Intelligent Control and Automation, 2026, 17(4), 212-224
https://www.scirp.org/journal/ica
ISSN Online: 2153-0661
ISSN Print: 2153-0653

# Where Can Artificial Intelligence Support Risk Prioritization in an Integrated Management System of a Testing Laboratory and a Certification Body? A Perspective and Validation Protocol

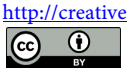
**Karina Rafailovna Zagidullina**

Tashkent State Technical University Named after Islam Karimov, Tashkent, Uzbekistan
Email: zkarina99r@gmail.com

**How to cite this paper:** Zagidullina, K.R. (2026) Where Can Artificial Intelligence Support Risk Prioritization in an Integrated Management System of a Testing Laboratory and a Certification Body? A Perspective and Validation Protocol. *Intelligent Control and Automation*, **17**, 212-224.
https://doi.org/10.4236/ica.2026.174010

**Received:** October 26, 2026
**Accepted:** November 3, 2026
**Published:** November 6, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

## Abstract

An organization that combines a testing laboratory and a product certification body must rank risks of different nature in one register. In earlier work, the author proposed a weighted risk priority index and showed by Monte Carlo simulation that its ranking is more sensitive to errors in expert probability and consequence scores than to uncertainty in the weights. This perspective paper asks where artificial intelligence (AI) could reduce that score uncertainty without replacing expert judgment. Three advisory functions are discussed: evidence for probability and consequence scores drawn from nonconformity and audit records, early-warning signals from key performance indicator (KPI) trends, and mapping of changed requirements to risks. Conditions for use are formulated from the literature on AI in risk management, laboratory quality control and AI risk governance. A simulation on the author's register of 19 risks then quantifies how accurate score support would need to be: halving the nominal share of scores that are off by one point (from 0.10 to 0.05 per direction; realized shares are lower because scores are limited to 1 - 5) raised the mean rank correlation with the reference ranking from 0.88 to 0.94 and the retention of the four leading risks from 37% to 62%, and improving only the laboratory risks helped the leading risks less than improving the impartiality and management risks. A retrospective validation protocol with pre-specified acceptance criteria is proposed. No AI method was implemented; the simulation does not evaluate any AI model, and the functions proposed are untested. The register, the random seed and the simulation script are given in the appendices.

**Keywords**

Artificial Intelligence, Risk Assessment, Risk Prioritization, Testing Laboratory, Certification Body, Integrated Management System, Validation

## 1. Introduction

Organizations that test products and certify them themselves face requirements of two different kinds. The adoption of ISO 9001 [1] is voluntary and defines a general quality management system. The requirements that apply specifically to accredited testing laboratories and to product certification bodies are ISO/IEC 17025 [2] and ISO/IEC 17065 [3]; they set the conditions for accreditation, including the treatment of risks and, for certification bodies, of impartiality. Each standard asks for risk-based thinking, the general risk management process is described in ISO 31000 [4], but the risks differ: business risks, risks to the validity of results and risks to impartiality. A single register therefore has to rank risks of different nature on one scale. The ICH Q9(R1) guideline [5] is a sector-specific document on quality risk management in the pharmaceutical industry; it does not govern testing and certification organizations in general. It is cited only because it states explicitly that subjectivity in risk assessment must be managed and minimized, a concern that applies equally to the scoring of risks in other sectors.

In practice, a risk is still often scored as the product of ordinal probability and consequence scores. Such products have coarse resolution and make different risks look equal [6]. Weighting by risk type and process criticality moves part of the subjectivity from scores to weights [7]. In a recent robustness study of such a weighted index on a register of 19 risks [8], the author found that the ranking was stable under moderate disagreement about weights (mean Spearman correlation 0.95 for ±20% weight deviation, against 0.45 for random weights), but was affected more by errors in the scores: when 20% of scores were shifted by one point, the correlation fell to 0.88. The study concluded that effort should be directed to the quality of the scores, for example, calibrated scales, independent assessors and data from the nonconformity log. The data came from one organization.

This paper builds on that conclusion. Its question is narrow: where could AI support the production of better scores and earlier signals, under what conditions, and how should such support be validated before it is relied upon? The paper reports no AI experiment. It adds one computation, a simulation on the author's register that shows how accurate score support would have to be to change the ranking noticeably, followed by a validation protocol. The functions proposed are untested.

