

Performance Measurement, Monitoring, and Employer Learning

Organizations, Personnel, and Internal Labor Markets
Week 3

Sai Zhang

A better dashboard number leaves several explanations open

Central question

How do firms learn about ability, effort, and match quality when observed performance is incomplete or strategically produced?

- Recorded idle time falls after a dashboard arrives. Did effort, task assignment, coding, or the observed workforce change?
- Recruitment creates an initial workforce and information set. Work and reviews then change both.
- Follow production, observation, interpretation, feedback, and decisions separately.

Field orientation: Hoffman–Stanton (2025); managerial functions: Benson–Shaw (2025).

Learning goals: validate the measure and bound the claim

- Separate production from its recording, and learning from changing productivity.
- Interpret repeated signals, evaluator differences, and monitoring records under explicit assumptions.
- Identify what a feedback intervention actually changes and who receives information.
- Build a design with two distinct outcome families and a worker-welfare margin.

Core points

Output, effort, rating, wage, ability, and welfare are different objects. Better information can change beliefs; monitoring can also change behavior, reporting, and selection.

Production combines capacity, match, effort, and conditions

$$q_{ijt} = \theta_{it} + \mu_{ij} + a_{it} - d_{jt} + u_{jt} + \epsilon_{ijt}.$$

- i, j, t : worker, job/task, and time; q : quality-adjusted production on a comparable scale.
- θ : productive capacity; μ : worker–job match; a : effort.
- d : task difficulty; u : common operating conditions; ϵ : other shocks.
- A difference in output does not tell us which component differs.

Pedagogical additive decomposition, not an estimated production function. Real inputs may be complementary and lack a common cardinal scale.

Recording is another process with its own rules

$$m_{ijrt}^{(v)} = \lambda_v q_{ijrt} + b_{rv} + \zeta_{ijrt} + \nu_{ijrt}, \quad O_{ijrt} = 1.$$

- m : recorded measure; v : rubric, sensor, software, or coding version; r : evaluator.
- λ_v : production-to-measure scaling; b_{rv} : evaluator calibration.
- ζ : strategic reporting or visibility; ν : recording noise; O : record observed.
- A rating can discretize this index. A move from 2 to 3 need not have the same value as a move from 4 to 5.

Teaching measurement equation. An accurate administrative event need not be a complete economic outcome.

Choose a second outcome that can challenge the first

Recorded object	Competing explanation	Independent check or key fields
Sales/completed units	Prices, returns, customer or task mix	Quality, rework, difficulty, availability
Time/attendance	Recoding, approved leave, support work	Stable taxonomy, output, total eligible time
Supervisor rating	Leniency, coaching, assignment, rubric	Crossed raters, independent audit, dates
Worker self-report	Reporting incentives and nonresponse	Confidential survey, response ledger
Wage/promotion	Pay bands, bargaining, vacancies	Contract and decision-time information

The second measure should have a different failure channel, not just another label from the same source.

Equal task counts can conceal different task contributions

Worker	Easy tasks	Hard tasks	Raw count	Weighted contribution
A	8	2	10	$8 \cdot 1 + 2 \cdot 2 = 12$
B	2	8	10	$2 \cdot 1 + 8 \cdot 2 = 18$

- The weights are hypothetical and require economic justification.
- If a policy changes assignment, controlling for post-policy difficulty can remove part of the policy effect.
- Report the total assignment-inclusive effect separately from a fixed-case-mix comparison.

Constructed teaching example. Neither arbitrary task weights nor rating ranks are cardinal welfare measures.

A signal model begins with restrictions on what is known

$$\theta \sim N(\mu_0, P_0), \quad s_t = \theta + \nu_t, \quad \nu_t \sim N(0, R).$$

- θ : fixed productive type; μ_0, P_0 : prior mean and variance; s_t : signal; R : known noise variance.
- Noise is independent over time and of the type; effort and task conditions have already been accounted for.
- Ask whether scores are copied from prior ratings, whether difficulty is observed, and who sees each signal.
- The assumptions define the exercise; they are not facts implied by the word “signal.”

Normal-signal teaching model, not a claim that managers compute Gaussian posteriors.

