Adv. J. Manag. Humanit. Soc. Sci. (2026), Volume 2, Issue 5, 527-544

Advanced Journal of Management, Humanity and Social Science

Journal homepage: https://www.ajmhss.com/

DOI: https://doi.org/10.5281/zenodo.23927666

Designing a Meritocracy System in Public Organizations Using Competency-Based Management and Performance Analytics

Reza Salimi Beni

Master of Science in Information Technology Management, Specialization in Advanced Information Systems, Islamic Azad University, Electronic Campus, Tehran, Iran

Article info

Received: 25.06.2026

Accepted: 01.11.2026

Available Online: 02.11.2026

Checked for Plagiarism: Yes

Keywords:

meritocracy; public organizations; competency-based management; performance analytics; procedural justice

ABSTRACT

Public organizations need appointment and advancement systems that connect demonstrable capability with public service results while protecting equal opportunity and administrative impartiality. Competency frameworks often describe desirable behavior without specifying how evidence should influence employment decisions. Performance dashboards can make decisions more consistent but may also reward favorable assignments, reproduce historical disadvantage, or obscure political intervention. This article develops an integrated meritocracy system through a targeted synthesis of recent academic and institutional publications from 2024-2026 and an explicit design analysis. The study produces a role-based competency architecture, a contextualized performance measurement model, an auditable decision process, and a prospective validation protocol. Its central design principle is to distinguish eligibility, evidence sufficiency, comparative assessment, and final authorization. Integrity requirements and essential competencies operate as noncompensatory conditions; a transparent composite supports comparison only after these conditions are satisfied. A worked example with six entirely hypothetical candidates demonstrates score calculation, uncertainty bounds, sensitivity to policy weights, and the consequences of missing evidence. The example shows why a small numerical advantage cannot establish a defensible appointment when rankings reverse under plausible weighting changes. The proposed system links decisions to development opportunities, independent review, documented reasons, and monitoring of distributional effects. The results are analytical design outputs, not findings from an implemented organization or evidence of causal effectiveness. Future evaluation should examine measurement validity, procedural justice, service quality, and administrative costs through a staged pilot with independent assessment. The article contributes a practical account of meritocracy as a governed evidentiary process rather than an unqualified ranking of employees.

Introduction

Employment decisions determine which capabilities public institutions can use and whose judgment carries authority. A promotion system that rewards personal sponsorship may retain formal recruitment procedures while weakening the connection between role requirements and appointment. A system that relies exclusively on output counts can create a different problem: employees learn to maximize visible transactions even when complex cases, legal safeguards, and equitable access require time.

Designing meritocracy therefore involves specifying what counts as relevant contribution, how evidence is produced, and who can challenge an interpretation. The organizational problem concerns both capability and legitimate authority.

Recent research provides an important foundation. Oliveira et al [1]. synthesize evidence connecting merit principles with government performance,

*Corresponding Author: **Reza Salimi Beni,** rezasalimi1977@gmail.com*

while Goulden [2] examines the purposes and risks of public sector competency frameworks.

These publications support treating merit and competence as institutional design questions. Neither implies that a single universal score can identify the best employee across unrelated administrative functions. A tax investigator, a policy analyst, and a disability benefits officer produce different forms of value under different constraints. A credible system must respect these distinctions while applying consistent standards of evidence and fairness.

Performance management supplies information that competency descriptions alone cannot provide. Competence concerns the ability to undertake relevant work; performance concerns observed contribution under actual conditions. Their relationship is contingent. An employee may possess strong analytical skills but lack access to usable data, or may deliver high transaction volumes because a manager allocates uncomplicated cases. The proposed system therefore considers opportunities to perform alongside observed outcomes. It also separates technical proficiency from legally and ethically acceptable conduct. Strong output cannot automatically compensate for a substantiated failure to meet an essential integrity requirement.

Digitalization makes integration technically feasible while increasing the consequences of design mistakes. Administrative systems can connect recruitment, work samples, service records, and development histories. However, a database can preserve a biased decision as efficiently as a sound one. The availability of a variable does not establish its relevance to merit. Monitoring communication patterns or computer activity may generate abundant observations without measuring public service contribution. The present design limits data collection to defensible purposes and treats model complexity as a cost that requires justification.

The research question is how public organizations can combine competency-based management and performance analytics into a transparent, context-sensitive, and contestable meritocracy system. Three subsidiary questions guide the analysis. What constructs and evidence should be distinguished? How should evidence inform decisions without concealing uncertainty or allowing unacceptable trade-offs? What governance and evaluation arrangements are necessary before the system influences careers? These questions are answered through a conceptual design, supported by recent sources and tested for internal consistency using an illustrative calculation.

The contribution is an integrated specification connecting measurement to authorization and review. Its components are deliberately usable without machine learning. The article does not report a field survey, expert consensus exercise, or organizational intervention. Instead, it provides design propositions, a reproducible numerical demonstration, and a protocol for subsequent empirical testing. This distinction matters because technical coherence cannot establish fairness in practice. A framework becomes credible only when evidence from implementation supports its measurement assumptions and institutional safeguards.

Materials and Methods

Study design and epistemic status:

The study uses a conceptual design approach with a targeted narrative evidence synthesis. The object of design is an organizational decision system rather than a predictive algorithm. The materials consist of recent peer-reviewed research, an academic working paper, international organization publications, and government guidance. The analysis translates these materials into requirements and then examines whether the proposed components fit together logically. The resulting architecture is a prescriptive contribution whose usefulness remains open to empirical assessment.

No primary human participant data were collected. No organization supplied personnel records, and no interviews, Delphi rounds, structural equation models, or significance tests were conducted. The worked example is deliberately constructed and has no population interpretation. Its purpose is to make the scoring rule inspectable, reveal rank reversals, and demonstrate the effect of governance conditions. Statements about likely mechanisms are presented as propositions, not estimated effects. This study design was selected because no organizational dataset or approved access to employee information was available.

Source identification and temporal scope:

Source discovery was undertaken on 3 November 2026 using web searches and publisher or issuing-body pages. The publication window was the three calendar years 2024–2026, up to the search date. Search concepts combined merit principles, public sector competencies, performance management, people analytics, skills-based hiring, fairness, and algorithmic accountability. Targeted searches also examined OECD, United Nations, NIST, UK government, and information regulator publications. Publisher metadata were used to check titles, dates, author information, and persistent identifiers when available. The synthesis is not an exhaustive systematic review. It does not claim comprehensive coverage of subscription databases, a registered protocol, independent duplicate screening, or a PRISMA flow. The reference set comprises 25 distinct publications selected for direct relevance to at least one design requirement. Institutional publications were retained because they address administrative implementation and assurance

arrangements that journal articles may discuss less directly. Their inclusion does not convert recommendations into evidence that a particular system works. Where metadata distinguish online publication and issue year, the reference follows the version specified in the bibliography.

Inclusion and exclusion criteria:
Eligible materials were published within the stated window, had an identifiable publisher or issuing institution, and addressed public employment, competency development, performance measurement, human resource analytics, or safeguards relevant to consequential personnel decisions. General AI guidance was eligible only where it informed a concrete control, such as provenance checks for generated summaries. Sources were excluded when they consisted only of vendor promotion, lacked a clear publication identity, or offered claims without enough accessible information to characterize their relevance. Multiple pages describing the same report were treated as one publication.
The review distinguishes several evidence roles. Public administration studies inform the interpretation of merit and role fit. Analytics research informs organizational adoption and bias management. International reports provide comparative administrative context. Government guidance supplies examples of assurance and transparency practices. These roles are not interchangeable. Evidence from private financial firms, for example, can identify a bias-management mechanism but cannot establish its magnitude or effectiveness in a civil service. Such transfer limits are made explicit in the discussion.

