

The Optimal Intensity of Performance Evaluation

Zhongjun Wang

zhong.junwang@126.com

University of Science and Technology Liaoning <https://orcid.org/0000-0002-7357-6313>

Research Article

Keywords: Goodhart's Law, performance evaluation, threshold model, goal displacement, qualitative judgment, Fokker–Planck equation

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11206939/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

The Optimal Intensity of Performance Evaluation: A Threshold Model of Incentive Alignment and Goal Displacement in Higher Education

Abstract

Performance evaluation has become a central governance instrument in universities worldwide, yet how much evaluation is too much remains unanswered. This paper converts Goodhart's and Campbell's Laws from rhetorical warnings into a parameterised threshold model: under diminishing marginal returns, evaluation intensity and organisational effectiveness follow an inverted-U, but beyond a critical threshold ($k < k_0(w)$) it triggers escalating dysfunction. It makes three contributions to higher-education research. First, although Goodhart and Campbell have guided the field for decades, they have remained interpretive metaphors; to our knowledge this is the first within higher-education scholarship to render the counter-productive point measurable, building on economics (Rablen, 2010) and reinforcement learning (Skalse et al., 2022) while adding an institutional threshold and mitigation mechanism those accounts lack. Second, it decomposes qualitative review into three non-interchangeable levers—peer review, value orientation, procedural justice—that respectively shift the threshold, the benefit ceiling, and the baseline dysfunction cost. Third, it adds a drift–diffusion layer showing that above-threshold evaluation pushes faculty effort toward metric optimisation, and distils the argument into six falsifiable propositions with identification strategies. Because the thresholds and levers are parameterised, the same diagnostic travels across national regimes, from China's Break Five Onlys to cross-national research assessment.

Keywords: Goodhart's Law; performance evaluation; threshold model; goal displacement; qualitative judgment; Fokker–Planck equation

1. Introduction

Performance evaluation has become a central governance instrument in contemporary higher education. Since the 1980s, neoliberal restructuring, fiscal austerity, and global competition have pushed governments to treat universities as instruments of economic competitiveness, demanding measurable accountability; global rankings (ARWU, QS, THE) have amplified this metric pressure, especially on peripheral and semi-peripheral systems. Universities now tie publications, citations, and grants to promotion, tenure, and resources, and the marginal cost of measuring further dimensions has fallen

sharply—encouraging expansion of evaluation intensity almost regardless of whether marginal benefits remain commensurate.

These developments expose a fundamental tension: evaluation is meant to improve effectiveness, yet pushed beyond a point it can undermine the very goals it advances. The tension is widely framed through Goodhart’s Law (1975)—a measure that becomes a target ceases to be a good measure—and Campbell’s Law (1979), and a large higher-education literature uses them to interpret how tightly coupled indicators induce gaming and goal displacement (Anderson et al., 2022; Murphy, 2009). Yet Goodhart originated in monetary economics and was imported as an interpretive lens rather than developed natively; qualitative studies document metric overload, but formal models of the non-monotonic relationship between evaluation intensity and outcomes—especially ones combining threshold effects with endogenous mitigation—remain scarce. Without such a model, reformers across China’s Break Five Onlys, the UK REF, and continental Europe cannot say when evaluation helps or harms. This paper addresses that question directly.

The argument integrates organisational economics, motivation psychology, and higher-education policy analysis. This cross-disciplinary orientation is deliberate: formal modelling is distinctive in a field where Goodhart’s Law has been cited rhetorically rather than developed as a theoretical engine. The contribution is both substantive and methodological—translating qualitative claims into testable structure within a subfield where formal modelling of evaluation effects lags behind empirical documentation.

This paper develops a formal threshold model in which evaluation intensity has an optimum: below k_0 , evaluation improves coordination, while beyond k_0 , dysfunctional costs grow disproportionately. To higher-education scholarship it makes three compact additions. (i) It turns Goodhart’s (1975) and Campbell’s (1979) Laws into a parameterised, falsifiable threshold proposition rather than an interpretive metaphor, going beyond Rablen (2010), who derives an inverted-U from an individual reference point without an institutional threshold, and Skalse et al. (2022), whose reward-hacking result decays monotonically with no threshold or mitigation. (ii) It splits qualitative review into three non-interchangeable levers—peer review raising k_0 , value orientation raising A , procedural justice reducing B —replacing the undifferentiated “quantitative versus qualitative” debate (Power, 1997; Wilsdon et al., 2015; de Rijcke et al., 2016) with lever-specific, testable design. (iii) It adds a Fokker–Planck drift–diffusion layer and six falsifiable propositions, extending Holmström and Milgrom’s (1991) smooth, static substitution to a social-norm tipping point and identification strategies for public panel data. Together these yield an operational design constraint $k < k_0(w)$ for evaluation-system designers.

Section 2 reviews the literature; Section 3 develops three formulations; Section 4 derives six propositions; Section 5 discusses policy; Section 6 outlines limitations; Section 7 concludes.

2. Theoretical foundations and literature review

2.1 Goodhart’s Law, Campbell’s Law, and the metrics critique

Among the ways a measure ceases to be a good measure once it becomes a target, this paper focuses on the adversarial case: agents strategically optimise the metric at the expense of the true objective, with the threshold k_0 marking where gaming becomes dominant. This is well documented in academia—researchers inflate self-citations when metrics drive careers (Seeber et al., 2019). Skalse et al. (2022) formalise analogous reward hacking in reinforcement learning, showing reward–objective correlation decays monotonically as a reward model is optimised; parallel to our goal displacement, but our formulation adds an explicit threshold and a mitigation mechanism absent from that literature. Campbell’s (1979) Law likewise notes that quantitative social indicators used for decision-making become corruption-prone.

In higher education, these principles manifest when publications, citations, and grants drive promotion and allocation (Hyle, 1999; Morgan Jones et al., 2022; Anderson et al., 2022). Kerr’s (1975) warning about rewarding A while hoping for B foreshadows Goodhart, and the Matthew effect (Merton, 1968) plus the reactivity of public measures (Espeland & Sauder, 2007; Sauder & Espeland, 2009) amplify the distortion. Broader critiques frame this as an audit explosion (Power, 1997; Shore & Wright, 2000), performativity (Ball, 2003), and the Leiden Manifesto’s metric caution (Hicks et al., 2015; Aksnes et al., 2019). Most applications remain qualitative; formal treatments exist in adjacent fields—reference-dependent targets (Rablen, 2010) and performance pay crowding out motivation (Siciliani, 2009)—and Section 3.2 distinguishes our concave-utility foundation.

2.2 Motivation theory and the crowding-out effect

Self-determination theory distinguishes intrinsic from extrinsic motivation, identifying autonomy, competence, and relatedness as basic needs (Deci & Ryan, 1985): evaluation supporting these needs enhances intrinsic motivation, while evaluation thwarting them drives burnout (Sabagh et al., 2022). A meta-analysis of 128 experiments confirms controlling rewards crowd out intrinsic motivation (Deci, Koestner, & Ryan, 1999), especially damaging in academia (Siciliani, 2009). Crowding-out thus supplies the psychological micro-foundation for the dysfunction cost $U_2(k)$: evaluation perceived as controlling generates disproportionately large costs above a threshold (Frey & Jegen, 2001). Procedural-

justice research (O’Connell et al., 2021) further shows that perceived fairness moderates acceptance of metric systems, motivating our treatment of w_j as a distinct sub-dimension.

2.3 Multitask principal-agent theory and Chinese evaluation reform

Holmström and Milgrom (1991) show that when some tasks are unmeasurable, optimal contracts weaken incentives on measurable tasks to prevent effort distortion—but their trade-off is smooth, with no critical discontinuity, whereas our model adds an explicit threshold and a dynamic drift mechanism (Section 3.6). Cross-national variation is substantial: the UK REF and Italy’s ANVUR mix bibliometrics with expert assessment (de Rijcke et al., 2016; Ortagus et al., 2020), while US performance-based funding has produced disappointing completion effects and has been abandoned by several states (Rutherford & Rabovsky, 2014; Hillman et al., 2014; Li & Hu, 2025). Volume-heavy systems with little qualitative review show declining returns, whereas mixed systems report more sustainable outcomes—UK REF data show diminishing returns to publication volume, consistent with an inverted-U (Hammarfelt & de Rijcke, 2015). These observations motivate but do not substitute for formal identification.