The gain weights new information against prior uncertainty

$$\begin{aligned}K_t &= \frac{P_{t-1}}{P_{t-1} + R}, \\ \mu_t &= \mu_{t-1} + K_t(s_t - \mu_{t-1}), \\ P_t &= (1 - K_t)P_{t-1}.\end{aligned}$$

- K_t : updating gain; $s_t - \mu_{t-1}$: surprise; P_t : posterior uncertainty.
- A precise signal receives more weight. A long reliable history makes one noisy surprise less influential.
- Belief revision does not by itself change productive capacity or identify an effort response.

The same signal implies different updates at different precision

$$\mu_0 = 0, \quad P_0 = 1, \quad s_1 = 2.$$

Noise variance	Gain	Posterior mean	Posterior variance
$R = 4$	$1/5$	$0 + (1/5)2 = 0.4$	0.8
$R = 1$	$1/2$	$0 + (1/2)2 = 1$	0.5

What changed?

Signal precision changed. The worker's type did not. A more favorable posterior is not an estimate of newly created productivity.

Hypothetical numerical illustration; both updates follow the stated normal model.

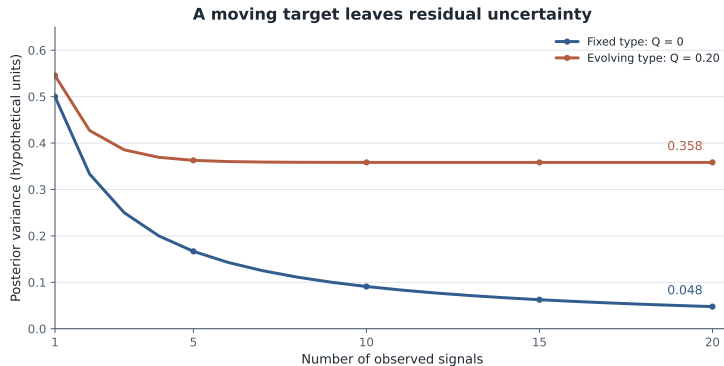
If productivity evolves, information becomes stale

$$\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, Q), \quad P_t^- = P_{t-1} + Q.$$

$$K_t = \frac{P_t^-}{P_t^- + R}, \quad \mu_t = \mu_{t-1} + K_t(s_t - \mu_{t-1}), \quad P_t = (1 - K_t)P_t^-.$$

- Q : process variance; P_t^- : uncertainty before observing the next signal.
- Independent normal process and measurement errors give this simple filtering benchmark.
- With $Q > 0$, the target moves: knowledge, health, motivation, or production conditions can change.
- Changing performance and declining forecast errors need not mean the same thing.

Repeated signals need not eliminate uncertainty



Prior mean 0, variance 1; signal-noise variance $R = 1$; repeated signal = 1.

SYNTHETIC_TEACHING_NOT_PUBLISHED_ESTIMATES — all parameters are hypothetical.

Fixed type: uncertainty falls. Moving type: new process noise replenishes uncertainty.

Synthetic teaching model: $P_0 = 1$, $R = 1$, $Q \in \{0, 0.2\}$, 20 signals. No empirical learning rate is estimated.

Classic learning papers use different information restrictions

Paper	Information/conditioning argument	Interpretation boundary
Farber–Gibbons	Wages and researcher-observed ability correlates over experience	Schooling coefficient can remain constant
Altonji–Pierret	Joint experience interactions: proxy weight rises, schooling weight falls	Conditional restrictions depend on what firms know
Lange	Model-based speed of resolving employer uncertainty	Estimated speed is not an observed manager updating rule

Do not impose a common declining-schooling-coefficient prediction on all three. A researcher-observed proxy need not be unknown to employers.

Farber–Gibbons (1996); Altonji–Pierret (2001); Lange (2007).

Kahn–Lange uses performance information to separate mechanisms

- **Setting:** a twenty-year single-firm panel of performance measures and earnings.
- **Variation and model:** serial information in ratings and pay disciplines a dynamic model of learning and productivity evolution.
- **Measured objects:** repeated performance measures and earnings; productive type remains latent.
- **Limit:** identifying covariances and measurement restrictions matter. Ratings are not directly observed true ability.

Kahn and Lange (2014), *Employer Learning, Productivity, and the Earnings Distribution: Evidence from Performance Measures*.