Extraction and synthesis procedure:
For each retained publication, the analysis recorded its bibliographic identity, publication type, relevant construct, practical implication, and transfer limitation. The synthesis used four organizing categories: role requirements, evidence quality, decision rules, and accountability. A source could inform more than one category, but its recurrence was not counted as independent support. The analysis did not pool effect sizes because the materials differ in research design, unit of analysis, and purpose. Descriptive reports and policy guidance were not assigned statistical weights.
Design requirements were formed by asking what would have to be true for a personnel decision to be defensible. A role must be specified before evidence is scored. A measure must represent the intended construct before its precision matters. A comparison must account for material differences in work conditions. An affected person must be able to examine relevant evidence and obtain an independent review. These requirements are analytical judgments informed by the selected literature; they are not a consensus standard or a validated measurement scale.

Table 1. Evidence categories and their design roles

Evidence category	Examples in reference set	Design use	Interpretive limit
Public administration research	Merit review [1]; public sector fit [6]	Define suitability and institutional purpose	Associations do not validate the proposed artifact
Competency and skills analysis	Framework review [2]; digital skills [3]	Develop role profiles and observable anchors	Local content validation remains necessary
HR analytics research	Adoption review [7]; bias research [8, 9]	Identify data and deployment controls	Private sector findings need contextual testing
Comparative government reporting	Workforce and assessment reports [4, 5]	Identify administrative implementation concerns	Country aggregates cannot rank individuals
Assurance and transparency guidance	NIST and UK materials [12-15]	Specify traceability and review practices	Guidance is not proof of effectiveness
Service and integrity reporting	UN and OECD reports [16-20, 25]	Align the system with public purposes	System-level outcomes have multiple causes

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

Artifact construction and analytical evaluation:
The artifact was constructed in five steps. First, the analysis defined merit at the decision level. Second,

it identified competency and performance components and specified how each could be observed. Third, it established admissibility, sufficiency, and authorization rules. Fourth, it constructed a synthetic example and examined alternative weights and measurement bounds. Fifth, it specified institutional controls and a future evaluation protocol. This sequence makes policy choices visible before numerical aggregation begins.
The analytical evaluation addresses coherence, traceability, interpretability, and sensitivity. Coherence asks whether rules contradict one another, such as permitting a high score to override an essential qualification. Traceability asks whether an assessor can connect a decision to an observation and an applicable rule. Interpretability asks whether an employee can understand what could change a decision. Sensitivity asks whether a conclusion depends heavily on uncertain scores or discretionary weights. These checks demonstrate properties of the proposed design; they do not establish organizational effectiveness.

Reproducibility and boundaries:
All values used in the worked example, including weights and thresholds, are reported in the article. Scores can be reproduced with ordinary spreadsheet arithmetic. No fitted model, proprietary dataset, or random-number procedure is required. Thresholds are illustrative policy parameters, not standards derived from prior studies. Their use in an organization would require job analysis, legal review, stakeholder consultation, and validation. The candidate labels identify fictional records only.
The study assumes a public organization with identifiable job families, lawful authority to assess role-relevant competence, and an available independent review function. It does not assume a particular civil service statute, national jurisdiction, or pay regime. Competitive recruitment, internal advancement, and developmental assignment require different evidence. The core architecture can support all three, but the example addresses internal promotion within one job family. Decisions about sanctions or dismissal require separate legal and evidentiary processes and are outside the scoring model.

Conceptual Foundations

Merit as a decision-specific construct:
Merit is defined here as demonstrated suitability for a specified public role, established through relevant and sufficiently reliable evidence under fair procedures. The definition avoids treating merit as an intrinsic measure of a person's general worth. It connects a particular decision to the capabilities and responsibilities of the destination role. Seniority, credentials, and prior ratings may supply evidence where justified, but none is automatically equivalent to suitability. A requirement is defensible only when its connection to the work can be explained.
This definition has three practical consequences. Criteria must precede candidate assessment. Evidence rules must apply consistently to all candidates in the same competition. Decision-makers must explain how the selected person meets the role requirements. An organization cannot establish merit merely by publishing a ranking after managers have privately negotiated the outcome. Nor can it claim equal opportunity if only favored employees receive assignments through which essential evidence can be acquired. Fair access to evidence-generating opportunities belongs within the system's scope.

Competence and role fit:
Competency-based management organizes the knowledge, skills, behaviors, and values required for work. Recent work on digital government skills [3] and person–environment fit [6] reinforces the need to relate capability to organizational demands. The present framework uses role fit narrowly: an evidenced match between a person's capabilities and the requirements of a particular position. It excludes personal similarity to existing leaders, stylistic conformity, and unsupported judgments about cultural compatibility.
A competency model should identify observable behavior at meaningful proficiency levels. “Strategic ability” is too vague to score consistently. A more useful indicator describes an employee identifying alternative policy options, testing assumptions, and explaining distributional consequences in a documented briefing. Such an indicator can be assessed through work samples and structured questioning. It also makes feedback actionable. Candidates can understand which aspect of their performance needs improvement rather than receiving a label that appears fixed and personal.

Performance and public value:
Performance is the observed contribution to role-relevant objectives within an identified period and context. It includes output, quality, timeliness, and the manner in which services are delivered. Government at a Glance 2025 [4] provides a comparative account of public sector performance assessment, while the OECD public servant survey report [5] places workforce experiences alongside organizational domains. These sources motivate a design that connects assessment to work conditions and institutional outcomes without assuming that employee ratings directly measure public value.
Public value is represented through an explicit statement of the service's mandate and obligations. For a licensing office, this might include accurate and timely decisions, accessible procedures, consistent legal treatment, and proportionate use of resources. These outcomes can conflict. Reducing

average processing time by rejecting difficult applications would undermine the intended purpose. A performance indicator must therefore specify its denominator, exclusions, quality conditions, and potential behavioral response. The model does not equate activity with achievement.

Procedural justice and institutional credibility:
A decision can be numerically reproducible yet procedurally unfair. Candidates may have had no reasonable access to the assessment task, may be unable to correct an incorrect record, or may face a reviewer who participated in the original decision. Procedural credibility therefore requires accessible criteria, consistent treatment, opportunities to respond, and reasoned review. The framework treats these as design conditions. It does not infer that employees will perceive fairness simply because the conditions are written into a policy.
The OECD trust reports [18] concern broader institutional relationships rather than a direct test of promotion procedures. They support attention to fairness and integrity as relevant governance issues, but no numerical trust effect is imported into the model. The proposed connection between employment procedures and institutional legitimacy remains a mechanism to test. A public organization should examine employee perceptions and citizen service experiences separately, since improvement in one does not guarantee improvement in the other.

Analytics as organizational practice:
People analytics includes the processes through which personnel data become organizational evidence. Alam et al [7]. Emphasize adoption challenges and stakeholder engagement, and Milanez et al [10]. Discuss algorithmic management and accountability concerns. Their relevance is that the usefulness of analytics depends on how information is interpreted and acted upon. A dashboard does not assign responsibility for data errors, establish appeal rights, or determine whether an observed association is a valid basis for promotion. The framework consequently begins with descriptive and diagnostic analysis. Descriptive analysis identifies evidence coverage and score distributions. Diagnostic analysis investigates differences between assessors, assignments, and groups. Predictive tools remain optional and require a separate justification. Prescriptive recommendations cannot supersede lawful authority. Even a simple weighted average can have consequential effects and therefore needs version control, change authorization, and inspection. Governance obligations arise from the decision's consequences, not only from the sophistication of the software.