China’s evaluation system is a particularly salient context. Since the early 2000s, publications, citations, and grants have been tightly coupled to promotion, producing widespread “index worship” (Chen, 2021; Yan, 2019b). In 2020 the Ministry of Education issued the flagship “Break Five Onlys” (po wu wei) reform, targeting over-reliance on papers, titles, degrees, salaries, and hats (Liu, Z., 2020; Jiang & Wang, 2021; Meng & Zhang, 2021). Implementation has been uneven, with symbolic adoption and capacity constraints as universities struggle to operationalise qualitative evaluation (Meng & Zhang, 2021; Chen, 2021). This motivates a model explaining both why intense quantitative evaluation produces dysfunctions and why qualitative reform is necessary but insufficient without capacity.

International variation maps directly onto the model’s parameters: the sustainable threshold is not a universal constant but a function of institutional design. Performance-based funding spans a spectrum from mechanical publication counting, through expert-weighted models, to performance agreements (Hicks, 2012). Australia’s undifferentiated publication counting was followed by declining citation impact (Schneider et al., 2016), whereas Norway’s tiered indicator preserved impact while raising output (Aagaard et al., 2015; Bloch & Schneider, 2016); Italy’s ANVUR claims informed peer review but relies heavily on bibliometrics in STEM (D’Angelo, 2025), and the UK REF and Hong Kong RAE sit toward the high-threshold end (Li, Lo, & Yang, 2024). These patterns are consistent with review infrastructure, not intensity alone, setting the sustainable threshold.

2.4 Positioning the threshold model within the higher-education literature

Higher-education scholarship offers three well-developed accounts of why quantitative evaluation can go wrong. The first is the normative audit critique (Power, 1997; Shore & Wright, 2000; Ball, 2003; Hicks et al., 2015; Wilsdon et al., 2015): it shows that measurement distorts academic purposes, but not where quantitatively the distortion begins. The second is reactivity and commensuration (Espeland & Sauder, 2007, 2009), powerful narratively but without comparative-static predictions about when effects turn negative. The third is organisational economics—multitask agency (Holmström & Milgrom, 1991) and motivation crowding-out (Deci, Koestner, & Ryan, 1999; Frey & Jegen, 2001)—which supplies an incentive logic but predicts smooth, continuous substitution rather than a discrete tipping point.

The threshold model's contribution is to convert the first two traditions' qualitative insights into one falsifiable structure: an explicit institutional threshold, a three-lever decomposition of qualitative review, and six estimable propositions. It departs from Skalse et al. (2022), whose reward-hacking result decays monotonically with no threshold and no moderating lever, and it replaces the third tradition's smooth trade-off with a discontinuous one—gradual below the threshold and sharply accelerated above it by a social-norm tipping point (Granovetter, 1978). It is best read not as an economics import but as a formalisation of claims the field has long made qualitatively—now measurable and testable.

3. Model development

3.1 Definition and operationalisation of evaluation intensity k . A complete notation table is provided in Table S1 of the Online Supplementary Materials.

Evaluation intensity k is the degree to which an organisation relies on quantitative metrics to monitor, evaluate, and reward faculty, operationalised as a composite of four dimensions: metric weight in personnel decisions; measurement granularity; consequence magnitude; and standardisation across disciplines. Each is scored 0–1, and k is their unweighted mean. Equal weighting is a transparent baseline, and the qualitative predictions depend only on the signs of the partial derivatives, not on the weighting scheme; each dimension is codable from promotion documents and policy records (cf. Ortagus et al., 2020).

Organisational effectiveness U is defined as true long-term organisational value creation—high-impact research, quality teaching, and service—distinct from short-term metric-based scores. This definition is itself contested, since what counts as valuable varies across missions and stakeholders. The model treats U from the standpoint of the organisation's stated educational and scholarly mission. As Section 5.4

discusses, the qualitative prediction—that excessive metric intensity undermines whatever U the organisation values—is robust across competing definitions, although the optimal intensity k^{**} and the threshold k_0 vary quantitatively with the choice of U .

3.2 Static univariate model: functional form and micro-foundations

The net effectiveness $U(k)$ is decomposed into an alignment benefit $U_1(k)$ and a dysfunction cost $U_2(k)$. Equation (1) specifies the alignment benefit as a concave exponential approaching an asymptote A , where A is the maximum benefit and α governs the diminishing-returns rate. The concavity reflects diminishing marginal returns to evaluation: the first units clarify expectations and provide basic feedback, while later units yield smaller marginal gains. This rests on standard concave-utility assumptions rather than behavioural reference-dependent preferences. As the comparison with Rablen (2010) below details, the inverted-U does not require prospect theory, loss aversion, or reference points; non-monotonicity arises from the cost side, not from kinked utility.

$$U_1(k) = A(1 - e^{-\alpha k}) \quad (1)$$

Equation (2) gives the piecewise dysfunction cost: below the threshold k_0 , it grows linearly at rate B ; above k_0 , an exponential term is added. The exponential form, which applies for $k > k_0$, is justified by two reinforcing feedback mechanisms. First, a social-norm effect: once a critical mass of faculty game the metric, academic norms shift, raising the social cost of compliance and the relative return to gaming (Espeland & Stevens, 1998).

Second, an organisational-amplification effect: management observing short-term metric growth intensifies quantitative evaluation, creating a feedback loop (Muller, 2018). Let $p(k)$ denote the proportion of gaming faculty. Above k_0 , social-norm feedback satisfies a feedback law (Eq. 28, Online Supplementary Materials): the gaming proportion grows exponentially at feedback rate r —the social-norm feedback strength—calibrated to track the dysfunction-acceleration parameter β . Because dysfunction cost is proportional to $p(k)$, this yields the exponential term directly. The qualitative prediction—that marginal dysfunction rises with k —is robust across exponential, quadratic, and S-shaped costs; only the precise location of k^{**} and k_{be} is function-form-dependent (Table S4).

$$U_2(k) = \begin{cases} Bk, & k \leq k_0 \\ Bk + C(e^{\beta(k-k_0)} - 1), & k > k_0 \end{cases} \quad (2)$$

$$U(k) = \begin{cases} A(1 - e^{-\alpha k}) - Bk, & 0 \leq k \leq k_0 \\ A(1 - e^{-\alpha k}) - Bk - C(e^{\beta(k-k_0)} - 1), & k > k_0 \end{cases} \quad (3)$$

A comparison with Rablen (2010) clarifies the role of preferences. Rablen derives non-monotonic effort under prospect-theoretic reference dependence—loss aversion and diminishing sensitivity relative to a reference point—whereas we obtain the inverted-U from concave benefits plus convex piecewise-exponential costs under standard expected utility, with no reference point. They agree on an inverted-U but differ in three ways: Rablen’s effort declines when a target crosses an individual reference point, whereas ours declines when evaluation intensity crosses the organisational threshold k_0 —an institutional rather than psychological boundary; Rablen’s decline is smooth with no acceleration or break-even point; and ours needs no loss aversion. We therefore position the model as a complement, distinguishable empirically by whether dysfunction accelerates sharply at an institutional threshold (ours) or smoothly at an individual reference point (Rablen).