## 2. Background

### 2.1. AI in Risk Management

A systematic review of 84 peer-reviewed articles on AI in construction project risk

management [9] found that machine learning is used mainly for predictive modelling, natural language processing supports document and compliance risk, knowledge-based reasoning supports decisions, and optimization algorithms help allocate resources. The authors list data quality, model interpretability and workforce skills as the main barriers to integration. The review concerns construction, not laboratories, but its barriers are generic, and they apply to the register discussed here, which is small (19 risks) and assembled by a single organization.

### 2.2. AI in Laboratory Quality Control

In laboratory practice, machine learning has been applied to patient-based real-time quality control to detect analytical and preanalytical errors [10]. For proficiency testing and method validation, a Python-based framework that uses machine learning for outlier detection and performance scoring has been presented to automate ISO 13528 and ISO 5725-2 procedures [11]. These works address measurement data, not management-system risks, but they show that statistical learning can be attached to existing quality procedures. For the risk side, the use of FMEA in an accredited testing laboratory has been described [12]; it relies on expert scoring and therefore on the uncertainty discussed above.

### 2.3. Governance of AI Itself

Using AI inside a management system creates its own risks. ISO/IEC 23894 [13] and the NIST AI Risk Management Framework [14] provide guidance on managing risks of AI systems. A technical report proposes a quality management system design for AI systems based on the EU Artificial Intelligence Act and a prototype of a risk management service [15]. Explainability of models is a recurring requirement and has been surveyed in [16]. For a laboratory, ISO/IEC 17025 [2] additionally requires control of data and information management, which includes validation of information systems.

## 3. Baseline Risk Model

The baseline is the weighted index $R = P \times C \times Wt \times Wp$ [7], where P and C are probability and consequence scores, Wt is the weight of the risk type and Wp the weight of the criticality of the process in which the risk arises. The register contains laboratory technical risks, impartiality risks of the certification body and management risks. The index separates risks that the unweighted product $P \times C$ scores equally and, in the cited study, classified nine of 19 risks as high [8]. In the integrated system, the register is connected to corrective actions and to the KPIs of the system [17]. The baseline is deliberately kept unchanged here; AI would act on its inputs, not on its structure.

## 4. Three Candidate Advisory Functions

Figure 1 shows the proposed arrangement. In all three functions, the output is advisory: a human assessor decides the final scores and actions, and the AI output

is validated and logged.

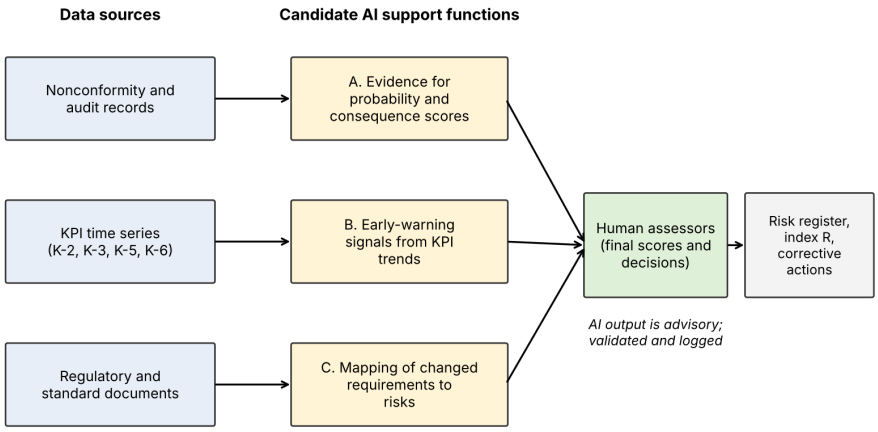



**Figure 1.** Proposed advisory use of AI in risk prioritization (conceptual; not implemented).

*Function A*: *evidence for probability and consequence scores*. Probability and consequence scores are currently assigned by experts. A model trained on the organization's nonconformity and audit records could show, for each risk, how often related events occurred and how serious they were, so that the assessor works with evidence and a documented range rather than an impression. This addresses directly the instability shown in [8]. It requires that records be coded to the same risk categories as the register; otherwise, the evidence cannot be linked to risks.

*Function B*: *early-warning signals from KPI trends*. The KPI system of the integrated management system includes indicators that are calculated from process data (for example metrological traceability, report issue time and share of decisions made on time). Anomaly detection on these time series, similar in idea to patient-based real-time quality control [10], could raise a signal before a quarterly or annual review. The signal would prompt a reassessment of the related risk; it would not change a score automatically.