Results

Integrated architecture:
The first result is an architecture with six connected components: role definition, competency evidence, performance evidence, contextual interpretation, decision authorization, and institutional learning. The system begins when a vacancy or advancement pathway is defined. It ends when a decision has been explained, reviewed where necessary, and translated into a development or placement plan. Records connect these components so that a later reviewer can reconstruct the decision using the rules and evidence available at the time.
The architecture separates the evidence repository from the authorization process. Assessors can enter observations without controlling the final appointment, and appointing officials cannot silently alter the underlying record. A change requires an identified author, reason, timestamp, and retained prior version. This separation reduces the opportunity to rewrite evidence to match an intended outcome. It also protects assessors from pressure by making their original judgments recoverable. Accountability is distributed across roles while the final decision retains a named owner.



Figure 1. Proposed architecture linking role definition, evidence, decisions, and learning. Arrows indicate process connections, not estimated causal effects.

The feedback path returns information about assessment problems, skill gaps, and service outcomes to the role framework. It does not automatically retrain a model using past promotion decisions as correct labels. Historical appointments may reflect the very favoritism the system seeks to address. Learning therefore requires independent examination of outcomes, including whether appointees perform destination-role tasks effectively and whether the system excludes capable people. A stable framework is desirable, but stability cannot justify retaining criteria that fail validation.

Role and competency architecture:

Each job family contains a limited set of essential competencies, additional differentiating competencies, and contextual descriptors. Essential competencies reflect requirements without which lawful and effective work is not possible. Differentiating competencies help compare eligible candidates. Contextual descriptors capture the work setting without turning it into a personal score. For example, legal reasoning may be essential for an adjudication role, advanced data visualization may be differentiating, and caseload complexity may be contextual.

Four proficiency levels are proposed: guided execution, independent execution, handling complex cases, and improving organizational practice. Levels describe the complexity and independence of demonstrated work. They do not correspond automatically to grade, years of service, or formal educational level. A person may demonstrate advanced technical proficiency while requiring support in supervisory practice. The profile therefore remains multidimensional. Aggregation supports a decision only after the organization has examined essential requirements individually.

Table 2. Proposed competency domains and observable evidence

Domain	Illustrative behavioral anchor	Possible evidence	Scoring caution
Public service and integrity	Explains conflicts of interest and applies required safeguards	Structured dilemma task; adjudicated records where lawful	No adverse inference from an allegation alone
Technical and legal capability	Applies relevant rules to an unfamiliar case and justifies exceptions	Anonymized work sample; structured oral defense	Separate knowledge from writing style
Analytical judgment	Compares options and tests material assumptions	Policy briefing; case analysis	Assess reasoning rather than policy agreement
Service and accessibility	Explains a decision clearly and identifies an access barrier	Observed service task; quality review	Account for communication accommodations
Collaboration and supervision	Gives usable feedback and allocates work fairly	Supervisory simulation; documented observations	Avoid popularity or similarity judgments
Digital and data capability	Checks data quality and recognizes limits of automated outputs	Practical data task; process audit	Do not substitute software certificates for ability
Learning and adaptation	Applies feedback and shares a documented improvement	Before-and-after work samples; learning record	Consider access to developmental opportunities

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

The competency dictionary should contain a definition, behavioral anchors, acceptable evidence sources, assessment conditions, and a review date. A small dictionary that assessors use consistently is preferable to a large catalogue that encourages impressionistic scoring. Each indicator should be tested with contrasting examples. If assessors cannot distinguish adjacent proficiency levels using realistic work, the indicator needs revision. Complexity in the dictionary should follow from the work, not from a desire to make the framework appear comprehensive.

Digital and AI-related capabilities are included where relevant to the role. The OECD AI workforce brief [21] identifies capability development as a public administration concern, while its skills-first report [22] addresses the recognition and use of skills. The design implication is to assess what employees can demonstrate rather than merely whether they attended training. A certificate can identify an opportunity to acquire knowledge; a work sample can show whether that knowledge is

applied appropriately. The two forms of evidence serve different purposes.

Evidence collection and assessment quality:
Evidence is collected from structured work samples, observed work, validated administrative records, and structured interviews. Multiple sources are useful only when their limitations are understood. Five ratings based on the same supervisor's impression do not provide five independent observations. The evidence register identifies both the immediate source and the underlying event. This prevents duplicate counting when a single successful project appears in a manager rating, a portfolio, and a recognition record.
Assessment tasks should resemble destination-role demands. A promotion to supervisory work requires evidence of planning, feedback, conflict handling, and accountable resource allocation, even if the candidate's current role is primarily technical. Existing employees need accessible alternatives when current assignments do not allow demonstration. A standardized simulation may supply such evidence. Accommodations should preserve the construct being assessed and be documented separately from scores so that necessary accessibility measures do not become a disadvantage.
Before scoring real candidates, assessors independently rate benchmark examples and discuss the grounds for disagreement. The calibration record should distinguish misunderstanding of criteria from legitimate differences in interpretation. For an operational pilot, reliability can be examined with an intraclass correlation or a generalizability analysis appropriate to the assessment structure. These are prospective methods, not analyses performed in this study. A favorable coefficient would still not establish content validity; assessors can agree on an irrelevant indicator.
Evidence quality is represented separately from assessed proficiency. A high competency rating based on one old example may be less secure than a slightly lower rating supported by recent observations across several tasks. The framework records recency, coverage, independence, and relevance, but does not conceal them in a single confidence score. Decision-makers receive a profile of weaknesses. Where an essential requirement lacks adequate evidence, the response is an additional assessment or a deferred determination rather than an assumed low ability.

Performance indicator design:
The performance module combines service outputs with quality, timeliness, and contribution to collective work. Its unit of comparison is a role family and assessment period, not the entire organization. Measures should distinguish an employee's contribution from team results and external conditions. A policy analyst's briefing may be technically sound even when political authorities select another option. An inspector may identify more violations because of assigned sector risk rather than superior detection skill. Attribution must therefore be examined before consequences are attached. For each indicator, the organization specifies the event counted, denominator, time window, data source, validation procedure, and foreseeable gaming behavior. A service quality measure might use independently reviewed errors per completed case, with the sample selection process stated in advance. A timeliness measure should identify legally required waiting periods and dependencies on third parties. Both the raw indicator and any adjustment remain visible. Adjustments are explanations that require evidence, not opportunities to modify inconvenient results informally.

Table 3. Illustrative performance measurement specification

Dimension	Operational measure	Context to record	Control against distortion
Service volume	Eligible completed cases per available work period	Case mix; part-time schedule; system downtime	Exclude procedurally invalid closures
Quality	Independently verified errors per reviewed case	Sampling frame; task complexity	Blind review where feasible; retain denominator
Timeliness	Share completed within the applicable service standard	External waits; statutory pauses; peak demand	Monitor difficult cases and overdue tail
Accessibility	Observed compliance with service access requirements	Language needs; channel availability	Use structured evidence rather than popularity
Collective contribution	Documented support to team delivery and knowledge reuse	Assigned duties; mentoring opportunities	Avoid double counting the same project
Resource stewardship	Role-relevant use of authorized resources	Budget authority; staffing; mandated constraints	No reward for unlawful or harmful cost cutting

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

Performance should be measured over a period long enough to represent recurring work and short enough to remain relevant to the destination role. No universal duration is imposed. A six-month transaction record may be informative for a high-volume service but inadequate for a policy development cycle. The appraisal plan records why its window is suitable. Where demand is seasonal, comparisons use comparable periods. Where a crisis changes duties, the record identifies the change rather than forcing normal targets onto exceptional work. A quality constraint prevents speed from dominating a service measure. For instance, cases closed outside required legal procedures should not contribute positively to a throughput score. More generally, the model examines whether an employee can improve a metric while making the underlying service worse. If so, the measure needs a balancing indicator or a narrower use. A performance system should make this test routine because high-stakes metrics can change the behavior that originally made them informative.

