IMPROVING METHODS FOR DETECTING SYSTEMATIC BIAS IN ASSESSMENT TO ENSURE THE OBJECTIVITY OF DIGITAL CONTROL RESULTS IN HIGHER EDUCATION INSTITUTIONS

Abduganiyeva Mavjuda Asatullayevna
Senior Lecturer, Department of Digital Economy Tashkent State University of Economics
ORCID: 0009-0005-0256-236X
E-mail: mavjudaabduganiyeva02.27@gmail.com
DOI:https://doi.org/10.5281/zenodo.24051252

Abstract

The thesis proposed a workflow to detect systematic bias in digital assessment. The author combined many-facet Rasch measurement, blind second marking, control charts, and false discovery rate control. Simulated tests distinguished rater leniency, severity, central tendency, halo effects, and drift over time. Each signal triggered proportionate feedback or moderation. The author recommended scheduled HEMIS reports, rotating blind reviews, and calibration sessions, and identified validation on real higher-education data as the next step.

Keywords: digital assessment, rater effects, many-facet Rasch model, HEMIS monitoring, blind second marking, calibration, objective grading.

Annotatsiya

Tezisda raqamli baholashdagi tizimli og‘ishlarni aniqlash va nazorat qilish metodikasi ishlab chiqildi. Muallif ko‘p fasetli Rasch modeli, ko‘r ikkilamchi baholash, vaqt nazorati jadvallari va xatolar ulushini boshqarish qoidasini birlashtirdi. Simulyatsiya natijalarida metodika baholovchining yumshoqligi yoki qattiqligi, markaziy tendensiya, halo ta’siri va mezon siljishini aniqladi. Har bir signal uchun mutanosib javob va fikr-mulohaza bosqichlari belgilandi. Muallif HEMIS hisobotlari, tanlab qayta baholash va kalibrlash mashg‘ulotlarini tavsiya qilib, metodikani haqiqiy ma’lumotlarda tekshirish zarurligini qayd etdi.

Kalit so‘zlar: raqamli baholash, baholovchi ta’siri, ko‘p fasetli Rasch modeli, HEMIS monitoringi, ikkilamchi baholash, kalibrlash, xolis baholash.

Аннотация

В тезисе была разработана методика выявления систематических смещений при цифровом оценивании. Автор объединила многофасеточную модель Раша, слепое повторное оценивание, контрольные карты и контроль доли ложных обнаружений. Симуляция показала выявление мягкости и строгости оценивания, центральной тенденции, эффекта ореола и дрейфа критериев. Для сигналов были предусмотрены пропорциональная обратная связь и меры реагирования. Автор рекомендовала отчёты HEMIS, выборочную модерацию и калибровочные занятия, отметив необходимость последующей проверки методики на реальных данных, собранных в университетах.

Ключевые слова: цифровое оценивание, влияние оценщика, многофасеточная модель Раша, мониторинг HEMIS, повторное оценивание, калибровка, объективная оценка.

INTRODUCTION

A grade that a teacher types into a university information system does not stay where it was entered. It feeds the grade point average, decides scholarships and transfers, and ends up on the diploma supplement. Once grades are digital, every mark becomes a permanent and comparable record. This

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

raises the cost of a biased mark, but it also creates the first real opportunity to find bias systematically rather than by complaint.

Uzbekistan has reached this point quickly. According to the National Statistics Committee, the number of students rose from 808.4 thousand in the 2021/2022 academic year to 1,535.0 thousand in 2025/2026, served by 215 higher education institutions and 54.9 thousand teaching staff [1]. The Concept for the development of the higher education system until 2030, approved by Presidential Decree PF-5847 of 8 October 2019, set the gradual transition to the credit-module system and the goal of bringing national universities into the leading international rankings [2]. Assessment records now flow through HEMIS, the national higher education management information system, in which teachers enter marks into an electronic rating book and students see their results; the Digital Government portal lists 207 universities connected to the platform [3]. The data needed to monitor the quality of marking therefore already exist in one place.