*Function C*: *mapping of changed requirements to risks*. The register of the studied organization contains a risk of nonconformity with accreditation requirements caused by changes in the normative base. Natural language processing could compare new or revised standards and regulations with the organization's documents and propose which processes and risks are affected, which is the kind of document and compliance support described in [9]. A person would check every proposed link.

*Functions not proposed*. AI is not proposed for making certification decisions or for judging impartiality of persons. These decisions require competent persons who can be held accountable, and the risk of misjudgment from a model that cannot explain itself [16] would itself be a threat to the impartiality that the integrated system must protect [17].

## 5. Conditions for Use

Four conditions follow from the literature and from the standards cited above.

First, data: the barriers named in [9] apply, and records must be complete, coded to risk categories and kept separately for the laboratory and the certification body, because a model trained on pooled data could carry information across the boundary between the two activities. Second, validation: an AI function used in the system is an information system whose fitness has to be demonstrated before use [2]. Third, risk management of the AI function itself, using the approach of [13] and [14], including logging of inputs and outputs and a defined fallback to the manual procedure. Fourth, explainability: an assessor must be able to see why the model suggests a value [16].

## 6. How Accurate Must Score Support Be? A Simulation on the Register

*Design*. The question can be partly answered without any AI model by asking how much the ranking improves if the share of wrong scores falls. The register of 19 risks of the earlier study [8] was taken as the reference. Following the scenario of score errors in that study, each probability score P and each consequence score C was shifted down by one point with probability q and up by one point with probability q (within the range 1 - 5), so that a nominal share 2q of the scores is shifted in a run. Scores were limited to the range 1 - 5: a shift that would leave the range was not applied, so the score stayed unchanged. The realized share of altered scores is therefore lower than 2q (Table 1); the values of q in this paper are nominal shift probabilities, not realized error rates. For each q, 20,000 runs with a fixed random seed (7) were made with the weights unchanged. In each run, the index R was recalculated and compared with the reference ranking by the Spearman rank correlation, the probability that the same four risks stay the highest, and the share of risks that change their level (low, medium, high, critical), with the cutoffs low $R < 3$, medium $3 \le R < 9$, high $9 \le R \le 36$ and critical $R > 36$. Ties were handled as follows: the Spearman correlation uses average ranks, and the four leading risks are selected by a stable sort, so that equal values keep the register order.

Table 1. Effect of the share of wrong scores on the ranking (register of 19 risks, 20,000 runs per row, seed 7).

| Share of scores shifted by one point | Mean rank correlation | 5th percentile | Four leading risks retained, % | Risks changing level, % | Realized altered scores, % |
|---|---|---|---|---|---|
| q = 0.25 (all risks) | 0.74 | 0.56 | 9 | 28.2 | 44.1 |
| q = 0.15 (all risks) | 0.83 | 0.68 | 23 | 18.4 | 26.5 |
| q = 0.10 (all risks) | 0.88 | 0.76 | 37 | 12.7 | 17.6 |
| q = 0.05 (all risks) | 0.94 | 0.84 | 62 | 6.5 | 8.8 |
| q = 0.025 (all risks) | 0.97 | 0.90 | 79 | 3.3 | 4.4 |
| q = 0.05 for the 8 laboratory risks, 0.10 for the other 11 | 0.91 | 0.80 | 43 | 9.6 | 13.9 |
| q = 0.05 for the 11 impartiality and management risks, 0.10 for the 8 laboratory risks | 0.91 | 0.80 | 54 | 9.6 | 12.5 |

*Check*. The simulation reproduces the earlier results: mean rank correlation 0.88 at q = 0.10 and 0.74 at q = 0.25 [8].

*Reproducibility*. The 19-risk register (P, C, Wt, Wp; Table A1), the cutoffs, the tie rules, the seed and the complete Python script (Appendix B; Python 3, NumPy 2.5.3, SciPy 1.18.1) are given in the appendices, so the results of Table 1 can be regenerated independently. With seed 7 the script reproduces every value of Table 1; the last column of Table 1 gives the realized share of altered scores (the mean over the 20,000 runs).