Contextual interpretation:

Contextual variables describe opportunity and task difficulty. Examples include caseload complexity, staffing, availability of functioning systems, language support needs, and the employee's share of newly introduced tasks. They are not explanations to add after an unpopular score appears. Material variables and their treatment should be specified prospectively, then reviewed when genuine unforeseen conditions arise. Contextualization should improve comparability without insulating poor performance from scrutiny. A practical starting point is stratification into work conditions that can be explained. Employees handling comparable case types can be compared before more complex statistical adjustments are considered. Regression-based adjustment may become appropriate with sufficient observations, stable definitions, and independent validation. Its assumptions must be inspected, including whether the adjustment removes a genuine performance difference or preserves an unjust assignment practice. An adjusted score should never appear without the original observation and a plain-language account of the adjustment. The opportunity record is particularly important for promotion. Access to acting leadership, specialist training, and high-visibility projects can shape both competence and the evidence available to demonstrate it. The organization therefore monitors who receives these opportunities, how selection occurs, and whether alternative assessment routes exist. If evidence coverage differs systematically by group or location, the problem cannot be solved by improving the ranking formula alone. Assignment and development policies need examination.

Decision rules and no compensatory conditions:

The decision process has four stages. Eligibility establishes lawful minimum requirements and the applicable competition. Evidence sufficiency determines whether the available record can support assessment. Comparative assessment evaluates eligible candidates on disclosed role-related criteria. Authorization produces a decision with reasons and identifies any departure from the analytic recommendation. A candidate who lacks evidence is distinguished from one who demonstrably fails a requirement. This distinction directs the appropriate remedy and prevents missing data from becoming a covert judgment of ability.

Integrity is handled through an essential-condition rule grounded in relevant, substantiated evidence and due process. A complaint or allegation does not become a negative score. An unresolved matter may require a separate procedure according to applicable rules, but the analytics system must not infer guilt. When a legally relevant integrity requirement is demonstrably unmet, numerical strengths elsewhere cannot cancel it. The same no compensatory logic applies to other genuinely essential requirements, such as a mandatory professional authorization. For candidates who satisfy all applicable conditions, a composite can summarize differentiating evidence. Let C denote the competency score, P the contextualized performance score, and D demonstrated development and knowledge-sharing contribution, each on a 0-100 scale. An illustrative rule is $M=0.45C + 0.45P + 0.10D$. The weights are policy choices for the worked example. They are neither estimated from the reference set nor recommended as universal defaults. Each organization must justify weights against the destination role and inspect the consequences of reasonable alternatives. A score is a structured aid to judgment, not an entitlement to appointment. The authorized panel should examine component profiles, uncertainty, and role-specific requirements before making a final comparison. However, discretion cannot mean unlimited departure from the criteria. An override needs a reason linked to the published requirements, the evidence supporting it, and independent review. Repeated overrides in favor of particular sponsors, units, or groups become an audit signal.

Numerical demonstration with synthetic candidates:

The demonstration considers six fictional internal candidates for one supervisory role. All values are invented to expose design properties. Candidates A, B, and C have complete evidence and meet illustrative eligibility conditions. Candidate D has an

unresolved essential-competency evidence gap. Candidate E is assumed, solely for this fictional example, to have failed a properly adjudicated essential integrity requirement. Candidate F meets basic eligibility but has incomplete performance evidence. No record represents an actual person, agency, or observed employment decision.

The baseline rule assigns A score of 81.80, B a score of 82.00, and C a score of 81.35. B has the highest arithmetic value, but the difference between B and A is only 0.20 points. The correct analytic conclusion is a near tie under the stated assumptions. It would be indefensible to describe B as materially superior without examining measurement resolution, role fit, and sensitivity. The constructed values intentionally place the leading candidates close together to demonstrate this issue.

Table 4. Synthetic candidate inputs and decision status

Candidate	C	P	D	Baseline M	Assessment status
A	88	76	80	81.80	Eligible; complete evidence
B	77	87	82	82.00	Eligible; complete evidence
C	83	80	80	81.35	Eligible; complete evidence
D	90	89	85	Withheld	Essential competency evidence incomplete
E	93	92	90	Not ranked	Fictional adjudicated integrity condition unmet
F	84	Missing	78	Withheld	Performance evidence incomplete

Note: All inputs and candidate statuses are hypothetical. C=competency; P=contextualized performance; D=development and knowledge sharing. M=0.45C + 0.45P + 0.10D. Withheld values are not treated as zero.

An alternative competency-priority rule uses weights of 0.60, 0.30, and 0.10. Under that rule, A scores 83.60, B 80.50, and C 81.80. A performance-priority rule with weights 0.30, 0.60, and 0.10 produces 80.00 for A, 83.50 for B, and 80.90 for C. The preferred candidate changes when the relative importance of competence and observed performance changes. This is a mathematical result for the synthetic records, not empirical evidence that promotion panels behave in this way.

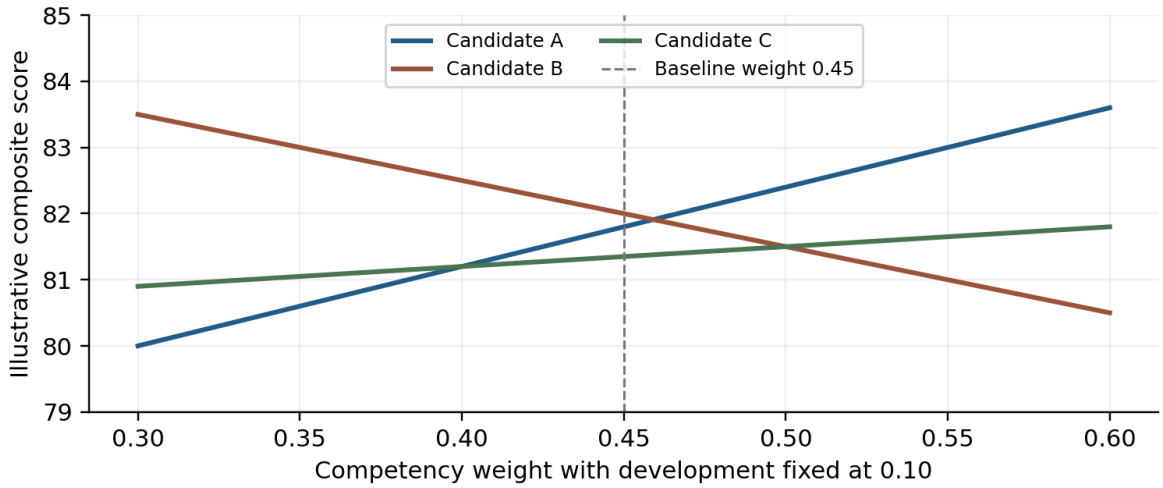


Figure 2. Weight sensitivity in the synthetic example. The performance weight equals 0.90 minus the competency weight. The development weight remains 0.10. Curves are calculated values, not observed data.