Random marking error tends to cancel out across many scripts. Systematic error does not. A rater may be consistently lenient or severe, may avoid the ends of the scale, may let one strong impression colour every criterion, or may slowly change standards from term to term. These patterns, described a century ago as constant errors in ratings [4] and later organised into a taxonomy of rating quality [5], shift whole groups of students up or down. European quality assurance standards ask institutions to apply assessment consistently and fairly to all students [6], and quality understood as fitness for purpose and as transformation of the learner [7] is hard to claim when the grade partly reflects who marked the work.

Universities usually look at average grades by group or by teacher. That check cannot tell a lenient teacher from a strong group of students, because both raise the average. The aim of this thesis is to improve the method for detecting systematic bias in digital assessment results so that rater effects are separated from student ability, tracked over time and turned into a proportionate response. It asks three questions. RQ1: which combination of measurement and monitoring tools separates leniency, severity, central tendency, halo and drift in routine HEI data? RQ2: what flagging rule keeps false alarms low when dozens of raters are screened at once? RQ3: how should results return to raters so that the next round of marking improves? The contribution is an integrated detection workflow with explicit formulas and thresholds, demonstrated on a clearly labelled simulated dataset.

LITERATURE REVIEW

Rater effects and the limits of assessment criteria

E. L. Thorndike observed that officers rating their subordinates gave correlated marks to traits that should have been independent, which he called a constant error and which later became known as the halo effect [4]. F. E. Saal, R. G. Downey and M. A. Lahey reviewed the indices used to judge rating data and grouped the main problems into leniency or severity, halo, central tendency and restriction of range, and low reliability [5].

Detailed rubrics were expected to remove these problems. S. Bloxham, B. den-Outer, J. Hudson and M. Price showed in higher education marking that assessors read and weigh the same criteria differently and that tacit standards survive even well-written rubrics [8]. Consistency therefore has to be checked in the marks themselves rather than assumed from the criteria.

Measurement models for rater-mediated assessment

J. M. Linacre extended the Rasch model to several facets, so that the probability of a score depends at once on the student, the task and the rater [9]. G. Engelhard applied this many-facet model to written composition and showed that raters trained in the same way still differed significantly in severity [10]. C. M. Myford and E. W. Wolfe set out how many-facet Rasch measurement identifies severity, central tendency, halo and randomness through rater measures and fit statistics [11]. S. A. Wind and M. E. Peterson reviewed methods for evaluating rating quality and found that studies draw on several families of indices [12], which supports combining model-based measures with simpler checks.

Monitoring over time and multiple testing

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

Rater measures computed once give a snapshot, while statistical process control follows change. E. S. Page proposed the cumulative sum scheme, which accumulates small deviations until they become evidence of a shift [13], and J. M. Lucas and M. S. Saccucci studied the properties of exponentially weighted moving average charts and gave design guidance for their parameters [14]. When many raters are tested at once, some will look extreme by chance. Y. Benjamini and Y. Hochberg introduced control of the false discovery rate, which limits the expected share of false flags among all flags [15]. The three strands are mature, yet they are rarely joined in one routine for a national digital grade platform. The present work joins them.

METHODOLOGY

The proposed workflow has four steps: assemble linked records, estimate rater effects, monitor change over terms and apply a decision rule that leads to feedback. Figure 1 shows the sequence.

Figure 1 should be read from top to bottom. Grade records from HEMIS and a benchmark sample of blind second marks enter the same dataset, in which every mark is linked to a student, a criterion, a rater and a term. Three detection lenses then look at the data in parallel, and their signals meet in a single decision rule. The rule chooses one of three levels of response, and the dashed loop on the right shows that the outcome returns to raters before the next term, so the system is designed to correct marking rather than only to record problems.

Many-facet Rasch model

For student n, criterion i and rater j, the probability of score category k rather than k−1 is modelled as:

$$\ln \frac{P_{nijk}}{P_{nij(k-1)}} = \theta_n - \delta_i - \alpha_j - \tau_k \quad (1)$$

where θ_n is the ability of the student, δ_i is the difficulty of the criterion, α_j is the severity of the rater and τ_k is the threshold between adjacent categories, all in logits. In plain words, the model asks how likely a given score is once we know how able the student is and how hard the criterion is, and attributes what remains systematic to the rater. A positive α_j means a severe rater and a negative value a lenient one. Each severity estimate is tested against the group mean of zero:

$$z_j = \frac{\hat{\alpha}_j}{SE(\hat{\alpha}_j)}; \quad SE(\hat{\alpha}_j) = \frac{1}{\sqrt{\sum_{n,i} W_{nij}}} \quad (2)$$

where $\hat{\alpha}_j$ is the estimated severity, W_{nij} is the model variance of each observed score and SE is the standard error. Large positive or negative z_j values point to raters whose standard differs from colleagues by more than chance. Consistency is judged with the information-weighted mean square:

$$Infit_j = \frac{\sum_{n,i} (x_{nij} - E_{nij})^2}{\sum_{n,i} W_{nij}} \quad (3)$$

where x_{nij} is the observed score and E_{nij} its expected value under the model. Values near 1 indicate that the rater uses the scale as the model expects. Values below about 0.7 indicate marks that are too predictable, the typical trace of central tendency or halo, and values above about 1.3 indicate erratic marking.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

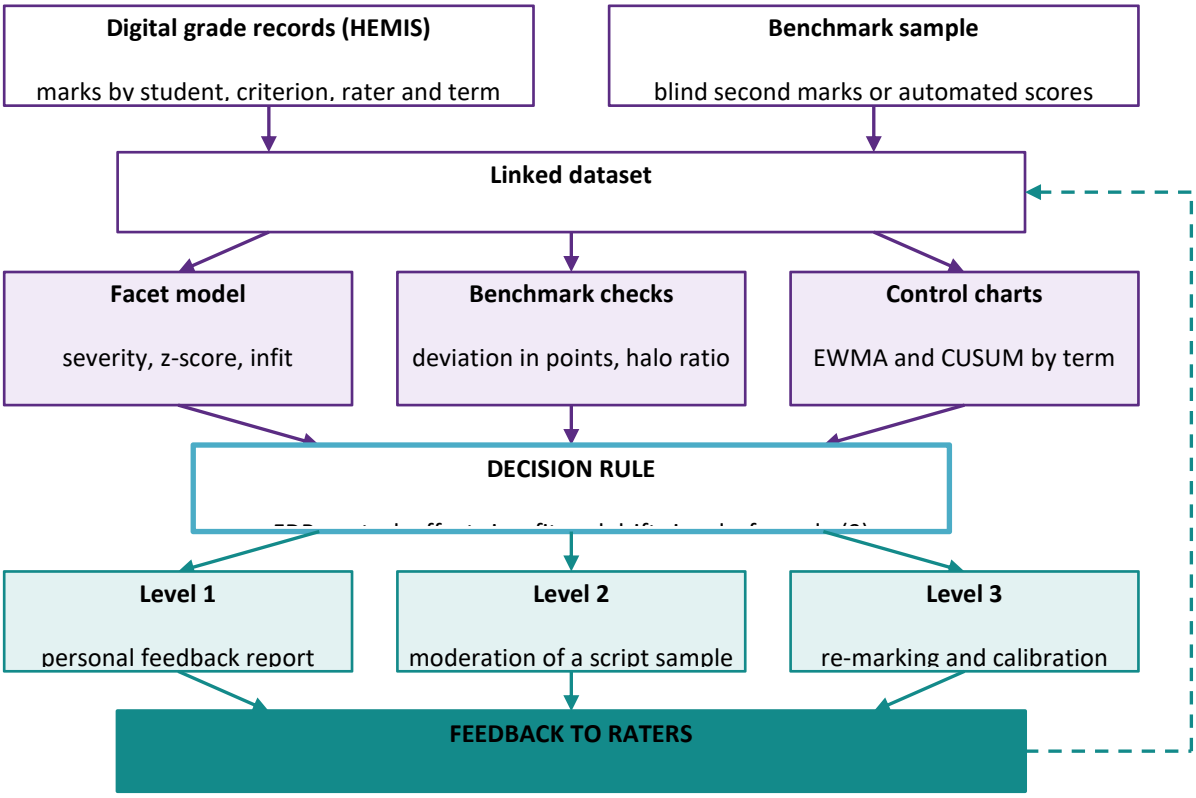


Figure 1. Workflow for detecting systematic assessment bias with feedback to raters[1]