*Results*. Halving the error rate from q = 0.10 to q = 0.05 raised the mean correlation from 0.88 to 0.94, the retention of the four leading risks from 37% to 62%, and cut the share of risks that change level from 12.7% to 6.5%. In terms of realized altered scores, these two rows correspond to 17.6% and 8.8% of all scores. Reaching a mean correlation of 0.97 needs q around 0.025. The gain is therefore not small, but it requires score errors to be reduced by half or more, not by a few percent. The last two rows of Table 1 give a second result: improving only the eight laboratory risks and improving only the eleven impartiality and management risks give the same mean correlation (0.91), but the second keeps the four leading risks much more often (54% against 43%), because three of the four leading risks in the register are impartiality or management risks.

*Interpretation and limits*. The result is a requirement, not a performance: it says how accurate AI-supported scores would have to be, and it does not say that any AI model can reach this accuracy. The errors were modelled as independent random shifts of one point around the register scores, which were treated as correct; real errors may be systematic, correlated between risks, or larger than one point, and the register scores themselves are expert judgments. The second result suggests a practical problem: data-driven evidence is most likely to exist for frequent laboratory events, while the leading risks of the register are impartiality and management risks for which records are probably scarce. If this is true, Function A would help most where it is least likely to have data, and the benefit of AI support in this register would be smaller than the first rows of Table 1 suggest. This is an inference from the structure of the register and has to be tested on real records.

## 7. Proposed Validation Protocol

The author's previous study [8] used rank correlation with a baseline ranking and the share of simulation runs in which a risk stays at the high level. The same measures can be used to validate AI support. The protocol is retrospective and uses the organization's register and its records of events (nonconformities, complaints, impartiality incidents, audit findings and results of proficiency testing). Its steps are: 1) fix a scoring date and build the baseline ranking from expert scores alone, using only information available on that date (Section 7.1); 2) produce a second ranking in which the expert scores are replaced, or supplemented, by the AI evidence of Function A, again using only information available on that date; 3) build the reference ranking from the events recorded after the scoring date (Sec-

tion 7.2) and compare both rankings with it, not with each other; 4) report the measures of Section 7.4.

### 7.1. Chronological Separation

A scoring date t0 is fixed, for example the date of a past annual review. The expert-only assessment and the AI-supported assessment may use only information available at t0: the register as it was, records dated before t0 and documents in force at t0. Models are trained and tuned only on records dated before t0. Records dated after t0 are kept closed and are used only to build the reference ranking in the outcome evaluation, so that no information from the observation period can reach either assessment. Assessors who produce the expert-only scores must not see the AI output or the later records. If the data allow, several scoring dates may be used, each with its own observation period, provided that the observation periods do not supply records that are used for scoring at a later date in the same comparison.

### 7.2. Reference Ranking from Later Events

The reference ranking is built from events recorded in an observation period after t0. The length of the period is fixed in advance and should not be shorter than one full audit cycle of the organization (for example 12 months). Each event is linked to a risk of the register independently by two persons; disagreements are resolved by a third. For each risk, the observed loss is the sum, over its events, of a severity score on the same 1 - 5 scale as the consequence score C, assigned from the event record by a person who does not see the rankings. The loss is divided by the exposure of the period, defined as the volume of the activity in which the risk arises (for example the number of tests performed, reports issued or certification decisions taken), so that risks attached to large and small activities are comparable. Risks are ranked by this exposure-adjusted, severity-weighted event rate; risks without events share the lowest rank, and average ranks are used in all correlations. Frequency (events per unit of exposure) and mean severity are also reported separately, because they may rank the risks differently.

### 7.3. Suppression of Events by Preventive Action

Observed events are not a clean measure of the underlying risk. A risk that was correctly ranked high at t0 may receive preventive or corrective action and then produce few events, so that a correct prioritization could look like a wrong one. The protocol therefore 1) records every preventive and corrective action taken after t0 with its date and the risk it addresses; 2) compares the rankings separately for risks with and without a recorded action, and does not count a risk with an action as a false alarm because it shows few events; 3) counts detected nonconformities, near misses and out-of-control quality control results as events, together with failures that reached the customer, because they remain observable when prevention works; and 4) repeats the analysis without the risks that received

an action as a sensitivity check. The effect cannot be removed completely and is listed among the limitations.

### 7.4. Measures

*Ranking*. For the ranking as a whole, the rank correlation of each assessment with the reference ranking, the retention of the four leading risks and the share of risks that change level are reported for the expert-only and the AI-supported assessment at the same t0 and for the same risks.