Summing gives Equation (3). For $k \leq k_0$, the first-segment marginal condition (Eq. 4, Online Supplementary Materials) is positive when $A\alpha > B$, with interior optimum k^* (Eq. 5, Online Supplementary Materials); k^* is the unconstrained first-segment optimum, and the global optimum lies there iff $k^* \leq k_0$ and $U'(k_0) \leq 0$. Under the baseline parameters $k^* = 1.498 > k_0 = 0.5$, the global optimum $k^{**} = 0.6063$ lies in the second segment. For $k > k_0$, the second-segment marginal condition (Eq. 6, Online Supplementary Materials) gives k^{**} solving Equation (7) numerically. As $k \rightarrow \infty$, the exponential term dominates and $U(k) \rightarrow -\infty$, so sufficiently excessive evaluation leaves the organisation worse off than none at all. The break-even point k_{be} (Equation 8) solves the zero-net-utility condition; under baseline parameters $A = 1.0, \alpha = 2.0, B = 0.1, C = 0.2, \beta = 2.0, k_0 = 0.5$, numerical solution yields $k^{**} \approx 0.606$ and $k_{be} \approx 1.303$, marked in Figure 1 (k^{**} and k_{be}).

$$A\alpha e^{-\alpha k^{**}} = B + C\beta e^{\beta(k^{**}-k_0)} \quad (7)$$

$$A(1 - e^{-\alpha k_{be}}) = Bk_{be} + C(e^{\beta(k_{be}-k_0)} - 1) \quad (8)$$

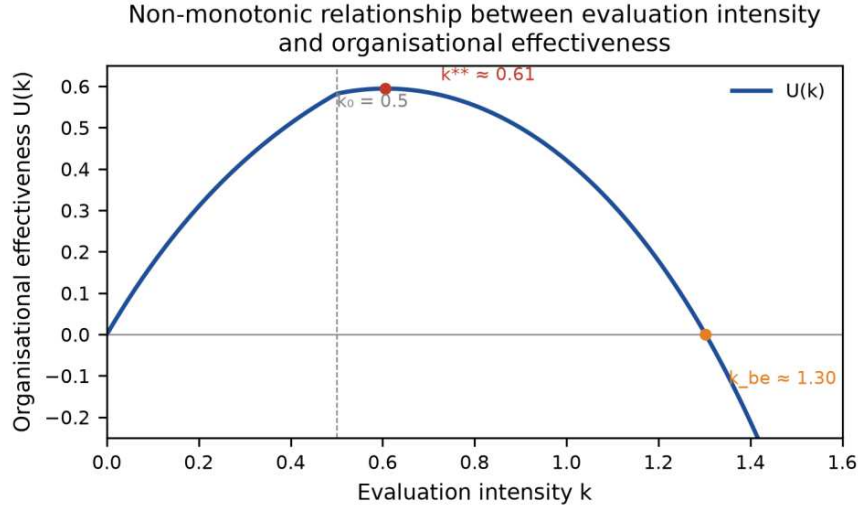


Figure 1. Non-monotonic relationship between evaluation intensity k and organisational effectiveness $U(k)$. The solid curve shows $U(k) = U_1(k) - U_2(k)$, marking the critical threshold $k_0 = 0.5$, optimal intensity $k^{**} \approx 0.61$, and break-even point $k_{be} \approx 1.30$. Parameters: $A = 1.0$, $\alpha = 2.0$, $B = 0.1$, $C = 0.2$, $\beta = 2.0$, $k_0 = 0.5$. Parameter values are illustrative; see Section 3.5 for empirical anchoring. Reproducibility code and high-resolution figures are available in the Online Supplementary Materials.

3.3 Bivariate extension: qualitative judgment as a three-dimensional mitigation mechanism

We introduce $w \in [0,1]$ as the weight of qualitative judgment. The threshold-raising function is Equation (9) and the benefit-enhancement function Equation (10), both linear in the qualitative-judgment weight. Equation (11) gives the bivariate utility; the marginal conditions and the decomposition of this weight into peer review w_p , value orientation w_v , and procedural justice w_j , together with their extended specifications, are derived in the Online Supplementary Materials (Eqs. 12–15). We propose these mechanism assignments as testable hypotheses. Peer review primarily raises k_0 because expert assessment reduces information asymmetry: reviewers can distinguish genuine high-impact work from strategic publication, blunting metric optimisation and pushing the threshold rightward. Value orientation primarily raises A because

alignment of metrics with institutional mission recognises hard-to-quantify outputs (teaching quality, service, interdisciplinary work), raising the maximum achievable alignment benefit. Procedural justice primarily reduces B because transparent criteria, faculty representation, and formal appeal mechanisms mitigate resentment and resistance, lowering baseline administrative friction and disengagement (O’Connell, O’Siochru, & Rao, 2021).

$$k_0(w) = k_{00}(1 + \gamma w) \quad (9)$$

$$A(w) = A_0 + \delta w \quad (10)$$

$$U(k, w) = A(w)(1 - e^{-\alpha k}) - \begin{cases} Bk, & k \leq k_0(w) \\ Bk + C(e^{\beta(k-k_0(w))} - 1), & k > k_0(w) \end{cases} \quad (11)$$

Table S3 summarises the three sub-dimensions and their mechanisms. These identify primary channels, not exclusive effects: procedural justice may also raise k_0 , peer review may also raise A , and value orientation may reduce B . If a mechanism is misassigned—peer review operating mainly through A rather than k_0 —the lever-to-parameter predictions are revised, but the inverted-U and threshold existence remain intact because they depend on the sign of the partial derivatives. Proposition 4 tests dominant against secondary effects. Figure 2 plots $k_0(w)$ and $A(w)$ on separate axes; the levers are not interchangeable, and misassignment is correctable through Proposition 4’s interaction tests.

quantities have different units and are not directly comparable. The decomposition transcends the undifferentiated “quantitative versus qualitative” dichotomy: rather than asking broadly how much qualitative judgment to add, policymakers target specific levers—raising the threshold via peer review, lifting the ceiling via mission alignment, or lowering the floor via procedural justice—each with distinct cost profiles and institutional prerequisites, and the model shows they are not interchangeable. Misassignment is correctable through Proposition 4’s interaction tests.

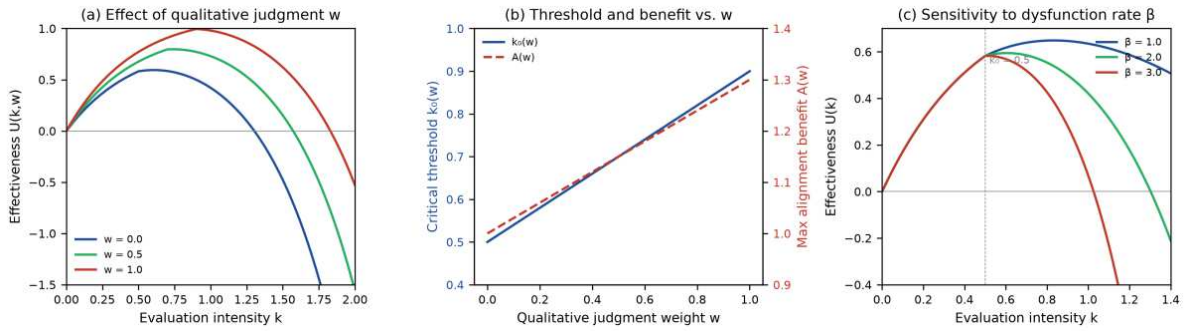


Figure 2. Effect of qualitative-judgment weight w and dysfunction rate β . (a) $U(k, w)$ for $w = 0, 0.5, 1.0$; higher w shifts the threshold rightward. (b) $k_0(w)$ (left axis, evaluation intensity) and $A(w)$ (right axis, maximum alignment benefit) as functions of w ; the two quantities have different units and are plotted on separate axes, not for direct comparison. (c) Sensitivity: $U(k)$ for $\beta = 1.0, 2.0, 3.0$, confirming a robust inverted-U. Parameters: $A = 1.0, \alpha = 2.0, B = 0.1, C = 0.2, k_{00} = 0.5, \gamma = 0.8, \delta = 0.3$. Parameter values are illustrative.

3.4 Dynamic model: Fokker–Planck drift–diffusion and the threshold

It is worth stating explicitly what this dynamic layer adds. The inverted-U already establishes that net effectiveness eventually declines; what it does not capture is how metric scores and true value drift apart over time. The drift–diffusion layer answers that process question: it predicts that above the threshold,

even as publication counts continue to rise, forward-looking true value declines monotonically (Proposition 5). This divergence is the layer's genuinely new dynamic prediction, and the equations that follow are its derivation rather than an additional claim; full derivations are in the Online Supplementary Materials. The Heaviside step used below is a parsimonious device, and replacing it with a smooth logistic transition leaves all qualitative predictions unchanged, as the robustness check below details.