The no compensatory conditions change what the table means. D's apparent numerical total is withheld because an essential requirement cannot yet be evaluated. E does not enter comparative ranking because the fictional adjudicated condition is unmet. F requires additional performance evidence or a formally approved alternative assessment. None should receive an invented zero to complete a spreadsheet. Their statuses concern admissibility and sufficiency, whereas the near tie among A, B, and C concerns comparison after admissibility.

Uncertainty and ranking stability:
To illustrate measurement bounds, suppose each of the three complete candidates has an uncertainty allowance of ±3 points for C and P and ±5 points for D. These are stipulated bounds for a stress test, not estimated standard errors or confidence intervals. Under the baseline weights, the composite allowance is 0.45(3) + 0.45(3) + 0.10(5) = 3.20 points. A's bounded range is 78.60–85.00, B's is 78.80-85.20, and C's is 78.15–84.55. All ranges overlap extensively.

The overlap means that the synthetic ranking cannot distinguish the candidates reliably under those allowances. It does not imply that the candidates are identical or that a statistical null hypothesis has been accepted. The practical response is to seek evidence that addresses the destination role's most important unresolved distinction. A structured supervisory exercise could be appropriate if supervisory competence is the uncertainty that matters. Repeating a general interview without a defined purpose would add burden without necessarily improving the decision.

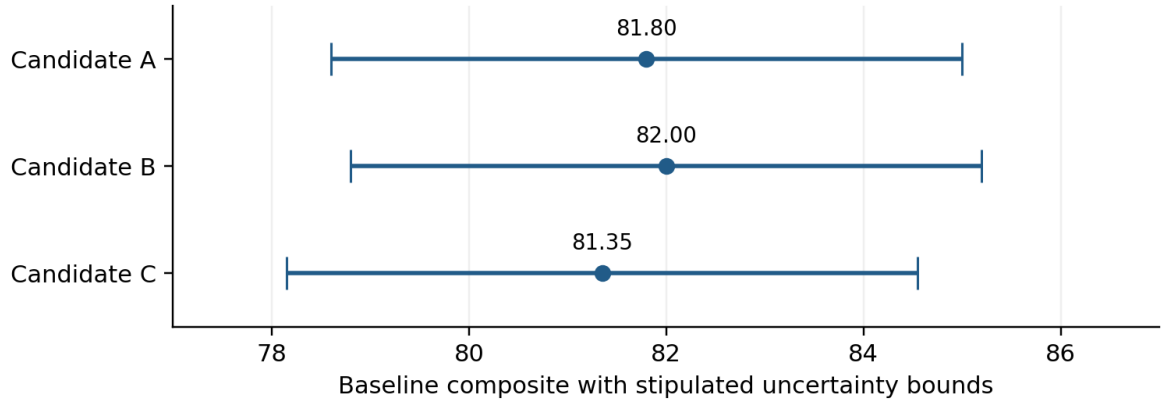


Figure 3. Stipulated bounds for the three complete synthetic records. Bars represent ±3.20 points under the stated stress-test assumptions; they are not statistical confidence intervals.

Organizations should report rank stability across a predeclared set of policy weights rather than searching for weights that favor a preferred person. For this example, a simple continuum can shift the competency weight from 0.30 to 0.60 while keeping D at 0.10 and setting P to the remainder. A and B are equal when the competency weight is approximately 0.4591. Above that value A leads B; below its B leads A. This crossover identifies the policy judgment on which their comparison depends. C remains useful as a distinct balanced profile rather than disappearing into a binary narrative.

Fairness assessment and decision explanations:
Fairness monitoring examines stages rather than only final appointment rates. The organization compares access to assessment, evidence completeness, assessor ratings, eligibility determinations, shortlist decisions, and appointments. A difference at one stage may be amplified or obscured at another. Group-level selection rates should therefore use clearly defined denominators and sufficient counts. Small cells require confidentiality controls and careful interpretation. A descriptive disparity is a reason to investigate; it is not by itself a complete legal finding or proof of its cause.
Bandara et al [8]. Distinguish capabilities for managing biases across data, models, and deployment. Naoum et al [9]. Review the potential benefits and risks of AI for inclusion. These works inform the separation of fairness checks across the system. The proposed design extends those concerns to simple rule-based analytics. An organization can discriminate through a threshold, a missing-data convention, or access to opportunities even when no machine learning is involved. The relevant audit object is the full decision process. An individual explanation should identify the applicable role requirements, evidence used, component assessments, material uncertainty, and the reason for the outcome. It should also identify how factual errors can be corrected and how judgment can be reviewed. Disclosure must protect other candidates' personal information. An unsuccessful candidate may receive their own assessment and an explanation of the selected standard without receiving a competitor's confidential portfolio. An appeal can then focus on a concrete issue rather than a general dissatisfaction with the result.

Governance and accountability:
A merit governance board owns the policy and approves material changes. Operational HR manages competitions and maintains records. Trained assessors evaluate evidence. Data stewards maintain definitions and data quality. An independent review function examines challenges

and exceptional decisions. The appointing authority retains responsibility for lawful authorization. These roles may be combined in a small organization only where conflicts are actively controlled; the same person should not be the sole assessor, final decision-maker, and appellate reviewer for one case. The framework draws governance examples from UK recruitment assurance guidance [13], its Data and AI Ethics Framework [14], and the Algorithmic Transparency Recording Standard scope policy [15]. These are jurisdiction-specific materials, not universally binding rules. Their design relevance lies in documenting purposes, responsibilities, and assurance evidence. Any adopting organization must map these controls to its own legal obligations. Publication should describe how the system works without releasing employee-level information.

Table 5. Accountability allocation in the proposed system

Function	Primary responsibility	Required record	Conflict safeguard
Policy ownership	Merit governance board	Role framework and approved rule versions	Declare interests; publish aggregate changes
Assessment	Trained panel members	Individual ratings and evidence references	Independent initial ratings before discussion
Data quality	Designated data steward	Validation checks and correction log	No silent overwrites of source records
Appointment	Lawful appointing authority	Reasons and any justified override	Independent scrutiny of departures
Review	Reviewer outside the original decision	Findings, remedy, and notification	Recusal where prior involvement exists
Assurance	Internal audit or independent equivalent	Sample-based reconstruction and risk findings	Access to source evidence and rule history
Employee participation	Employee representatives and participating staff	Usability feedback and unresolved concerns	Protected reporting and accessible engagement

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

Governance also requires a record of what the system is not permitted to do. The proposed system does not infer political loyalty, emotional state, or general personal worth. It does not use social network closeness to senior officials as a positive feature, and it does not treat employee silence as consent to unrestricted monitoring. It collects information for identified employment purposes with defined access and retention. These limits should appear in procurement requirements and internal operating procedures so that they remain enforceable when personnel or vendors change.

Data architecture and optional AI:

The minimum data model links an employee or candidate identifier, role version, assessment event, evidence source, score, context, assessor, decision, and review record. Sensitive equality-monitoring information is held separately and accessed for authorized aggregate analysis. The scoring interface should not expose such fields unless a lawful and necessary purpose requires them. Audit records preserve who accessed and changed a consequential field. Retention periods follow applicable requirements rather than an assumption that all personnel information should be stored indefinitely. A practical architecture has a source layer, a validated evidence layer, an analytical layer, and a decision register. Source records remain recoverable. The validated layer records corrections and exclusions. The analytical layer calculates disclosed measures and generates quality checks. The decision register preserves the rules, evidence snapshot, reasons, and authorization used at a particular time. This design allows the organization to reproduce an earlier decision even after role definitions or software have changed. Generative AI may assist with drafting summaries of evidence, but it should not invent observations or create unsupported justifications. The NIST generative AI profile [12] is relevant where such tools are used. Each generated factual statement should be traceable to an authorized source and checked by a responsible human. A summary remains distinct from primary evidence. A reviewer must be able to inspect the original work sample rather than relying on a fluent account produced by a model.

