	A result that changes the partner's decision and a result that separates two explanations

Separate observed, elicited, modeled, and unavailable cells. A list of controls cannot replace a comparison; include an explicit attrition plan and spillover plan.

A credible reform evaluation follows implementation, outcomes, and burdens

- Define what was assigned, which components are observed, and which comparisons are missing.
- Measure where the bottleneck moved and who paid for implementation and maintenance.
- Keep output, quality, time, worker conditions, composition, and net value separate.
- Match inference to actual assignments and treatment histories, including spillovers and crossover.

Bridge to Week 10

Defend the reform proposal: what can this organization teach, what would change the decision, and which mechanisms or results could travel to another setting?

The conclusion is a documented organizational comparison, not a universal list of good practices. Week 10 extends feasibility, portability, confidentiality, and the course's research frontier.