Benchmark checks

A rotating sample of scripts receives a blind second mark from a randomly assigned colleague who does not see the first mark, or an automated score where the task allows it. For script n marked by its teacher, the deviation and the rater-level z-score are:

$$d_n = X_n^T - X_n^B; \quad z_j^B = \frac{\bar{d}_j - \bar{d}}{s_d/\sqrt{n_j}} \quad (4)$$

where X_n^T and X_n^B are the total scores given by the teacher and by the benchmark, $\bar{d}_j$ is the mean deviation of rater j, $\bar{d}$ and s_d are the mean and standard deviation over all raters, and n_j is the number of benchmark scripts of rater j. The deviation is expressed in score points, which makes it easy to explain to a teacher. Halo is captured by a spread ratio:

$$H_j = \frac{\bar{s}_j^T}{\bar{s}_j^B} \quad (5)$$

where $\bar{s}_j^T$ and $\bar{s}_j^B$ are the average standard deviations of criterion scores within a script under the teacher and under the benchmark. A ratio well below 1 means that the rater gives nearly the same score to every criterion of a script, which is what halo looks like in the data.

Control charts for drift

For each term t the benchmark deviation of rater j is standardised against that rater's own baseline of four terms, giving z_{jt}. The exponentially weighted moving average and its control limits are:

$$E_{jt} = \lambda z_{jt} + (1-\lambda) E_{j,t-1}; \quad E_{j0} = 0 \quad (6)$$

$$UCL_t = L\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]}; \quad LCL_t = -UCL_t \quad (7)$$

where λ is the weight given to the newest term and L sets the width of the limits. The one-sided cumulative sums are:

[1] Source: Developed by the author on the basis of [9]; [11]; [13]; [14]; [15].

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

$$C_{jt}^{+}=\max\left(0,C_{j,t-1}^{+}+z_{jt}-k\right)$$
$$C_{jt}^{-}=\max(0,C_{j,t-1}^{-}-z_{jt}-k) \quad (8)$$

where k is the allowance that ignores very small changes and a signal is given when either sum exceeds the decision interval h. The parameters were set to $\lambda = 0.2$, $L = 3$, $k = 0.5$ and $h = 5$, following common practice [14]. In everyday terms, the EWMA is a smoothed running average of how far a rater has moved from his or her own past standard, and the CUSUM adds up small, repeated moves in the same direction until they become too large to ignore.

Decision rule and illustrative dataset

The signals are combined into one flag:

$$F_j=\max(S_j,C_j,H_j^{*},D_j) \quad (9)$$

where $S_j = 1$ if the severity test survives false discovery rate control at $q = 0.05$ [15] and $|\hat{\alpha}_j|$ is at least 0.30 logits, $C_j = 1$ if infit lies outside 0.7–1.3 or the share of extreme categories falls below half of the median, $H_j^{\cdot} = 1$ if the spread ratio is below 0.6 without signs of central tendency, and $D_j = 1$ if a control chart signals. A flag leads to Level 1 (a personal feedback report) when only a modest severity effect below 0.50 logits is present, and to Level 2 (moderation of a sample of scripts) in all other cases; Level 3 (re-marking of the affected group and calibration training) follows only when moderation confirms the bias. The effect size guard keeps statistically significant but trivial differences out of the process.

Because institutional micro data were not available, the method was demonstrated on an illustrative dataset simulated in Python (NumPy, SciPy) that contains no real student or staff data. Assumptions: 900 scripts, 30 raters and four criteria scored 0–4; each script marked by its teacher and by a randomly assigned blind second marker; student abilities drawn from N(0, 1); rater severities drawn from N(0, 0.15); and five planted effects. R07 was lenient by 0.80 logits and R19 severe by 0.70 logits; R23 moved a score one step toward the middle with probability 0.7 (central tendency); R28 copied one global impression to each criterion with probability 0.8 (halo); R12 became more lenient by 0.25 logits per term from term 5 in a separate panel of eight terms with 30 benchmark scripts per rater and term (drift). The model was estimated by joint maximum likelihood.