More advanced predictive analytics should be considered only after basic data quality and decision governance are functioning. The OECD government

AI report [20] and Digital Government Outlook [25] identify governance and implementation capacity as relevant conditions for AI use. In the proposed architecture, a predictive model must have a specified target, validation sample, error analysis, and documented use limit. Predicting historical manager ratings is a particularly weak objective when those ratings are themselves under question. Prediction accuracy against a flawed target can reward the reproduction of institutional bias.

Implementation sequence:
Implementation starts with one job family and a limited decision purpose. The organization maps current procedures, identifies sources of favoritism or inconsistency, and defines an improvement objective. It then develops role requirements with employees and service managers, tests assessment tasks, and establishes the review process. A shadow period calculates assessments without affecting careers so that data and interpretation problems can be corrected. Moving directly from policy approval to consequential scoring would make employees bear the cost of design errors. The pilot then uses the framework in a bounded set of authorized decisions, with an accessible alternative process for unresolved technical problems. Independent reviewers examine documentation quality and selected cases. Scale-up depends on acceptable measurement performance, reasonable administrative burden, and functioning correction mechanisms. The calendar depends on organizational complexity; no claim is made that a fixed number of weeks is adequate for every institution. Milestones should be judged by evidence of readiness rather than by completion of software installation.

Table 6. Staged implementation and readiness evidence

Stage	Main activity	Evidence required to proceed	Reason to pause
Scope	Choose one job family and decision purpose	Named owner; legal mapping; defined problem	Authority or review route unclear
Design	Specify roles, tasks, indicators, and evidence	Documented task relevance and accessibility checks	Criteria cannot be observed consistently
Calibration	Train assessors using benchmark cases	Recorded disagreements and revised anchors	Unexplained dependence on assessor identity
Shadow use	Calculate without career consequences	Reproducible records and manageable burden	Missing data materially alter decisions
Bounded pilot	Apply approved rules to limited decisions	Independent case review and working appeals	Serious fairness or integrity concern unresolved
Evaluation	Assess measurement, process, and outcomes	Transparent report including negative findings	Success inferred only from satisfaction or speed
Scale and monitor	Expand with controlled versions	Resources, review capacity, and change governance	Controls weaken as volume increases

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

Development follows every assessment cycle. Selected candidates receive a transition plan for the destination role. Unselected candidates receive targeted feedback and access to development that addresses evidenced gaps. Candidates whose evidence was insufficient should be offered a reasonable route to demonstrate the requirement. The organization also examines whether recurring gaps originate in job design, staffing, or unequal access to assignments. This prevents the system from treating every organizational weakness as an individual deficit.

Prospective validation model:
A future empirical study should test four linked questions: whether measures represent role requirements, whether assessments are sufficiently dependable, whether procedures are experienced as fair, and whether appointments support better service outcomes. These questions require different observations and time horizons. A reliable assessment may still be invalid, and favorable employee perceptions may coexist with poor service quality. The validation plan therefore keeps the outcomes separate rather than collapsing all evidence into a single success score.
Content validity can be examined through structured review by job experts, employees, and service stakeholders, with dissent documented. Dependability requires repeated or multiple

assessments appropriate to the instrument. Criterion evidence should compare assessment results with independently evaluated destination-role performance after appointment, taking account of restriction of range and assignment differences. Fairness analysis examines differential evidence availability, measurement behavior, and decision outcomes. No universal coefficient or cut-off is declared sufficient in advance without reference to decision consequences. A causal evaluation of implementation should, where feasible, compare changes with a suitable counterfactual. Randomized phased introduction across comparable units is one possibility when lawful and practical. Otherwise, a carefully specified comparison design can examine pre-intervention trends, concurrent reforms, and case-mix changes. Simple before-and-after improvement is inadequate because staffing, budgets, leadership, and demand may change simultaneously. Process evaluation should explain whether the intervention was actually used as intended and why particular controls failed or succeeded.

Table 7. Proposed validation questions and study methods

Question	Observation or outcome	Suitable method	Main threat
Content relevance	Coverage of actual destination-role tasks	Structured expert and employee review	Dominant voices exclude less visible work
Dependability	Agreement across assessors and tasks	Repeated ratings; variance decomposition	Agreement mistaken for validity
Criterion evidence	Independent destination-role performance	Prospective follow-up of appointees	Range restriction and assignment differences
Procedural justice	Understanding, voice, and review experience	Anonymous survey plus interviews	Response bias and fear of identification
Distributional effects	Stage-specific access and decision patterns	Disaggregated descriptive and adjusted analyses	Small counts; invalid adjustment assumptions
Service contribution	Quality, timeliness, and accessibility	Phased comparison with process evaluation	Concurrent reforms and changing case mix
Feasibility	Staff time, cost, and appeal burden	Activity records and cost analysis	Unrecorded employee opportunity costs

Note: Proposed design specification developed in this article; it is not a validated standard or an empirical result.

The pilot should be preregistered where appropriate, including outcomes, analysis rules, exclusion criteria, and handling of missing data. A power analysis would need realistic assumptions about clustering, expected effects, and the number of participating units. This article provides no numerical sample-size claim because those inputs are unknown. A pilot can still be useful for feasibility and measurement refinement without being powered to establish a causal effect. Its reporting should state that purpose clearly and avoid converting an exploratory estimate into a definitive policy conclusion.

Discussion

Interpretation of the design results:

The design results suggest that integration requires more than combining a competency score and a performance dashboard. The crucial connection is the sequence through which requirements authorize data collection, evidence supports assessment, and assessment informs a reasoned decision. If any link is absent, an organization can retain the language of merit while allowing discretionary sponsorship to determine outcomes. This interpretation follows from the architecture and stress test. It remains a theoretical proposition requiring organizational evidence. The synthetic example clarifies the limited meaning of numerical precision. B's baseline lead of 0.20 points appears exact because the formula is exact, but the inputs and policy weights do not have equivalent certainty. Changing the competency weight by less than one percentage point from 0.45 to about 0.4591 reverses the comparison between A and B. Reporting many decimal places would obscure this dependence. The proper response is to explain the policy trade-off and examine relevant evidence, rather than presenting a fragile ranking as an objective discovery. The separation of eligibility and comparison is another important result. A compensatory formula treats all weaknesses as exchangeable unless rules prevent it. That may permit an employee's output to offset an essential deficiency. Conversely, coding missing evidence as zero treats uncertainty as demonstrated failure. Both practices simplify software while distorting the

decision. The proposed sequence preserves distinctions that public authorities need to justify, especially where legal qualifications, integrity, and equal access are involved.

Relationship to recent research:

The architecture is consistent with recent attention to competency frameworks as management instruments [2] and with comparative reporting on public employment assessment [4]. It extends these discussions through an explicit link between the competency dictionary, evidence register, and decision record. That extension is a design contribution, not evidence that existing frameworks lack such links everywhere. The framework should be compared with actual institutional practice before any claim of practical novelty is made for a particular jurisdiction. The distinction between data, model, and deployment issues draws on analytics research [8], but the present study locates those issues within public employment authorization. A technically accurate model can still be used for an unauthorized purpose. A fair assessment task can still be inaccessible to some candidates. An independent appeal can still be ineffective if reviewers cannot reconstruct the original evidence. The integration therefore concerns dependencies across technical and administrative components, which should be tested as a system. The design also connects competency management with skills utilization. OECD Skills Outlook 2025 [23] addresses the importance of using and developing capabilities in the workplace. The proposed feedback process makes this relevant to meritocracy: an organization should ask whether employees had a reasonable chance to acquire and demonstrate the competencies it rewards. This does not remove expectations of individual performance. It makes the conditions of assessment visible and allows managers to address structural barriers alongside personal development needs.