Equation (16) is the Fokker–Planck equation. Let $e(x, t)$ be the density of faculty effort at quality $x \in [0,1]$, where $x=0$ is pure metric optimisation and $x=1$ pure true-value production. Equation (17) gives the drift velocity $v(x, k, w) = -\mu k \cdot x(1 - x) \cdot \Theta(k - k_0(w))$, with $\mu > 0$ drift strength, D diffusion, and Θ the Heaviside step; reflecting boundary conditions (Equations 18–19) and initial condition (Equation 20) close the system. The step encodes discontinuous norm change: below k_0 , gaming is stigmatised and low-return; above k_0 , it becomes normalised and high-reward—a counterpart to policy reforms that abruptly raise stakes and Granovetter-style tipping points.

behaviour can produce sudden norm changes once a critical mass of faculty game the metric (Granovetter-style threshold models). In practice, however, the transition may be smooth rather than discontinuous. Replacing the Heaviside step with the logistic smooth-transition function $\sigma((k - k_0)/\tau)$, where τ controls the transition width, leaves the qualitative results unchanged: for any finite τ , there is a neighbourhood of k_0 in which drift rises rapidly, and the steady-state distribution still shifts toward low x as k grows. The core predictions—drift direction, the divergence between $V(t)$ and $M(t)$, and the existence of a rapid-transition region—are robust to smoothing; only the boundary's sharpness changes. The full check is reported in Online Supplementary Materials Section 3.

$$\frac{\partial e}{\partial t} = - \frac{\partial}{\partial x} [v(x, k, w)e(x, t)] + D \frac{\partial^2 e}{\partial x^2} \quad (16)$$

$$v(x, k, w) = -\mu k \cdot x \cdot (1 - x) \cdot \Theta(k - k_0(w)) \quad (17)$$

For $k \leq k_0(w)$, drift vanishes and Equation (16) reduces to pure diffusion (Eq. 21, Online Supplementary Materials); above the threshold, setting $\partial e / \partial t = 0$ gives the steady-state ODE (Eq. 22, Online Supplementary Materials). True value $V(t)$ is Equation (24) and metric score $M(t)$ Equation (25). For $k > k_0(w)$, Equation (26) gives a steady-state distribution densest at low effort quality, and average quality $\bar{x}^*(k)$ (Eqs. 23 and 27, Online Supplementary Materials) declines monotonically with k above the threshold: as faculty shift effort to metric optimisation, metric scores rise while true value falls—Goodhart's Law dynamically. Note that $M(t)=1-V(t)$ (Equation 25) holds by construction under normalised effort quality; the empirical content of Proposition 5 is therefore not this identity but that

$V(t)$ falls monotonically above the threshold. Figures 3–4 plot the distributions and the divergence between $V(t)$ and $M(t)$. Both layers share the threshold $k_0(w)$; $U(k)$ can be read as steady-state true value $\bar{x}^*(k)$ minus administrative costs Bk ; with $\bar{x}^*(k)$ declining above $k_0(w)$, this supplies the micro-behavioural basis for the exponential dysfunction term. The mapping does not assert a closed-form identity between $U(k)$ and $\bar{x}^*$; it establishes that the static decline has a behavioural foundation in the drift of effort quality.

$$V(t) = \int_0^1 x \cdot e(x, t) dx \quad (24)$$

$$M(t) = \int_0^1 (1 - x) \cdot e(x, t) dx = 1 - V(t) \quad (25)$$

$$e_*(x) = C \cdot \exp \left[-\mu k / D \left(\frac{x^2}{2} - \frac{x^3}{3} \right) \right], k > k_0(w) \quad (26)$$

A complementary micro-foundation for the threshold drift comes from social-norm dynamics. Let $p(k)$ denote the proportion of faculty engaging in strategic metric-gaming. As evaluation intensity rises above the threshold, peer effects and organisational amplification create a positive feedback loop: as more faculty game metrics, the publish-or-perish norm strengthens, lowering the psychological cost of gaming for others. Formally, Eq. (28) (Online Supplementary Materials) expresses how the gaming proportion grows as evaluation intensity rises above the threshold, at rate r once k exceeds k_0 ; r is the social-norm feedback strength glossed in Section 3.2, and p_0 the baseline share of gaming faculty.

Equation (28) (Online Supplementary Materials) shows that $p(k)$ grows exponentially above k_0 , providing a micro-level rationale for both the exponential dysfunction cost in Equation (2) and the threshold-activated drift in Equation (17).

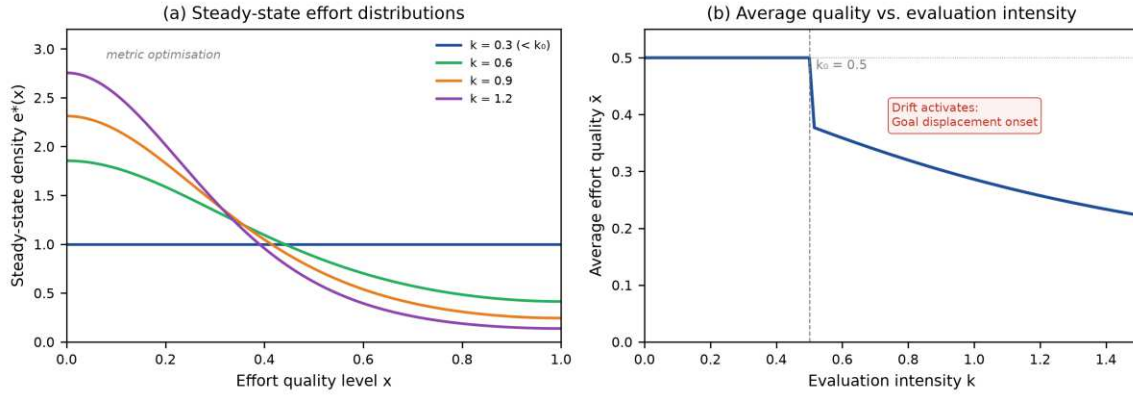


Figure 3. Steady-state effort distributions from the Fokker–Planck model. (a) Probability density $e_*(x)$ for different evaluation intensities k : uniform for $k \leq k_0 = 0.5$, while drift skews effort toward low x (metric optimisation) for $k > k_0$. (b) Average effort quality $\bar{x}$ declines as k rises beyond k_0 . Parameters: $\mu = 1.5, D = 0.1$. Qualitative patterns are robust.

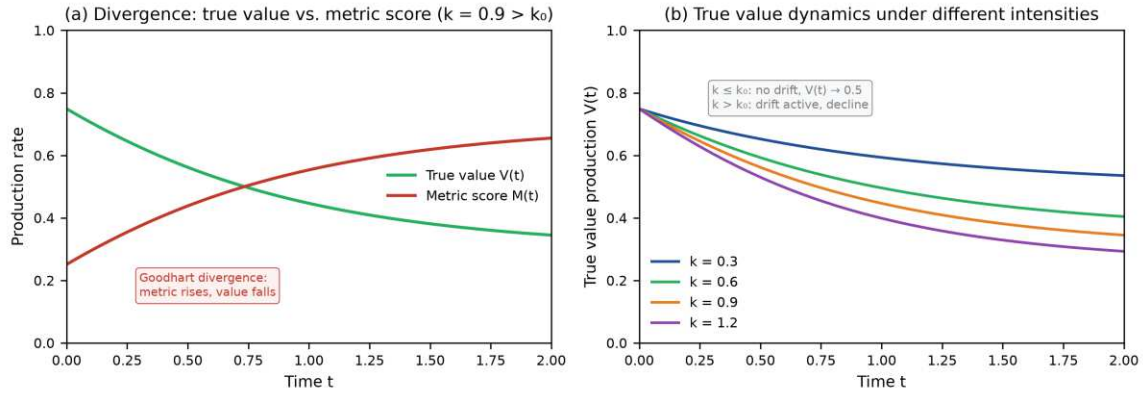


Figure 4. Dynamic divergence of true value and metric score. (a) For $k = 0.9 > k_0$, true value production $V(t)$ declines while metric score $M(t)$ rises, embodying Goodhart’s Law. (b) True-value dynamics $V(t)$ under different evaluation intensities: $k \leq k_0$ yields convergence to uniform quality, while $k > k_0$ yields systematic decline. Parameters: $\mu = 1.5, D = 0.1, \lambda = 5.0$. The divergence pattern is insensitive to moderate parameter changes.