ANALYSIS AND RESULTS

Context: the scale of marking in Uzbek higher education

Table 1

Students, teaching staff and higher education institutions in Uzbekistan[2]

Academic year	HEIs	Students, thousand	Teaching staff, thousand	Students per teacher
2021/2022	154	808.4	37.4	21.6
2022/2023	191	1,042.1	41.7	25.0
2023/2024	219	1,314.5	45.2	29.1
2024/2025	222	1,432.8	49.6	28.9
2025/2026	215	1,535.0	54.9	28.0

Table 1 shows that the student body almost doubled in four years (by a factor of 1.90), while teaching staff grew by a factor of 1.47. The number of students per teacher rose from 21.6 to 28.0. For marking, this means that each teacher now produces a larger share of the grades that enter HEMIS, so a single biased standard reaches more students. Manual review of every script is no longer realistic, so screening the data already stored on the platform becomes the practical way to protect fairness.

Static diagnosis of rater effects (illustrative data)

[2] Source: National Statistics Committee of the Republic of Uzbekistan [1]. Students per teacher is the author's calculation based on [1]. Teaching staff exclude part-time instructors; figures are at the beginning of each academic year.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

Table 2

Rater diagnostics for selected raters (illustrative simulated data)[3]

ID	Simulated effect	Severity, logits	z	Infit	Deviation, points	H	Extreme, %	Flag and level
R07	Leniency (−0.80)	−1.06	−13.21	0.93	+3.10	0.89	20	S, Level 2
R19	Severity (+0.70)	0.93	9.85	1.12	−1.93	1.20	30	S, Level 2
R23	Central tendency (+0.12)	0.27	3.30	0.61	−0.80	0.54	6	C, Level 2
R28	Halo (−0.32)	−0.46	−5.26	0.81	+1.67	0.41	19	S and H, Level 2
R05	None (−0.23)	−0.44	−5.47	0.95	+1.00	0.89	20	S, Level 1
R17	None (+0.28)	0.40	4.92	1.00	−1.97	0.97	18	S, Level 1
R02	None (−0.18)	−0.28	−3.58	1.08	+0.93	1.03	26	No flag
–	Median of 22 unflagged raters	0.02	0.15	1.05	−0.10	1.04	20	–

Table 2 shows that the model recovered the simulated rater severities well, with a correlation of 0.95 between true and estimated values. The lenient rater R07 gave on average 3.10 points more than the blind mark on a 16-point total, and the severe rater R19 gave 1.93 points less; both were flagged with very large z-scores. Central tendency in R23 did not show up in the severity measure but was clear in the fit statistics: infit fell to 0.61 and only 6% of marks were at the ends of the scale, against a median of 20%. Halo in R28 produced a spread ratio of 0.41, so criterion scores within a script were less than half as varied as under the benchmark. Each distortion left its own fingerprint, and no single statistic caught all four.

Two further results support the design of the rule. Rater R02 was statistically significant after correction, yet its severity of −0.28 logits was below 0.30 in absolute value, so it fell under the effect size guard and it was not flagged. Four raters without planted effects (R05, R06, R17 and R29) were flagged at Level 1; their simulated true offsets ranged from 0.15 to 0.28 logits, so these were genuine but mild differences that suit a feedback report rather than an intervention. Used alone, the benchmark z-score also signalled rater R03 (z = −2.71), whose true severity was close to zero, because its deviation

[3] Source: Author's calculation on an illustrative dataset simulated in Python (900 scripts, 30 raters, four criteria scored 0–4); not real student or staff data. Severity from formulas (1)–(2), infit from formula (3), deviation from formula (4), spread ratio H from formula (5). Deviation is teacher total minus blind total on a 0–16 scale. Values in brackets are the simulated true severities. S = severity flag, C = central tendency flag, H = halo flag.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

reflected the severities of the colleagues who second-marked its scripts. The facet model, which adjusts for every rater at once, did not flag R03, and this is the main argument for joining the two lenses.

Monitoring drift over terms (illustrative data)

Table 3

Drift monitoring for rater R12 by term (illustrative simulated data)[4]

Term	Simulated drift, logits	Deviation from blind mark, points	z	EWMA	EWMA limit	CUSUM (upper)
1	0	0.13	0.30	0.06	0.60	0.00
2	0	−0.23	−0.52	−0.06	0.77	0.00
3	0	0.47	1.05	0.16	0.86	0.55
4	0	−0.37	−0.82	−0.03	0.91	0.00
5	−0.25	0.37	0.82	0.14	0.94	0.32
6	−0.50	1.60	3.59	0.83	0.97	3.42
7	−0.75	1.83	4.12	1.49	0.98	7.04
8	−1.00	2.63	5.92	2.37	0.99	12.45

Table 3 traces rater R12. During the four baseline terms the deviation from the blind mark stayed within ±0.5 points. The drift began in term 5 and was too small to stand out at first; both charts signalled in term 7, when the EWMA of 1.49 exceeded its limit of 0.98 and the cumulative sum of 7.04 passed the decision interval of 5. Across the other 29 raters and four monitored terms there was one false alarm (R27 in term 7), which moderation of a small sample would clear. The charts tell the quality office not only who changed but also when, so only the affected terms need review.