Draw the information timeline before the wage regression

Work occurs → score entered → signal released
→ employer decision → later validation outcome

- Record which information the worker, supervisor, outside employer, and researcher possessed at each date.
- A later score can validate an earlier expectation retrospectively.
- Using that later score as an earlier prospective decision input leaks future information.
- Wages reflect beliefs together with pay rules, bargaining, and outside options.

Subjective ratings can contain information beyond sales

- **Setting:** Takahashi and coauthors study car-sales personnel records.
- **Comparison:** subjective evaluations alongside contemporaneous and later sales.
- **Object:** information about activities that a narrow count may miss, such as relationships or assistance.
- **Limit:** predictive content does not prove unbiased ability measurement. Ratings can change opportunities, effort, assignment, or retention.

Avoid a false measurement hierarchy

“Objective” means rule-recorded, not complete. “Subjective” means judgment enters, not that the information is useless.

Takahashi et al. (2021 issue; online 2019).

A high-rating supervisor need not be merely lenient

- **Setting:** Frederiksen–Kahn–Lange use personnel records and a model of performance management.
- **Variation:** worker/supervisor outcomes and ratings help distinguish rating generosity from managerial ability and firm information.
- **Measured outcomes:** evaluation scores and personnel outcomes, not a randomized supervisor trait.
- **Limit:** removing a supervisor effect may remove coaching or allocation differences along with generosity.

Validation design

Use common tasks, crossed raters, or independent quality outcomes. A rotation requires an assignment argument and may change more than ratings.

Changing evaluator identity uncertainty changes the process

- **Setting:** local civil servants in China.
- **Variation:** de Janvry and coauthors' field experiment changes uncertainty over who will evaluate the worker.
- **Outcomes:** evaluator-specific influence activities and work behavior under subjective evaluation.
- **Limit:** randomizing an evaluation rule does not independently randomize every behavioral channel proposed to explain the result.

Reading task

Separate the assigned intervention, measured behavior, and interpretation. Which outcome was recorded independently of the evaluator?

de Janvry et al. (2023), *Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior*.

Evaluation can alter how the signal is produced

- Workers can increase productive effort, move toward visible tasks, seek favorable evaluators, or influence ambiguous interpretations.
- Career incentives make performance endogenous even when the current wage is fixed.
- A more precise signal differs from a differently produced signal.
- Record task priorities, visibility, feedback, pay rules, and stable independent outcomes alongside evaluations.

Holmström (1999), dynamic career concerns. Week 4 develops explicit rewards and the desirability of the resulting effort allocation.

Correlated reviewers add less independent information

$$\text{Var}(\bar{\nu}) = \sigma^2 \left[\rho + \frac{1 - \rho}{R_a} \right].$$

- R_a : number of equal-variance raters; σ^2 : individual error variance; ρ : common pairwise error correlation.
- Four independent unit-variance raters: $\text{Var}(\bar{\nu}) = 1/4 = 0.25$.
- Four raters with $\rho = 0.75$: $0.75 + 0.25/4 = 0.8125$.
- A shared systematic bias is not averaged away. Agreement alone does not establish construct validity.

Covariance identity under the stated assumptions. These numbers are hypothetical, not effects of a multirater reform.

Evaluation experiments address different environments

	Ockenfels–Sliwka–Werner	Kusterer–Sliwka
Setting/design	Model plus controlled laboratory evidence	Online experiment
Variation	Multirater evaluation and incentives	Evaluation incentives and information
Object	Aggregation, precision, and incentives	Rating informativeness and social preferences
Limit	More reviewers do not remove common bias	Measured preferences are not randomly assigned traits

Neither is a firm field RCT. Ask what independent information a reviewer adds and what consequences the reviewer anticipates.

Ockenfels, Sliwka, and Werner (2025); Kusterer and Sliwka (2025 issue; online 2024).

Monitoring changes observation and may change behavior

$$\Pr(\text{recorded incident}) = \Pr(\text{latent incident}) \\ \times \Pr(\text{detected} \mid \text{latent incident}).$$

- This teaching identity assumes no false positives; otherwise include their contribution.
- Monitoring can change detection, effort, salience, trust, discretion, or coordination.
- A single treatment indicator identifies the component or bundle actually varied.
- A stable independent audit can challenge an interpretation based solely on detection-dependent records.