3.5 Parameter setting, empirical anchoring, and systematic sensitivity analysis

Table S4 (Online Supplementary Materials) lists the baseline parameter values and their empirical anchors; the baselines are as followed:

$$A = 1.0, \alpha = 2.0, B = 0.1, C = 0.2,$$

$$\beta = 2.0, k_{00} = 0.5, \gamma = 0.8, \delta = 0.3, \mu = 1.5, D = 0.1, \lambda = 5.0.$$

These normalised values are illustrative, and empirical calibration requires institution-level data. The marginal-returns parameter α governs how quickly alignment benefits saturate: education-production-function estimates typically find strongly diminishing returns to instructional inputs, supporting α in the

range 1–3. The dysfunction-growth rate β governs how rapidly gaming accelerates above threshold: studies of goal displacement suggest dysfunction compounds over a 2–5 year horizon, supporting β in the range 1–4. The baseline threshold $k_{00} = 0.5$ corresponds roughly to a promotion system in which quantitative metrics carry about 50–60% of weight, consistent with research-intensive practice. The threshold-raising coefficient $\gamma = 0.8$ implies that adding qualitative judgment shifts k_0 by roughly 0.04 per 0.1 increase in w .

Sensitivity ranges are in Table S5. Varying β 1.0–3.0 shifts $k^{**} \sim 0.83$ –0.51 (at $w = 0$) while preserving the inverted-U; varying C 0.1–0.5 shifts $k_{be} \sim 1.60$ –0.96; and varying α 1.0–3.0 shifts k^* without eliminating the optimum. The inverted-U, threshold discontinuity, and drift direction hold throughout the region $\alpha, C, \beta, \mu > 0$ and $A\alpha > B$. Qualitatively robust predictions (inverted-U, threshold, drift direction) hold across all admissible parameters; quantitative predictions such as $k^{**} \approx 0.61$ or the location of k_{be} are illustrative, and policy recommendations rest on the robust class.

The robustness boundary is delimited by $A\alpha > B$ evaluated at $k = 0$: if marginal benefit at $k = 0$ does not exceed marginal cost, the inverted-U collapses. This predicts that systems with high fixed administrative costs (B large relative to A) may be net-negative even at low intensity (Woelert, 2023), consistent with documented compliance without learning (Rutherford & Rabovsky, 2014). When $A\alpha$ is large relative to B , k^{**} sits further above k_0 , implying that well-designed systems—where evaluation provides feedback that genuinely improves coordination—can sustain higher intensity before dysfunction sets in.

3.6 Relationship to the Holmström–Milgrom multitask principal-agent model

The model extends the Holmström–Milgrom (H-M) multitask framework. H-M already predicts that excessive measurable incentives are suboptimal when tasks are unmeasurable; we differ in three ways. First, H-M yields a smooth trade-off, whereas ours introduces an explicit threshold k_0 beyond which dysfunction accelerates—interpretable as a critical point in cross-task substitution elasticity: gradual below k_0 , then accelerated above k_0 once a social-norm tipping point is reached. Second, H-M is static, whereas our Fokker–Planck layer adds dynamic drift. Third, H-M omits social amplification, whereas our management feedback loop is an additional dynamic. The approaches are complementary: H-M asks what incentive strength the principal chooses, whereas we take k as given and ask how effective it is; nesting the optimal-contract choice inside the threshold framework is left to future work.

4. Six testable propositions and identification strategies

The model generates six testable propositions. For each, we specify the key variables, data sources, identification assumptions, and potential threats, so that subsequent empirical work can be designed directly from the formal structure.

Proposition 1 (Non-monotonicity). The relationship between evaluation intensity k and organisational effectiveness U is inverted-U-shaped. Identification: institutional panel data over 10–15 years with institution fixed effects. k is measured as the composite four-dimension index of Section 3.1, coded from promotion documents and policy records; U is proxied by forward-citation-weighted research output (five-year forward citations per publication, capturing long-run impact rather than short-run quantity) and student learning outcomes (e.g., graduate employability, alumni surveys). The specification is $U_{it} = \beta_1 k_{it} + \beta_2 k_{it}^2 + \gamma X_{it} + \alpha_i + \delta_t + \varepsilon_{it}$; an inverted-U requires $\beta_1 > 0$ and $\beta_2 < 0$. Threats include omitted variables (leadership changes, resource shocks) and reverse causality (declining universities intensifying evaluation); both are addressed by institution and year fixed effects and lagged k .

Proposition 2 (Threshold effect). A critical threshold k_0 exists beyond which dysfunctions emerge rapidly. Identification: exploit evaluation reforms—China’s 2020 “Break Five Onlys,” the UK REF exercises, or US performance-funding adoption—as quasi-natural experiments using difference-in-differences (DID). The treatment group comprises institutions whose pre-reform quantitative intensity was above the sample median, the control group those below; the reform year defines treatment timing. Because implementation is gradual and varies across provinces and tiers, a staggered DID with bias-corrected estimators (Callaway & Sant’Anna, 2021) is appropriate. Outcomes include forward-citation-weighted research quality and graduate outcomes, with institutional size, disciplinary composition, and regional economic conditions as controls. The key identifying assumption is parallel pre-trends, testable via event-study coefficients. Further details are in the Online Supplementary Materials.

Proposition 3 (Qualitative judgment moderates the threshold). k_0 is higher in systems with substantial qualitative judgment. Identification: mediation models across institutions, estimating the paths from peer-review weight w_p , value-orientation weight w_v , and procedural-justice weight w_j to threshold shifts. w_p is the share of promotion criteria devoted to external peer review; w_v the explicit reference to institutional mission in evaluation documents; w_j the presence of transparent appeals and faculty representation. The model predicts w_p exerts the strongest threshold-raising effect ($\gamma_p > \gamma_v, \gamma_j$),

contingent on peer-review quality. The key threat—endogeneity of qualitative-judgment adoption—is addressed by institution fixed effects and within-institution variation over time.

Proposition 4 (Differential effects of qualitative-judgment sub-dimensions). Peer review raises k_0 more strongly per unit than procedural justice, because it directly reduces the returns to metric gaming; the proposition is designed to distinguish dominant from secondary mechanisms. Identification: hierarchical regression across institutions with varying designs, interacting w_p and w_j with the quadratic term in k . If the mechanism assignment is correct, the interaction $w_p \times k^2$ should attenuate the negative curvature more than $w_j \times k^2$; if instead w_j dominates, the misassignment is exposed. A falsification test predicts that w_v should primarily affect the intercept (A) rather than the curvature k_0 . This is the most directly testable of the three-sub-dimension hypotheses and serves as a within-model robustness check.

Proposition 5 (Dynamic adaptation). The divergence between metric scores and true value accelerates over time when $k > k_0$. Identification: event-study methods tracking institutional metric scores (annual publication counts, grant income) and quality indicators (forward citations per paper, graduate outcomes) over several post-reform years. The divergence slope is predicted to rise monotonically with post-reform k . Synthetic control methods can construct counterfactual trajectories for matched untreated institutions. The key outcome is the ratio $M(t)/V(t)$: the model predicts it exceeds 1 for treated institutions with high pre-reform k , but not for those with low pre-reform k . Threats include secular trends in publication practices (e.g., open-access mandates) affecting all institutions, which year fixed effects absorb.

Proposition 6 (Heterogeneity). k_0 is lower for early-career researchers (where career precarity is higher) but higher in STEM fields, where outputs are more quantifiable. Identification: cross-disciplinary and cross-career-stage subsample analysis. The model predicts k_0 for early-career faculty, because precarious employment increases responsiveness to metric incentives, and k_0 in quantifiable fields, because metrics track true value more closely. Gender heterogeneity is also expected: documented evaluation biases (Valencia, 2022) suggest female faculty may face lower effective thresholds owing to higher service burdens and stereotype threat. Expected directions are specified a priori.