*Function B*. The measures are detected and missed events and the lead time before the quarterly review. In addition, the false-alarm burden is reported: the number of signals per month, the share of signals not followed by a relevant event within a window fixed in advance, and the assessor time spent on checking them. A function that detects every event but floods assessors with signals is not useful, and the burden shows this.

*Function C*. The share of proposed links accepted by assessors measures precision only and cannot show which requirements were omitted. Recall is therefore measured against a reference set: for a sample of revisions of standards, accreditation criteria and regulations, two competent persons independently mark all processes and risks that are affected, without seeing the AI output, and the reference set is their reconciled union. The proposed links are compared with this set, and the missed relevant links, the recall and the precision are reported.

### 7.5. Pre-Specification

To avoid choosing criteria after seeing the results, the acceptance criteria and analysis plan should be written down before the data are analyzed, following the logic of preregistration [18]. The criteria themselves, for example how much agreement with later events counts as an improvement over expert-only scores, must be set by the organization and cannot be taken from this paper; no threshold is claimed here. The observation period, the exposure measure, the severity scoring, the false-alarm window and the sample of revised requirements belong to the criteria that must be fixed in advance.

## 8. Limitations

The paper has important limitations. No AI method was built and no AI result is reported; the simulation of Section 6 evaluates the ranking, not a model, and rests on assumed random errors and on the register scores taken as correct. The baseline model has been tested in one organization with a register of 19 risks, which is small for statistical learning, and the number of past events available for training is probably small as well. The literature reviewed comes from other fields, construction and clinical laboratories, and its transfer to testing and certification organizations is an assumption. The validation protocol has not been run, and the suggestion that records are scarce for the leading risks has not been checked on real data. The reference ranking built from later events is affected by preventive

actions that suppress events (Section 7.3), and this effect can be reduced but not removed. The values of q are nominal shift probabilities; the realized share of altered scores is lower because of the score limits.

## 9. Conclusion

The author's previous robustness study suggests that the weakest point of the weighted risk index is the quality of expert scores. Three advisory AI functions are proposed to address it: evidence for scores, early-warning from KPI trends, and mapping of changed requirements to risks, all with a human assessor retaining the decision. The conditions for use are data quality, validation, risk management of the AI function and explainability. A simulation on the register shows that halving the nominal share of shifted scores would raise the rank correlation from 0.88 to 0.94 and the retention of the four leading risks from 37% to 62%, and that the benefit depends on which risks are improved. A retrospective validation protocol with pre-specified criteria is proposed. The next step is to collect coded records and to run the protocol on them; until that is done, the proposals remain hypotheses.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

[1] ISO 9001:2015 (2015) Quality Management Systems-Requirements. ISO. https://www.iso.org/standard/62085.html

[2] ISO/IEC 17025:2017 (2017) General Requirements for the Competence of Testing and Calibration Laboratories. ISO. https://www.iso.org/standard/66912.html

[3] ISO/IEC 17065:2012 (2012) Conformity Assessment-Requirements for Bodies Certifying Products, Processes and Services. ISO. https://www.iso.org/standard/46568.html

[4] ISO 31000:2018 (2018) Risk Management-Guidelines. ISO. https://www.iso.org/standard/65694.html

[5] ICH (2023) ICH Harmonised Guideline Q9(R1): Quality Risk Management. Step 4. International Council for Harmonisation. https://www.ich.org/page/quality-guidelines

[6] Hubbard, D.W. (2020) The Failure of Risk Management: Why It's Broken and How to Fix It. 2nd Edition, Wiley. https://doi.org/10.1002/9781119521914

[7] Zagidullina, K.R. (2026) Improving the Efficiency of Testing Laboratories on the Basis of an Extended Risk-Oriented Model with Weight Coefficients. *Muhandislik va Iqtisodiyot*, No. 4, 400-404. (In Russian)

[8] Zagidullina, K.R. (2026) Weighted Risk Prioritization Index for Organizations Combining Testing and Certification: Robustness to Weight Uncertainty and Retrospective Validation. *Bulletin of TUIT*: *Management and Communication Technologies*, **3**, 580-595. (In Russian)