Recorded incidents can rise while underlying incidence falls

Hypothetical regime	Latent incidence	Detection	Recorded incidence
Before	0.10	0.25	$0.10 \cdot 0.25 = 0.025$
After	0.06	0.75	$0.06 \cdot 0.75 = 0.045$

- An apparently worse administrative number is consistent with improved behavior.
- A regression on the same detection-dependent outcome cannot by itself reveal the latent incident rate.
- Propose an audit or separate variation in detection, and state its coverage and error process.

Constructed numerical example, with no false positives.

Restaurant monitoring adoption requires a comparison argument

- **Setting:** Pierce–Snow–McAfee study monitoring technology in restaurants.
- **Variation:** observational adoption over time, not randomized deployment.
- **Outcomes:** recorded theft and sales, which can respond through different channels.
- **Limit:** adoption timing, store trends, spillovers, and changing observation rules affect interpretation.

Methods translation

Specify treatment cohorts and eligible untreated comparisons. A modern group-time estimator does not make endogenous adoption exogenous.

Pierce, Snow, and McAfee (2015); Callaway and Sant'Anna (2021).

Privacy and control are distinct mechanisms and evidence types

	Bernstein	Falk–Kosfeld
Setting	Factory field study and privacy experiment	Laboratory principal-agent game
Variation	Visibility/privacy	Minimum-performance restriction and control
Objects	Adaptation, concealment, and organizational learning	Behavioral response to being controlled
Limit	No universal harm from monitoring	Not an electronic-surveillance field effect

State which component is changed before moving a mechanism to another workplace.

Bernstein (2012), *The Transparency Paradox*; Falk and Kosfeld (2006), *The Hidden Costs of Control*.

Teacher observations link monitoring to service outcomes

- **Setting:** Phipps studies unannounced teacher observations.
- **Variation:** exogenous observation variation; reconstruct the institutional assignment rule before characterizing it more precisely.
- **Outcomes:** pedagogy and student outcomes provide distinct measurement margins.
- **Limit:** do not infer an RCT from the abstract. Original microdata are proprietary; a code/data-description archive is a different access object.

Phipps (2026), *Does Monitoring Change Teacher Pedagogy and Student Outcomes?*

Bernstein–Li changes who can directly see existing data

- **Setting:** frontline mechanics and supervisors.
- **Variation:** field experiment assigning four of eleven work centers access to own and peers' performance data.
- **Information institution:** workers gain direct access; supervisors remain informed.
- **Outcomes:** productive, support, and nonproductive time.
- **Limit:** reduced nonproductive time did not establish an overall productive-time increase or a welfare gain.

Bernstein and Li (2026 issue; online 2025), study of direct performance-data access and supervisor feedback intermediation.

Data access, activities, and valuable output are separate targets

- ① **Knowledge:** did workers understand their own performance or expectations better?
- ② **Activities:** did recorded productive, support, or nonproductive time change?
- ③ **Value:** did completed work, quality, rework, or worker welfare improve?

Treatment-component boundary

Bernstein–Li supplies own and peer information together. It does not independently identify self-monitoring, peer comparison, or every managerial response.

A supplement listing is not evidence that original microdata and redistribution rights have been inspected. Our lab uses synthetic records.

Relative input feedback is different from an output rank

	Blanes i Vidal–Nossol	Rilke et al.
Setting/design	Personnel records; quasi-experimental relative-rank information	Sales field experiment
Information change	Relative-performance feedback	Relative input feedback on top of own-input information
Objects	Performance in an existing piece-rate environment	Effort inputs, sales outputs, and initially low performers
Limit	No new prize does not mean no monetary incentives	More calls/licenses does not establish higher conversion efficiency

Blanes i Vidal and Nossol (2011); Rilke et al. (2025).

A factorial proposal can separate information components

- Independently vary access to one's own history and access to peers' histories.
- Add an unchanged blinded quality measure that can disagree with recorded activity.
- A further supervisor-feedback arm can test interpretation beyond raw access.
- Choose assignment units that address peers sharing information with controls.
- Measure privacy concerns, pressure, and understanding separately from output.

Proposed research design, not the full treatment structure of any cited study.