The six propositions are testable with publicly available institutional panel data. Proposition 1 needs institution-level evaluation-policy codes paired with public research-output and student-outcome data; Proposition 2 exploits national reforms (po wu wei, UK REF, US performance-funding adoption) as quasi-exogenous variation; Propositions 3 and 5 need cross-sectional design and longitudinal

metric/quality data. None requires proprietary personnel data, although structural estimation of k_0 will need sufficient policy variation or a structural-break estimator. The propositions are the hypotheses future studies will confront with data.

5. Policy implications

5.1 *The threshold constraint $k < k_0(w)$*

The model yields a clear operational design constraint: evaluation intensity should remain below $k_0(w) = k_{00}(1 + \gamma w)$, ideally near k^{**} . Institutions should audit their systems against the four-dimension k index and benchmark them against peers. The precise numerical value of k_{00} requires empirical calibration, but the directional constraint is robust: do not push metric intensity beyond the point at which qualitative judgment can no longer buffer it. Increasing qualitative judgment w is the primary lever for expanding safe quantitative intensity, but only where qualitative systems are sufficiently resourced and procedurally fair to resist colonisation by metric logic.

5.2 *Three-dimensional policy levers and differentiated recommendations*

Because the three sub-dimensions operate through distinct mechanisms, policy levers are not interchangeable: peer review raises the threshold; value alignment raises the maximum benefit; procedural justice reduces baseline dysfunction. Over-reliance on metrics erodes academic citizenship (Smith & Walker, 2024). Differentiated recommendations follow.

Through the academic-profession literature, the parameters map onto documented features of academic work: rising administrative burden (Woelert, 2023) corresponds to a high fixed-cost term that can push systems net-negative at low intensity; burnout under audit (Lei et al., 2021; Sabagh et al., 2022) is the experiential face of the convex dysfunction cost; precarious early-career structures (Tian & Lu, 2017; O’Brady & Cooper, 2025) imply a lower effective threshold; and faculty’s strategic responses to managerial metrics (Wang et al., 2024) qualify the representative-agent drift. Placing this labour-process literature inside the parameters is what makes the model an account of academic work.

For research-intensive STEM universities, cap quantitative metrics at 50–60% of promotion weight (40–50% qualitative); for humanities and social sciences, 40–50% metrics with 50–60% qualitative; for teaching-focused institutions, 30–40% with peer evaluation of teaching prioritised. Each option carries costs—peer review demands reviewer time (Miller, 2019)—and implementation should proceed from

procedural justice toward peer review. Early-warning indicators (rising publications without forward citations, declining motivation surveys) let institutions detect approach toward $k_0(w)$ before dysfunction entrenches.

The percentages above are illustrative benchmarks, not empirically estimated norms. Consistent with Section 3.5, the model predicts only that the sustainable threshold is systematically higher where outputs are readily quantifiable (research-intensive STEM) and lower where they are not (humanities, teaching-focused institutions); the exact cut-offs should be estimated through Proposition 6 rather than adopted as standing targets. The same logic travels across national systems: because the thresholds and levers are parameters rather than constants, one diagnostic framework applies to every research-assessment regime.

5.3 China’s “Break Five Onlys” reform as an illustrative case

China’s 2020 Break Five Onlys reform is best read not as a critical test—the study offers no primary data—but as an illustrative case surfacing the model’s boundary conditions, especially the institutional capacity required for qualitative judgment to raise the threshold. The same logic applies to the regimes in Section 2.3: systems with resourced peer review sustain higher intensity, whereas formally cutting metric weight without building review capacity yields only symbolic change. China’s uneven implementation—elite universities with deeper review capacity moved furthest, while local institutions retained metrics de facto (Meng & Zhang, 2021; Chen, 2021; Zhang & Wang, 2025)—is consistent with that prediction.

Read through the model, the reform corresponds to decreasing k while increasing w : a higher $k_0(w)$, a larger $A(w)$, and reduced goal displacement. Implementation gaps show why this does not follow automatically: raising w requires capacity—trained reviewers, time, transparent procedures—and lowering k without raising w merely shifts the problem. Two mechanisms explain the gap: symbolic adoption (keeping indicators as de facto screens; Chen, 2021) and information asymmetry under central goals (Meng & Zhang, 2021). Where w is backed by capacity, $k_0(w)$ rises; otherwise reform remains symbolic.

5.4 The contested definition of U and the normative question

The definition of U as “true long-term organisational value creation” is contested, which limits prescriptive force. If U is defined narrowly as research output, the model predicts an inverted-U with a higher k_0 ; defined broadly as student development and social service, the threshold is lower. The

qualitative prediction—that excessive metric intensity undermines whatever U the organisation values—is robust, but the quantitative prescription depends on whose U is optimised. Evaluation systems distribute recognition and resources, so their design should involve faculty in deliberation rather than managerial optimisation alone. Because U is organisation-specific, no universal k^{**} exists across institutions with different missions.

Global rankings (ARWU, QS, THE) are an exogenous pressure outside organisational control, weighting publications and citations and pushing k upward regardless of internal preferences. In model terms they are an additive shock to k : $k_{observed} = k_{internal} + \rho \cdot \text{RankingPressure}$, where ρ captures ranking sensitivity. When this forces $k_{observed}$ above $k_0(w)$, the institution enters the dysfunctional regime even if its internal $k_{internal}$ would have been optimal—ranking-driven intensification (Espeland & Sauder, 2007). Reducing k will not succeed in isolation unless ranking pressure is also addressed through league-table reform or national coordination; institutions cannot unilaterally lower k because ranking competition penalises those that do.

6. Limitations and future research

Several limitations bound the model’s scope. First, homogeneous faculty abstract from disciplinary, career-stage, and gender heterogeneity. The parameters α, β , and k_{00} likely vary systematically: early-career researchers may have lower k_{00} , STEM fields higher k_{00} because outputs are more quantifiable, and female faculty lower thresholds owing to evaluation bias; future work should extend to a heterogeneous-agent framework. Second, k is treated as exogenous, yet universities choose evaluation intensity endogenously as a function of their characteristics, resources, and ranking pressure. This endogeneity complicates causal estimation, and remedies include instrumental variables and the policy-exploiting quasi-experimental designs of Proposition 2.

Third, the model omits ranking-induced competition, which can push universities above their own $k_0(w)$ regardless of internal efficiency; if it forces k above k^{**} , the system can become trapped in a suboptimal equilibrium. Fourth, normalised parameters require calibration; the reported values ($k^{**} \approx 0.61, k_{be} \approx 1.30$) are illustrative. Fifth, the reduced-form $M(t) = 1 - V(t)$ is a structural identity needing independent measurement of true value and metric scores, and the homogeneous-faculty assumption abstracts from departmental politics and power asymmetry. Future work should estimate k_0 from panel data and test the six propositions.

Three further points frame how the model should be read. It is deliberately theory-first: the contribution is a falsifiable apparatus whose six propositions (Section 4) state what empirical work must estimate rather than a verdict on any one system. Formalising Goodhart's and Campbell's Laws structures long-standing qualitative insight; it does not replace qualitative evidence of how evaluation is experienced on the ground, and is best read alongside the ethnographic reactivity literature (Espeland & Sauder, 2007). Finally, the absolute location of k_0 is illustrative—parameters are normalised and uncalibrated, so no institution should read 0.5 as a target. What transfers is the directional diagnosis: safe intensity rises with resourced peer review and procedural fairness, not a universal numeric cut-off.

7. Conclusion

This paper develops a formal threshold model of performance evaluation in higher education, grounded in Goodhart's Law and standard concave-utility assumptions. Its central move is to turn Goodhart from metaphor into a falsifiable, parameterised proposition with an explicit threshold mechanism. Three formulations establish that the inverted-U emerges under standard preferences, qualitative judgment acts through three testable sub-dimensional levers, and above-threshold regimes produce progressive divergence between metric scores and true value.