[9] Tian, K., Zhu, Z., Mbachu, J., Ghanbaripour, A. and Moorhead, M. (2025) Artificial Intelligence in Risk Management within the Realm of Construction Projects: A Biblio-

metric Analysis and Systematic Literature Review. *Journal of Innovation & Knowledge*, **10**, Article 100711. https://doi.org/10.1016/j.jik.2025.100711

[10] Lorde, N., Mahapatra, S. and Kalaria, T. (2024) Machine Learning for Patient-Based Real-Time Quality Control (PBRTQC), Analytical and Preanalytical Error Detection in Clinical Laboratory. *Diagnostics*, **14**, Article 1808. https://doi.org/10.3390/diagnostics14161808

[11] AlMuhlaki, F.A., Bin Qoud, S.A., AlYousefi, R.A., AlFaleh, I., Qahtani, N., AlEnizi, K., *et al.* (2025) SASO Proficiency Test Machine-Advanced Pythonic AI Algorithms for Automating and Validating ISO 13528 and ISO 5725-2 at Saudi Standards, Metrology, and Quality Organization. *Proceedings of the IMEKO TC6 Conference on Digitalization*, Benevento, 3-5 September 2025, 308-313. https://doi.org/10.21014/tc6-2025.057

[12] Blaško, P., Šolc, M., Petrík, J., Girmanová, L. and Blašková, A. (2023) Application of the FMEA Tool in an Accredited Testing Laboratory in the Context of the ISO/IEC 17025:2017 Standard. *Standards*, **3**, 57-69. https://doi.org/10.3390/standards3010006

[13] ISO/IEC 23894:2023 (2023) Information Technology-Artificial Intelligence-Guidance on Risk Management. ISO/IEC. https://www.iso.org/standard/77304.html

[14] NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology.

[15] Mustroph, H. and Rinderle-Ma, S. (2024) Design of a Quality Management System Based on the EU AI Act. In: Savelka, J., Harasta, J., Novotna, T. and Misek, J., Eds., *Frontiers in Artificial Intelligence and Applications*, IOS Press, 314-320. https://doi.org/10.3233/faia241261

[16] Adadi, A. and Berrada, M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6**, 52138-52160. https://doi.org/10.1109/access.2018.2870052

[17] Makhmudov, Kh. and Zagidullina, K. (2026) Conceptual Architecture of an Integrated Management System for Certification Bodies and Testing Laboratories. *STANDART*, No. 3, 132-137. (In Russian)

[18] Nosek, B.A., Ebersole, C.R., DeHaven, A.C. and Mellor, D.T. (2018) The Preregistration Revolution. *Proceedings of the National Academy of Sciences*, **115**, 2600-2606. https://doi.org/10.1073/pnas.1708274114

## Appendix A. Baseline Register Used in the Simulation

Table A1 lists the 19 risks of the register of [8] with their scores and weights. The index is $R = P \times C \times Wt \times Wp$; risk levels: low $R < 3$, medium $3 \le R < 9$, high $9 \le R \le 36$, critical $R > 36$. Risks 1 - 8 are laboratory risks, 9 - 15 impartiality risks and 16 - 19 management risks.

Table A1. Baseline register (19 risks).

| No. | Group | P | C | Wt | Wp | R | Level |
|---|---|---|---|---|---|---|---|
| 1 | laboratory | 2 | 3 | 1.5 | 2.0 | 18.0 | high |
| 2 | laboratory | 2 | 3 | 1.2 | 1.5 | 10.8 | high |
| 3 | laboratory | 2 | 2 | 1.7 | 1.0 | 6.8 | medium |
| 4 | laboratory | 1 | 3 | 1.5 | 1.5 | 6.8 | medium |
| 5 | laboratory | 2 | 2 | 1.7 | 1.0 | 6.8 | medium |
| 6 | laboratory | 1 | 3 | 1.5 | 1.5 | 6.8 | medium |
| 7 | laboratory | 1 | 2 | 1.5 | 0.8 | 2.4 | low |
| 8 | laboratory | 1 | 3 | 1.2 | 1.5 | 5.4 | medium |
| 9 | impartiality | 2 | 3 | 2.0 | 2.0 | 24.0 | high |
| 10 | impartiality | 2 | 3 | 2.0 | 2.0 | 24.0 | high |
| 11 | impartiality | 1 | 3 | 2.0 | 2.0 | 12.0 | high |
| 12 | impartiality | 1 | 3 | 2.0 | 2.0 | 12.0 | high |
| 13 | impartiality | 2 | 2 | 2.0 | 1.0 | 8.0 | medium |
| 14 | impartiality | 1 | 2 | 2.0 | 1.5 | 6.0 | medium |
| 15 | impartiality | 1 | 3 | 2.0 | 1.5 | 9.0 | high |
| 16 | management | 1 | 3 | 1.2 | 1.0 | 3.6 | medium |
| 17 | management | 2 | 3 | 1.0 | 1.5 | 9.0 | high |
| 18 | management | 2 | 3 | 1.7 | 1.5 | 15.3 | high |
| 19 | management | 2 | 2 | 1.0 | 1.0 | 4.0 | medium |