Public value and the risk of narrow optimization:

A merit system may unintentionally favor work that is easily counted. Transaction counts can be extracted cheaply, whereas sound policy advice, preventive work, and respectful handling of vulnerable service users require more difficult assessment. If data convenience determines criteria, the organization may select candidates who are effective at the visible part of the job and overlook necessary but less measurable contributions. The proposed indicator specification responds by beginning with the mandate and testing foreseeable gaming behavior before assigning weights.

Human-centered service innovation [17] and the UN E-Government Survey [16] provide broader context for this concern. Digital transformation is relevant because public service quality depends on organizational capabilities and access, not simply the adoption of tools. The present design does not claim that a meritocracy system will improve national digital government rankings. Instead, it identifies an internal mechanism through which workforce assessment could support service priorities if role requirements accurately reflect them. Some outcomes are too distant from individual control to support high-stakes scoring. Public trust, policy compliance, and social inclusion reflect many influences beyond one employee's work. They are useful at the program or organizational level, but attaching them directly to individual promotion can create unjust attribution. The framework uses a chain of contribution: observable work supports intermediate service results, which may contribute to wider outcomes. Each link needs examination. Greater social importance does not make an outcome more suitable for individual appraisal.

Political neutrality and accountable responsiveness:

Public organizations must respond to lawful policy direction while maintaining professional integrity. A meritocracy system should distinguish disagreement about policy from evidence of poor role performance. Employees may be required to provide accurate analysis, implement lawful instructions, and explain feasibility. Their promotion should not depend on personal affinity with a political sponsor or on suppressing inconvenient evidence. The relevant boundary depends on the role and jurisdiction, which is why the architecture requires explicit criteria and legal mapping.

Anti-corruption and integrity reporting [19] supports attention to the relationship between formal rules and actual implementation. The proposed system addresses this issue through retained evidence, conflict declarations, and review of overrides. None can guarantee insulation from political pressure. Senior officials may influence role definitions, appointment timing, or access to development before an assessment begins. Audits must therefore look upstream of final scores. A formally independent panel can still operate within a competition whose requirements were tailored to a preferred candidate. The system should record material changes to a role during an active competition and require justification for them. Where requirements change for legitimate operational reasons, affected candidates should receive notice and a fair opportunity to respond. Changes that alter the eligible pool should receive heightened independent scrutiny. This is a governance proposition derived from the threat model, not a claim that all adjustments are improper. Administrative flexibility remains necessary, but its use must be visible.

Fairness without mechanical parity:
Group comparisons can reveal exclusion, but a fair system cannot be reduced to equal average scores or identical appointment rates. Differences can arise from unequal opportunity, irrelevant criteria, measurement errors, sample variation, or genuine differences in role-relevant evidence. Distinguishing these explanations requires process records and substantive analysis. A dashboard that colors a disparity green or red without context risks both false reassurance and unjustified accusation. The proposed audit investigates causes rather than treating a summary statistic as a verdict. The ICO's 2024 intervention on recruitment tools [24] supplies a practical example of regulatory attention to data protection in employment analytics. Its relevance here is the need for transparency and careful handling of personal information. The design does not infer that UK regulatory conclusions automatically apply elsewhere. An organization should identify the lawful basis and purpose for each data element, establish correction routes, and determine which equality analyses can be conducted under its applicable rules. A second fairness issue concerns the meaning of developmental contribution. Giving points for mentoring or knowledge sharing may recognize collective work, but opportunities can depend on grade, workload, and managerial invitation. A compulsory high score in this component could disadvantage new employees or staff in understaffed units. The illustrative D component should therefore be adapted to the decision and supported by accessible opportunities. The example's ten-percent weight demonstrates aggregation; it does not establish that this component belongs in every competition.

Privacy and employee voice:
The framework favors purpose-limited evidence over continuous surveillance. Rani et al [11]. describe algorithmic management in conventional workplaces and its implications for work organization. The design implication is to assess whether data collection changes the employment relationship as well as whether it improves measurement. A metric based on constant activity tracking can encourage visible busyness, reduce autonomy, or expose sensitive personal patterns. Such effects should be considered before deployment rather than treated as incidental consequences of a software purchase.
Employees should participate in defining assessment tasks, reviewing explanations, and identifying burdens. Participation does not give any group unrestricted authority to choose favorable criteria. Its purpose is to test whether the system represents actual work and can be used without avoidable harm. Management remains accountable for policy choices and must explain unresolved disagreements. Consultation is most useful when it occurs before key design decisions are locked into procurement or when changes remain feasible. The system also needs a protected channel for reporting manipulation. An employee may observe pressure to inflate scores, delete adverse evidence, or provide privileged access to a work sample. Such reports require handling outside the ordinary supervisor chain where conflicts exist. Aggregate reporting should distinguish substantiated manipulation from unverified allegations. Otherwise, transparency can itself become a means of reputational harm. The design must protect both the integrity of the process and the rights of those named in a complaint.

Administrative feasibility and cost:
An elaborate framework can consume the capacity it is intended to improve. Work samples require assessors, data checks require staff time, and appeals require independent expertise. Small organizations may have too few specialists to separate every role internally. The appropriate response is proportionate implementation, shared assessment services, or external independent review were lawful. Automation can reduce clerical effort, but it does not eliminate the need for interpretation, calibration, and correction. The economic evaluation should count the time of employees and managers, not only software licensing. It should also examine opportunity costs when assessment activities interrupt services. Potential benefits include fewer avoidable appointment disputes, better role matching, and more focused development. These remain hypotheses until measured. A business case should present assumptions and sensitivity analyses rather than claiming savings from generic AI productivity estimates. Costs are more readily observed early; service and capability benefits may require longer follow-up. A minimum viable implementation can use standardized role documents, structured assessment sheets, a controlled evidence register, and reproducible calculations. Its limitations should be explicit, including manual access controls and capacity constraints. A sophisticated platform becomes worthwhile when scale, integration, or audit requirements justify it. The conceptual model therefore permits a spreadsheet-based pilot but requires the same substantive rules about evidence, authorization, and review. Technical simplicity does not excuse weak governance.

Institutional transfer and contextual adaptation:
The proposed architecture is intended for public organizations but is not a universal civil service reform package. Administrative traditions differ in the balance between career and position-based employment, central and agency control, and statutory protections. A local authority with a small workforce faces different evidence and anonymity challenges from a national ministry. Adaptation

should begin by identifying which decisions the organization can lawfully redesign and which are constrained by central rules. Contextual adaptation should preserve the distinction between core principles and adjustable parameters. Relevance, evidence sufficiency, reasoned authorization, and review are core design commitments. Competency domains, weights, assessment windows, and review arrangements are parameters requiring local justification. Treating every feature as fixed would create needless rigidity. Treating every feature as optional would allow the system's safeguards to disappear. An adaptation record should state what was changed, why it was changed, and how the consequences will be evaluated. The source set is weighted toward English-language and international institutional materials. This supports a broad design discussion but limits sensitivity to jurisdiction-specific practices and scholarship in other languages. Future applications should add local legal materials, validated instruments, and research on administrative culture. Such additions may identify requirements that alter the decision architecture rather than merely translate its terminology. Transfer should be investigated as a research question rather than assumed from common managerial vocabulary.