The choice of k and w reflects political judgments about what universities should value. Increasing qualitative judgment w is the primary lever for expanding safe quantitative intensity, but only when qualitative systems are sufficiently resourced and procedurally fair. The model thus supports a middle path in which quantitative metrics are one component of a broader evaluative framework anchored in professional judgment. Qualitative predictions (inverted-U, threshold, drift direction) are robust to functional-form and parameter variation; quantitative ones require calibration (Section 3.5).

For higher-education research, the payoff is a reframing of the central debate: rather than asking whether metrics are broadly good or bad, the model asks under what combination of intensity, review capacity, and procedural fairness measurement begins to undermine its own goals. The six falsifiable, identifiable propositions invite the field to move past rhetorical condemnation or uncritical adoption toward measured, design-aware evaluation policy

This is a diagnostic rather than a mathematical exercise: it turns whether evaluation is “too much” from a positional dispute into a measurable institutional-design problem—where a sustainable threshold lies and whether measured intensity is about to outrun it. The increment is located precisely within the higher-evaluation literature: Goodhart and Campbell are otherwise interpretive metaphor, Espeland and Sauder's reactivity is powerful but yields no comparative statics, and the quantitative–qualitative debate

has treated qualitative judgment as one lever rather than the three testable sub-dimensions separated here. Because the threshold is a function of institutional design, one template diagnoses China's Break Five Onlys, the UK REF, Italy's ANVUR, and US performance-based funding ; the empirical agenda follows from the six propositions.

References

- Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian Publication Indicator. *Research Evaluation*, 24(2), 106–117. <https://doi.org/10.1093/reseval/rvv003>
- Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and limitations. *SAGE Open*, 9(1). <https://doi.org/10.1177/2158244019829575>
- Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). *Trends in global higher education: Tracking an academic revolution*. Paris, France: UNESCO.
- Anderson, C. G., McQuaid, R. W., & Wood, A. M. (2022). The effect of journal metrics on academic resume assessment. *Studies in Higher Education*, 47(11), 2310–2322. <https://doi.org/10.1080/03075079.2022.2061446>
- Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, 18(2), 215–228. <https://doi.org/10.1080/0268093022000043065>
- Bloch, C., & Schneider, J. W. (2016). Performance-based funding models and researcher behavior: An analysis of the influence of the Norwegian Publication Indicator at the individual level. *Research Evaluation*, 25(4), 371–382. <https://doi.org/10.1093/reseval/rvv047>
- Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>
- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90. [https://doi.org/10.1016/0149-7189\(79\)90048-X](https://doi.org/10.1016/0149-7189(79)90048-X)
- Chen, B. [陈斌]. (2021). 从“表现主义”到“本质主义”——大学学术评价指标化的支配及其超越 [From “expressionism” to “essentialism”: The domination and transcendence of indicator-based academic evaluation in universities]. *高等教育研究 [Journal of Higher Education]*, 42(5), 44–54. https://kns.cnki.net/kcms2/article/abstract?v=wFrWXqGzUjecj8iQAF5r1J_8G_4qvHy7AfXYjQ3Ss9uwnZXZA4MpJ386hpukBkKDmOUQyIW_hMzHlcV63HAIEuKiGzIyfkMMDfC7oBkDQ4QS6JhKrXAHdDm4EPA0pidjuYLj6XC95DDpujdHnh5xwn6AjrLg91ppaj_k6lBLaVJLYk0yFoKbg==&uniplatform=NZKPT&language
- Croucher, G., & Lacy, W. B. (2022). The emergence of academic capitalism and university neoliberalism: Perspectives of Australian higher education leadership. *Higher Education*, 83(2), 279–295. <https://doi.org/10.1007/s10734-020-00655-7>

- D'Angelo, C. A. (2025). The 2015–2019 VQR, Italy's third national research assessment: An intended informed peer review or a DIY bibliometric exercise? *Research Evaluation*, 34, rvaf045. <https://doi.org/10.1093/reseval/rvaf045>
- de Rijcke, S., Wouters, P., Rushforth, A. D., Franssen, T., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use: A literature review. *Research Evaluation*, 25(2), 161–169. <https://doi.org/10.1093/reseval/rvv038>
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668. <https://doi.org/10.1037/0033-2909.125.6.627>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. New York, NY: Plenum.
- Deem, R. (2001). Globalisation, New Managerialism, Academic Capitalism and Entrepreneurialism in Universities: Is the local dimension still important? *Comparative Education*, 37(1), 7–20. <https://doi.org/10.1080/03050060020020408>
- Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1–40. <https://doi.org/10.1086/517897>
- Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology*, 24, 313–343. <https://doi.org/10.1146/annurev.soc.24.1.313>
- Espeland, W. N., & Vannebo, B. L. (2007). Accountability, quantification, and law. *Annual Review of Law and Social Science*, 3, 21–43. <https://doi.org/10.1146/annurev.lawsocsci.2.081805.105908>
- Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5), 589–611. <https://doi.org/10.1111/1467-6419.00150>
- Goodhart, C. A. E. (1975). Monetary theory and practice: The UK experience. Bank of England Panel of Academic Consultants, 4, 1–14.
- Granovetter, M. S. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.
- Griffin, G. (2022). The 'work–work balance' in higher education: Between over-work, falling short and the pleasures of multiplicity. *Studies in Higher Education*, 47(11), 2190–2203. <https://doi.org/10.1080/03075079.2021.2020750>
- Hammarfelt, B., & de Rijcke, S. (2015). Accountability in context: Effects of research evaluation systems on publication practices. *Research Evaluation*, 24(1), 63–77. <https://doi.org/10.1093/reseval/rvu029>
- Hazelkorn, E. (2008). Learning to live with league tables and ranking: The experience of institutional leaders. *Higher Education Policy*, 21(2), 193–215. <https://doi.org/10.1057/hep.2008.1>
- Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251–261. <https://doi.org/10.1016/j.respol.2011.09.007>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>

- Hillman, N. W., Tandberg, D. A., & Gross, J. P. K. (2014). Performance funding in higher education: Do financial incentives impact college completions? *The Journal of Higher Education*, 85(4), 484–513.
<https://doi.org/10.1080/00221546.2014.11777349>
- Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7(Special Issue), 24–52.
https://doi.org/10.1093/jleo/7.special_issue.24
- Hyle, A. E. (1999). Faculty evaluation: A prickly pair. *Higher Education*, 38(3), 351–371.
<https://doi.org/10.1023/a:1003735227936>
- Jiang, X. [姜晓晖], & Wang, W. [汪卫平]. (2021). 高校教师学术声誉研究：一种探索性激励机制设计 [Research on faculty academic reputation: An exploratory incentive mechanism design]. *中国高教研究* [China Higher Education Research], 37(4), 42–47. <https://doi.org/10.16298/j.cnki.1004-3667.2021.04.08>
- Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, 18(4), 769–783.
<https://doi.org/10.2307/255378>
- Lei, W., Li, J., Li, Y., Castaño, G., Yang, M., & Zou, B. (2021). The boundary conditions under which teaching–research conflict leads to university teachers’ job burnout. *Studies in Higher Education*, 46(2), 406–422.
<https://doi.org/10.1080/03075079.2020.1811218>
- Li, A. Y., & Hu, X. (2025). Goodbye performance-based funding: Policy abandonment of performance funding for higher education in the United States. *Research in Higher Education*, 66(1), 1–25. <https://doi.org/10.1007/s11162-024-09828-0>
- Li, D., Lo, W. Y. W., & Yang, R. (2024). Unpacking the discourse surrounding the impact agenda in the Hong Kong Research Assessment Exercise 2020. *Research Evaluation*, 33, rva034. <https://doi.org/10.1093/reseval/rvae034>
- Liu, Z. [刘振天]. (2020). 教育评价破“五唯”重在立“四新” [Breaking the “Five Onlys” in educational evaluation: Focusing on establishing “Four News”]. *国家教育行政学院学报* [Journal of National Academy of Education Administration], (11), 13–15. <https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&filename=GJXZ202011003>
- Marginson, S. (2016). The worldwide trend to high participation higher education: Dynamics of social stratification in inclusive systems. *Higher Education*, 72(4), 413–434. <https://doi.org/10.1007/s10734-016-0016-x>
- Meng, W. [孟渥], & Zhang, Q. [张群]. (2021). 科研评价“五唯”何以难破——制度分析的视角 [Why is the “Five Onlys” in research evaluation so difficult to break: From an institutional analysis perspective]. *中国高教研究* [China Higher Education Research], 37(9), 51–58. <https://doi.org/10.16298/j.cnki.1004-3667.2021.09.08>
- Merton, R. K. (1968). The Matthew effect in science. *Science*, 159(3810), 56–63.
<https://doi.org/10.1126/science.159.3810.56>
- Miller, J. (2019). Where does the time go? An academic workload case study at an Australian university. *Journal of Higher Education Policy and Management*, 41(6), 633–645. <https://doi.org/10.1080/1360080X.2019.1635328>