Preserve measurement histories and information dates

- Build worker $\times$ task/job $\times$ evaluator $\times$ period records when the question requires them.
- Retain raw and standardized scores, rubric/software versions, assignment histories, and edits after decisions.
- Separate work date, recording date, release date, and decision date.
- Record who collected each signal and who could see it at each date.
- Keep the baseline roster when reviews disappear; exit, leave, exemption, and system failure are distinct missingness states.

Access routes determine what a student project can observe

Resource	Useful object	Access boundary
BLS NLS/NLSY79	Wage, employment, and background histories	Public-use and restricted tiers differ
IAB Linked Personnel Panel	Employer practices and worker job quality	Product-specific controlled access
Partner employer	Tasks, ratings, output, pay, exits	Specific agreement; no partnership assumed
de Janvry archive	Optional evaluation-study reproduction	Listing checked; actual files not executed here
Bernstein–Li supplement	Possible materials for data-access study	Listing does not establish original data availability
Phipps archive	Code and data description	Original microdata proprietary

Consult the chapter's resource ledger. The guaranteed lab requires no empirical microdata or access application.

Independent outcome families can disagree usefully

Headline measure	Challenging outcome	Worker-welfare margin
Completed tasks	Blinded quality and rework	Work intensity and predictable hours
Recorded time use	Output with unchanged taxonomy	Autonomy and pressure
Supervisor rating	External audit or customer outcome	Perceived fairness and clarity
Employer profit	Costs and quality outside the target	Pay, privacy, willingness to remain

A supervisor score and the same supervisor's narrative sentiment share a source. Two columns do not necessarily provide independent validation.

A worker average can contain the very noise it predicts

$$y_{it} = \theta_i + e_{it}, \quad \hat{\theta}_i^{\text{all}} = \frac{1}{T} \sum_{s=1}^T y_{is}.$$

$$\text{Cov}(\hat{\theta}_i^{\text{all}}, e_{it}) = \frac{\sigma_e^2}{T}.$$

- Assume independent mean-zero transitory errors, independent of θ_i ; T : observations per worker; σ_e^2 : error variance.
- The focal observation enters both the score and the outcome. Part of the predictive relationship is mechanical.
- Leaving out that record removes this covariance under the stated independent-error assumptions.

Teaching worker-average example; not a general fixed-effect bias formula.

Leave-out, cross-fitting, and variance correction differ

- A leave-out worker average addresses own-record overlap under its error assumptions.
- Cross-fitting separates estimation from evaluation; splits must respect worker, rater, team, and time dependence.
- Kline–Saggio–Sølvsten correct noisy fixed-effect variance components: the target and procedure differ from deleting one record from an average.
- None of these procedures resolves endogenous assignment, serial dependence by assumption alone, or unavailable prospective information.

Kline, Saggio, and Sølvsten (2020), *Leave-Out Estimation of Variance Components*. State the target before choosing the correction.

Follow the baseline workforce when monitoring changes selection

- Report retention and observation probabilities alongside performance.
- Ratings among survivors describe a selected population.
- Zero firm contribution after exit can be part of a fixed-employer accounting target; missing potential quality in the job is not observed zero.
- Lee-style bounds require the relevant assignment and monotone-selection assumptions. Balanced average attrition does not establish monotonicity.

Lee (2009). Carry Week 2's selection discipline into post-entry performance measurement.

With few centers, define the estimand before the test

$$\hat{\tau}_C = \frac{1}{4} \sum_{c:Z_c=1} \Delta \bar{y}_c - \frac{1}{7} \sum_{c:Z_c=0} \Delta \bar{y}_c.$$

- **Teaching design:** exactly four of eleven centers assigned treatment uniformly.
- Z_c : center assignment; $\Delta \bar{y}_c$: center mean outcome change.
- Each center gets equal weight within its arm. Unequal center sizes make this differ from a worker-weighted target.
- Many worker-days do not create more independently assigned centers.

Synthetic design motivated by Bernstein–Li. The complete-randomization protocol here is an explicit teaching assumption.

Exact assignment inference has a specific null and design

$$\binom{11}{4} = \frac{11!}{4!7!} = 330.$$

- Under the sharp null of no effect on any unit, recompute the center statistic for all 330 allowed assignments.
- Compare the observed statistic with that exact randomization distribution using the prespecified two-sided rule.
- This is not a generic test of a weak average-effect null.
- Do not permute worker-days. Bernstein–Li’s original allocation faced employee-count constraints; its assignment support cannot be inferred from four of eleven alone.