CONCLUSION AND RECOMMENDATIONS

The thesis joined three lenses to detect systematic bias in digital assessment. In answer to RQ1, the many-facet Rasch model separated rater severity from student ability, fit statistics and the share of extreme categories revealed central tendency, the spread ratio exposed halo, and control charts located drift in time. In answer to RQ2, a rule that combines false discovery rate control with a minimum effect of 0.30 logits flagged all planted effects while leaving trivial differences aside. In answer to RQ3, a three-level response turns each flag into feedback first and into re-marking only after moderation confirms it.

Four recommendations follow for Uzbek HEIs and for the developers of HEMIS. First, rotating blind second marking of a modest share of written scripts each term would create the benchmark. Second, the formulas can run as a scheduled HEMIS report that gives each teacher a one-page summary in score points. Third, thresholds and response levels agreed in advance make flags a routine part of quality assurance. Fourth, calibration sessions on anonymised flagged scripts would close the feedback loop in Figure 1.

The results come from simulated data and show what the method can detect under stated assumptions. The next step is validation on real HEMIS records, testing the thresholds, the number of benchmark scripts needed per rater and the speed of drift detection. Later work can add student groups, such as language of instruction, as a facet. These steps can make the expanding digital assessment system of Uzbekistan a source of trusted, comparable results.

[4] Source: Author's calculation on an illustrative dataset simulated in Python (30 benchmark scripts per rater and term); not real data. EWMA and limits from formulas (6)–(7) with λ = 0.2 and L = 3; CUSUM from formula (8) with k = 0.5 and h = 5. Terms 1–4 form the baseline. Negative drift and positive deviation both mean growing leniency.

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.

REFERENCES

1. National Statistics Committee of the Republic of Uzbekistan. (2026). Higher education in the Republic of Uzbekistan (as of the beginning of the 2025/2026 academic year). Press release, 26 May 2026. https://stat.uz/img/news/higher-education_p92388.pdf
2. Decree of the President of the Republic of Uzbekistan No. PF-5847 of 08.10.2019. On approval of the Concept for the development of the higher education system of the Republic of Uzbekistan until 2030. https://lex.uz/docs/4545884
3. Digital Government of the Republic of Uzbekistan. (n.d.). Higher Education Process Management – HEMIS. Retrieved November 6, 2026. https://dgov.uz/en/solution/detail/50/
4. Thorndike, E. L. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4(1), 25–29. https://doi.org/10.1037/h0071663
5. Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413
6. ENQA, ESU, EUA, & EURASHE. (2015). Standards and guidelines for quality assurance in the European Higher Education Area (ESG). Brussels. https://www.enqa.eu/wp-content/uploads/2015/11/ESG_2015.pdf
7. Harvey, L., & Green, D. (1993). Defining quality. Assessment & Evaluation in Higher Education, 18(1), 9–34. https://doi.org/10.1080/0260293930180102
8. Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. Assessment & Evaluation in Higher Education, 41(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607
9. Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA Press. https://www.winsteps.com/a/Linacre-MFRM-book.pdf
10. Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, 31(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x
11. Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement, 4(4), 386–422. https://pubmed.ncbi.nlm.nih.gov/14523257/
12. Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. Language Testing, 35(2), 161–192. https://doi.org/10.1177/0265532216686999
13. Page, E. S. (1954). Continuous inspection schemes. Biometrika, 41(1–2), 100–115. https://doi.org/10.1093/biomet/41.1-2.100
14. Lucas, J. M., & Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. Technometrics, 32(1), 1–12. https://doi.org/10.1080/00401706.1990.10484583
15. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

© 2026 International Conference on Artificial Intelligence and Applications . All Rights Reserved.