- Morgan, K., & Kearney, C. (2021). Governing by narratives: REF impact case studies and restrictive storytelling in performance measurement. *Studies in Higher Education*, 46(10), 2012–2028. <https://doi.org/10.1080/03075079.2021.1978965>
- Morgan Jones, M., Manville, C., & Chataway, J. (2022). Learning from the UK's research impact assessment exercise: Implications for evaluators. *The Journal of Technology Transfer*, 47(3), 722–746. <https://doi.org/10.1007/s10961-017-9608-6>
- Muller, J. Z. (2018). *The tyranny of metrics*. Princeton, NJ: Princeton University Press.
- Murphy, M. (2009). Bureaucracy and its limits: Accountability and rationality in higher education. *British Journal of Sociology of Education*, 30(6), 683–695. <https://doi.org/10.1080/01425690903235169>
- O'Connell, C., O'Siochru, C., & Rao, N. (2021). Academic perspectives on metrics: Procedural justice as a key factor in acceptance. *Studies in Higher Education*, 46(3), 548–562. <https://doi.org/10.1080/03075079.2019.1643306>
- O'Brady, S., & Cooper, B. (2025). Academic capitalism and precarity in the neoliberal university: Job insecurity and stress in two liberal market economies. *Industrial Relations Journal*, 56(4), 291–300. <https://doi.org/10.1111/irj.12466>
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown.
- Ortagus, J. C., Kelchen, R., Rosinger, K., & Voorhees, N. (2020). Performance-based funding in American higher education: A systematic synthesis of the intended and unintended consequences. *Educational Evaluation and Policy Analysis*, 42(4), 494–517. <https://doi.org/10.3102/0162373720953128>
- Power, M. (1997). *The audit society: Rituals of verification*. Oxford, UK: Oxford University Press.
- Pusser, B., & Marginson, S. (2013). University rankings in critical perspective. *The Journal of Higher Education*, 84(4), 544–554. <https://doi.org/10.1080/00221546.2013.11777301>
- Rablen, M. D. (2010). Performance targets, effort and risk-taking. *Journal of Economic Psychology*, 31(3), 372–382. <https://doi.org/10.1016/j.joep.2010.05.002>
- Rutherford, T. A., & Rabovsky, T. M. (2014). Evaluating impacts of performance funding policies on student outcomes in higher education. *The ANNALS of the American Academy of Political and Social Science*, 655(1), 185–208. <https://doi.org/10.1177/0002716214541048>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Sabagh, Z., Hall, N. C., Saroyan, A., & Trépanier, S.-G. (2022). Occupational factors and faculty well-being: Investigating the mediating role of need frustration. *The Journal of Higher Education*, 93(4), 559–584. <https://doi.org/10.1080/00221546.2021.2004810>

- Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63–82. <https://doi.org/10.1177/000312240907400104>
- Schneider, J. W., Aagaard, K., & Bloch, C. W. (2016). What happens when national research funding is linked to differentiated publication counts? A comparison of the Australian and Norwegian publication-based funding models. *Research Evaluation*, 25(3), 244–256. <https://doi.org/10.1093/reseval/rvv036>
- Seeber, M., Cattaneo, M., Meoli, M., & Malighetti, P. (2019). Self-citations as strategic response to the use of metrics for career decisions. *Research Policy*, 48(2), 478–491. <https://doi.org/10.1016/j.respol.2017.12.004>
- Shore, C., & Wright, S. (2000). Coercive accountability: The rise of audit culture in higher education. In M. Strathern (Ed.), *Audit cultures: Anthropological studies in accountability, ethics, and the academy* (pp. 57–74). New York, NY: Routledge. <https://doi.org/10.4324/9780203449721-7>
- Siciliani, L. (2009). Paying for performance and motivation crowding out. *Economics Letters*, 103(2), 68–71. <https://doi.org/10.1016/j.econlet.2009.01.022>
- Skalse, J., Howe, N. H. R., Krashennnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems 35* (NeurIPS 2022). https://proceedings.neurips.cc/paper_files/paper/2022/hash/40296e1b1e916d351b6d61ae113b53c8-Abstract-Conference.html
- Slaughter, S., & Leslie, L. L. (1997). *Academic capitalism: Politics, policies, and the entrepreneurial university*. Baltimore, MD: Johns Hopkins University Press.
- Smith, S., & Walker, D. (2024). The instrumental academic: Collegiality and the value of academic citizenship in contemporary higher education. *Higher Education Quarterly*, 78, e12551. <https://doi.org/10.1111/hequ.12551>
- Stupnisky, R. H., Larivière, V., Hall, N. C., & Omojiba, O. (2023). Predicting research productivity in STEM faculty: The role of self-determined motivation. *Research in Higher Education*, 64(4), 598–621. <https://doi.org/10.1007/s11162-022-09718-3>
- Tien, F. F., & Blackburn, R. T. (1996). Faculty rank system, research motivation, and faculty research productivity. *The Journal of Higher Education*, 67(1), 2–22. <https://doi.org/10.1080/00221546.1996.11780246>
- Tian, M., & Lu, G. (2017). What price the building of world-class universities? Academic pressure faced by young lecturers at a research-centered university in China. *Teaching in Higher Education*, 22(8), 957–974. <https://doi.org/10.1080/13562517.2017.1319814>
- Valencia, E. (2022). Gender-biased evaluation or actual differences? Fairness in the evaluation of faculty research performance. *Higher Education*, 83(6), 1315–1333. <https://doi.org/10.1007/s10734-021-00744-1>

- Wang, Y., Li, X., Cheng, H., & Zhang, L. (2024). Understanding responses to managerialism: Face culture and university faculty working under the shadow of ‘publish or perish’. *Culture and Organization*, 30(4), 405–424. <https://doi.org/10.1080/14759551.2023.2283072>
- Woelert, P. (2023). Administrative burden in higher education institutions: A conceptualisation and a research agenda. *Journal of Higher Education Policy and Management*, 45(4), 409–422. <https://doi.org/10.1080/1360080X.2023.2190967>
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kane, S., Pewall, D., Shapira, P., & Sutter, K. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: HEFCE. <https://doi.org/10.13140/RG.2.1.4929.1363>
- Yan, G. [阎光才]. (2019a). 学术影响力评价的是非争议 [Controversies over academic impact evaluation]. *教育研究* [Educational Research], 40(6), 16–26. <https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&filename=JYYJ201906003>
- Zhang, X., & Wang, P. (2025). The politics of higher education in China: The signal–response mechanism, downward tiered pressure escalation (cengceng jiama), and the Double First-Class University Initiative. *Comparative Education*, 61(2), 202–224. <https://doi.org/10.1080/03050068.2024.2376986>
- Yan, G. [阎光才]. (2019b). 谨慎看待高等教育领域中各种评价 [A cautious view of various evaluations in higher education]. *清华大学教育研究* [Journal of Education, Tsinghua University], 40(1), 1–4. <https://doi.org/10.14138/j.1001-4519.2019.01.000104>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [OnlineSupplementaryMaterials.docx](#)
- [ReproducibilityPackage.zip](#)