Uniform four-of-eleven complete randomization belongs to the stated synthetic lab design.

Research Lab: reconstruct the design with synthetic records

- **Mode A:** mandatory offline synthetic design reconstruction motivated by Bernstein–Li.
- Eleven centers, four treated, unequal center sizes, 76 workers, and 1,216 repeated records.
- Every worker, effect size, error process, and outcome is invented. The exhaustive 480-minute budget is a teaching simplification.
- Estimate center changes in nonproductive, productive, and support time plus independently recorded quality.
- Enumerate center assignments for the prespecified primary endpoint: nonproductive minutes. Secondary outcome tests are exploratory and unadjusted.

No original microdata were downloaded or executed. Optional empirical Mode B requires a separately verified release and reproduction target.

Diagnose behavior, recording, and model-based learning

- ① **Time allocation:** the synthetic intervention reallocates time without a direct productive-time or quality effect in the DGP.
- ② **Recording rule:** a separate score change leaves a blinded audit unchanged.
- ③ **Learning model:** uncertainty changes under explicit fixed/drifting-state assumptions.
- ④ **Measurement checks:** compare correlated raters and demonstrate own-record leakage.

Interpretation task

Which result comes from randomization, which from an accounting identity, and which from model assumptions? A nonsignificant quality estimate does not establish equivalence.

Methods Box: connect the construct to the comparison

- ① Define the construct, unit, denominator, and baseline population.
- ② Freeze measurement versions and preserve raw records.
- ③ State information sets and separate treatment components.
- ④ Justify the comparison: random access, crossed raters, observation rules, or adoption timing.
- ⑤ Align inference with dependence and assignment; retain missingness reasons.
- ⑥ Validate an independently recorded outcome and state worker costs.

A larger file, supervisor fixed effects, or a more flexible prediction method cannot replace those arguments.

Frontier proposals begin with a measurement distinction

Proposed question	Data or variation to add
Sensor/rubric changes	Parallel old/new measures and an independent audit
Who sees information?	Separately vary precision, recipients, and interpretation
Learning about a moving type	Repeated performance, tasks, and pay with explicit dynamic restrictions
Worker welfare	Quality plus confidential pressure, autonomy, and privacy measures
Correlated reviewers	Common tasks, independently assigned assessments, visibility rules

These are proposed directions, not declarations of empty literatures. Each needs a comparison and a result that could overturn the initial interpretation.

Three required readings address three different problems

Required reading	Student output
Kahn–Lange (2014)	Information timeline; observed moments versus latent learning/productivity components
de Janvry et al. (2023)	Treatment-component ledger; influence behavior and one alternative mechanism
Bernstein–Li (2026 issue; online 2025)	Outcome map; own/peer data access; productive, support, and nonproductive time

Rotating presentations: choose one evaluator paper and one monitoring/feedback paper. Use classic learning and methods readings for the project's particular restriction, not as 24 compulsory readings.

Transfer: write a measurement-validation plan

Two-page proposal

Baseline population; assignment or quasi-exogenous variation; who sees which information; primary measure; independent outcome family; worker-welfare outcome; alternative explanation; falsifying result.

- Attach a data dictionary with units, denominators, timing, versions, and access conditions.
- State what a changed score could mean besides changed production.
- Explain what remains unobserved after exit or selective review.
- Keep modeled counterfactuals and empirically identified contrasts separate.

Week 4 asks what the firm should reward

- Information can alter beliefs, behavior, recording, and selection through different channels.
- A stronger reward on a number need not increase the value of the underlying work.
- Baker–Gibbons–Murphy connects incomplete objective measures with subjective information and contractual incentives.
- Duflo–Hanna–Ryan combines cameras with attendance-linked pay: the treatment identifies a bundle, not monitoring alone.

Bridge forward

Compensation, risk sharing, multitasking, and relational contracts build on the measurement and information distinctions established here.

Baker, Gibbons, and Murphy (1994); Duflo, Hanna, and Ryan (2012). The latter's journal erratum must be checked before reusing original structural calculations; none are used here.