Limitations:

The principal limitation is the absence of field validation. The study establishes an internally described framework and demonstrates arithmetic consequences of its rules. It does not show that the framework increases performance, reduces favoritism, improves diversity, or strengthens trust. No empirical effect size can be inferred from the numerical example. The example was designed to illustrate ambiguity and therefore should not be used to estimate how often rankings would be unstable in real competitions. The targeted synthesis is another limitation. Sources were selected for design relevance through web and publisher searches, without exhaustive database retrieval or independent duplicate screening. The resulting bibliography is not a representative sample of all research on meritocracy. A formal systematic review could refine the evidence base and identify contradictory findings more completely. The present article consequently avoids counting thematic recurrence as proof of strength and separates normative controls from causal evidence. The framework also simplifies practical measurement problems. Converting competency evidence to a 0–100 score can imply interval properties that behavioral ratings may not support. The weighted example assumes commensurable component scales solely to make the calculation transparent. An operational system must test scaling assumptions or use ordinal decision rules where appropriate. Likewise, the bounded uncertainty exercise uses stipulated values and does not estimate measurement error, dependence between components, or the distribution of unknown outcomes. Finally, governance controls can fail through informal practices. Written reasons may be produced after a decision, reviewers may lack independence, and appeals may be too burdensome to use. These risks cannot be removed by adding fields to a database. Future evaluation must examine actual behavior, including whether people can challenge a decision without retaliation. The proposed controls are conditions for investigation and implementation, not proof that a meritocratic culture has been achieved.

Future research propositions:

Four propositions follow from the design and should be tested without assuming positive results. First, role-specific behavioral anchors combined with assessor calibration should improve agreement relative to unstructured global ratings. Second, transparent contextualization should reduce differences attributable to task allocation, provided adjustment variables are valid. Third, accessible explanations and independent review should improve perceived procedural justice, even when an individual is not selected. Fourth, monitoring developmental opportunities should reveal upstream barriers that appointment-rate audits alone miss. These propositions can be examined through complementary studies. Measurement studies can compare assessment formats within the same job family. Vignette experiments can isolate explanation and review features while recognizing their limited ecological validity. Longitudinal organizational studies can examine changes in opportunity access and appointment outcomes. A phased implementation can test service and workforce effects when an appropriate comparison is available. Qualitative work should examine how managers and employees interpret criteria, including strategies for gaming and informal resistance.

Research should also compare alternative decision models. A fully compensatory index, a threshold-plus-ranking model, and a structured panel without a final composite may produce different decisions and explanations. The relevant outcomes include validity, fairness, burden, and vulnerability to manipulation. There is no reason to assume the most quantitative model will perform best. Comparative evaluation can identify when aggregation helps consistency and when it masks consequential differences between candidates.

Conclusion

A defensible meritocracy system in a public organization requires an explicit connection between role requirements, evidence, performance interpretation, and accountable employment decisions. This article proposes a competency and analytics architecture that separates eligibility,

evidence sufficiency, comparative assessment, and authorization. Essential requirements remain no compensatory, contextual conditions are visible, and decisions produce reasons that can be independently reviewed. Development and opportunity allocation form part of the system because they shape what employees can later demonstrate. The synthetic demonstration shows that a precise composite can conceal a fragile comparison. The leading candidate changes under plausible alternative weights, and stipulated uncertainty bounds overlap across the complete records. These analytical results support using scores as transparent decision aids with explicit limits. They do not establish organizational effectiveness. The next step is a staged, independently evaluated pilot that tests measurement quality, procedural fairness, service contribution, and administrative feasibility before broad consequential use.

Disclosure Statement

No potential conflict of interest reported by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to data analysis, drafting, and revising of the paper and agreed to be responsible for all the aspects of this work.

References

[1] Oliveira, E., Abner, G., Lee, S., Suzuki, K., Hur, H., & Perry, J. L. (**2024**). What does the evidence tell us about merit principles and government performance? *Public Administration*, 102(2), 668-690.

[2] Goulden, A. (**2025**). Capability Building, Management and Reform: An Initial Review of Public Sector Competency Frameworks. UCL *Institute for Innovation and Public Purpose Working Paper* 2025-15.

[3] Burtscher, M., Piano, S., & Welby, B. (**2024**). Developing skills for digital government: A review of good practices across OECD governments. *OECD Social, Employment and Migration Working Papers*, 303. OECD Publishing.

[4] OECD. (**2025**). Government at a Glance 2025. *OECD Publishing.*

[5] OECD. (**2025**). Workforce Insights from Central Governments: Findings of the 2024 OECD/EU Survey of Public Servants. *OECD Publishing.*

[6] Wightman, G. B., & Christensen, R. K. (**2025**). A systematic review of person–environment fit in the public sector: Theorizing a multidimensional model. *Public Administration Review*, 85(2), 386-401.

[7] Alam, S., Dong, Z., Kularatne, I., & Rashid, M. S. (**2026**). Exploring approaches to overcome challenges in adopting human resource analytics through stakeholder engagement. *Management Review Quarterly*, 76, 729-787. Published online in 2025.

[8] Bandara, R. J., Biswas, K., Akter, S., Shafique, S., & Rahman, M. (**2025**). Addressing algorithmic bias in AI-driven HRM systems: Implications for strategic HRM effectiveness. *Human Resource Management Journal*, 35, 1047-1063.

[9] Naoum, R. F., Szakadáti, T., & Balogh, G. (**2026**). Artificial Intelligence (AI) in human resource management (HRM): A systematic review of its dual impact on diversity, equity, and inclusion (DEI). *Management Review Quarterly.*

[10] Milanez, A., Lemmens, A., & Ruggiu, C. (**2025**). Algorithmic management in the workplace: New evidence from an OECD employer survey. *OECD Artificial Intelligence Papers, 31. OECD Publishing.*

[11] Rani, U., Pesole, A., & González Vázquez, I. (**2024**). Algorithmic Management Practices in Regular Workplaces: Case Studies in Logistics and Healthcare. *Publications Office of the European Union.*

[12] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (**2024**). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. *National Institute of Standards and Technology.*

[13] Department for Science, Innovation and Technology. (**2024**). Responsible AI in Recruitment. UK Government.

[14] Department for Science, Innovation and Technology. (**2025**). Data and AI Ethics Framework. UK Government. *Updated* 18 December 2025.

[15] Government Digital Service. (**2024**). Algorithmic Transparency Recording Standard Mandatory Scope and Exemptions Policy. *UK Government.*

[16] United Nations Department of Economic and Social Affairs. (**2024**). United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development. *United Nations.*

[17] OECD. (**2024**). Global Trends in Government Innovation 2024: Fostering Human-Centred Public Services. *OECD Publishing.*

[18] OECD. (**2024**). OECD Survey on Drivers of Trust in Public Institutions 2024 Results: Building Trust in a Complex Policy Environment. *OECD Publishing.*

[19] OECD. (**2024**). Anti-Corruption and Integrity Outlook 2024. *OECD Publishing.*
[20] OECD. (**2025**). Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. *OECD Publishing*.
[21] OECD. (**2026**). Building an AI-ready public workforce: Implications and strategies. *OECD Publishing.*
[22] OECD. (**2026**). A Skills-First Labour Market. *Getting Skills Right. OECD Publishing.*
[23] OECD. (**2025**). OECD Skills Outlook 2025. *OECD Publishing.*
[24] Information Commissioner's Office. (**2024**). ICO intervention into AI recruitment tools leads to better data protection for job seekers. 6 *November* 2024.
[25] OECD. (**2026**). Digital Government Outlook 2026: From Foundations to Transformational Impact. *OECD Publishing.*