## Appendix B. Python Script Reproducing Table 1

The script below (file ica_sim_reproduce.py) uses NumPy and SciPy and the seed 7 for every row of Table 1. Run as “python ica_sim_reproduce.py”.

```
"""Reproduction script for Table 1 (Section 6): effect of score errors on the ranking
of the 19-risk register. Python 3, numpy, scipy. Fixed seed 7 for every row."""
import numpy as np
from scipy.stats import spearmanr

# Register: P, C, Wt, Wp for risks 1-19 (1-8 laboratory, 9-15 impartiality, 16-19 management)
REG = [(2,3,1.5,2.0),(2,3,1.2,1.5),(2,2,1.7,1.0),(1,3,1.5,1.5),(2,2,1.7,1.0),(1,3,1.5,1.5),
       (1,2,1.5,0.8),(1,3,1.2,1.5),(2,3,2.0,2.0),(2,3,2.0,2.0),(1,3,2.0,2.0),(1,3,2.0,2.0),
       (2,2,2.0,1.0),(1,2,2.0,1.5),(1,3,2.0,1.5),(1,3,1.2,1.0),(2,3,1.0,1.5),(2,3,1.7,1.5),
       (2,2,1.0,1.0)]
P, C, Wt, Wp = np.array(REG, float).T
n = len(P)
BASE = P * C * Wt * Wp                          # reference index R = P*C*Wt*Wp

def level(R):                                   # 0 low (<3), 1 medium [3,9), 2 high [9,36], 3 critical (>36)
    return np.where(R < 3, 0, np.where(R < 9, 1, np.where(R <= 36, 2, 3)))

def top4(R):                                    # ties broken by register order (stable sort)
    return set(np.argsort(-R, kind="stable")[:4])

BASE_LEVEL, BASE_TOP4 = level(BASE), top4(BASE)

def run(q, N=20000, seed=7):
    qv = np.asarray(q, float) * np.ones(n)    # per-risk shift probability (one direction)
    rng = np.random.default_rng(seed)
    rho, keep, chg, altered = [], [], [], []
    for _ in range(N):
        def shift(s):
            u = rng.random(n); d = np.zeros(n)
            d[u < qv] = -1                      # down with probability q
            d[(u >= qv) & (u < 2 * qv)] = 1 # up with probability q
            return np.clip(s + d, 1, 5)         # scores limited to 1-5
        p, c = shift(P), shift(C)
        R = p * c * Wt * Wp
        rho.append(spearmanr(BASE, R).correlation)      # ties get average ranks
        keep.append(top4(R) == BASE_TOP4)
        chg.append((level(R) != BASE_LEVEL).mean())
        altered.append(((p != P).sum() + (c != C).sum()) / (2 * n))
    rho = np.array(rho)
    return (round(rho.mean(), 2), round(float(np.percentile(rho, 5)), 2),
            round(100 * np.mean(keep)), round(100 * np.mean(chg), 1), round(100 * np.mean(altered), 1))

if __name__ == "__main__":
```

```
print("baseline R:", np.round(BASE, 1))
print("baseline top-4 (0-based index):", sorted(BASE_TOP4), "levels:", BASE_LEVEL.tolist())
print("q, mean rho, 5th pct rho, top-4 kept %, level change %, realized altered scores %")
for q in (0.25, 0.15, 0.10, 0.075, 0.05, 0.025, 0.0125):
    print(q, run(q))
lab = [i < 8 for i in range(n)]
print("lab q=0.05, others 0.10:", run([0.05 if l else 0.10 for l in lab]))
print("lab q=0.025, others 0.10:", run([0.025 if l else 0.10 for l in lab]))
print("non-lab q=0.05, lab 0.10:", run([0.10 if l else 0.05 for l in lab]))
```