Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI Assisted Video Analysis Reveals Divergent Laboratory Competence Profiles Despite Equivalent Practical Scores in Undergraduate Titration

Agung Tri Prasetya
agungchem@mail.unnes.ac.id
Universitas Negeri Semarang
Devy Lestari
Universitas Negeri Semarang
Siti Herlina Dewi
Universitas Negeri Semarang
Dimas Gilang Ramadhani
Universitas Negeri Semarang

Research Article

Keywords: digital assessment, practical laboratory skills, video-based assessment, artificial intelligence, higher education, formative assessment

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10876669/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Practical performance assessment in higher education often compresses multiple manipulative, perceptual, and decision-making actions into a single grade, potentially obscuring differences that matter for feedback. This study examined what an AI-assisted video-evidence workflow retained and lost when representing undergraduate strong acid–strong base titration performance. The verified implementation dataset comprised 26 chemistry-education undergraduates, each represented by one practical-performance video and a seven-domain, 25-micro-skill rubric. The evidence-processing workflow had been developed separately using 300 titration-related records and was used to locate and structure observable evidence rather than to replace human judgment. A measurement-integrity audit showed that 627 of 650 micro-skill records (96.46%) required manual review and 623 (95.85%) were produced through a fallback evidence route; the scores were therefore interpreted as evidence-derived diagnostic measures rather than validated summative grades. Within that boundary, domain scores differed overall, Friedman $\chi^2(6) = 41.47$, $p < .001$, Kendall's $W = .266$. Endpoint observation exceeded titration execution and repetition/consistency after Holm correction. Endpoint recognition was substantially more prevalent than timely stopping (80.8% versus 42.3%), and 13 of 15 exact-score pairs (86.7%) had non-identical domain profiles. The findings show that identical practical totals can represent different configurations of enacted performance. For digital assessment of practical STEM learning, aggregate grades are therefore more informative when accompanied by domain profiles, traceable evidence, and explicit measurement-quality information.

1. Introduction

Laboratory work occupies a distinctive position in chemistry education because students must coordinate chemical knowledge with material action. They select and manipulate apparatus, control quantities and rates, attend to visible and instrumental cues, decide when a process has reached an appropriate state, record measurements, and reproduce procedures with acceptable consistency. Reviews of university chemistry laboratory research therefore characterize laboratory learning as multidimensional, encompassing experimental competence, disciplinary understanding, higher-order reasoning, transversal capabilities, and affective outcomes [1]. Yet direct assessment of enacted practical skill remains less common than assessment through written reports or products, even when the intended learning outcome is the ability to perform chemistry [2, 3]. This mismatch creates a measurement problem: if practical competence is multidimensional, an assessment system must preserve enough information to distinguish different forms of successful and unsuccessful enactment. This concern is reinforced by work showing that laboratory learning involves interconnected domains of struggle and science practices, and that faculty goals span cognitive and psychomotor outcomes that are not captured well by product-only assessment [22, 25, 27, 35].

Titration makes this problem particularly visible. A conventional titration score can combine pipetting, solution preparation, buret reading, stopcock operation, titrant-flow control, swirling, endpoint recognition, stopping, final reading, reporting, and repetition. These actions are not educationally

interchangeable. A student who recognizes a phenolphthalein endpoint but cannot reduce the addition rate or stop the buret appropriately has a different learning need from a student who controls the buret well but reads the meniscus inconsistently. Likewise, completing one acceptable trial is not equivalent to demonstrating stable performance across repetitions. Research on hands-on laboratory assessment has responded to this challenge with direct observation, task inventories, rubrics, practical examinations, digital badges, and explicit performance criteria [4–9]. Student-goal studies and direct audits of basic laboratory technique further show that task completion, grades, or an apparently acceptable product do not necessarily imply stable practical performance, strengthening the case for explicit performance criteria [26, 28–31].

Video can add a useful evidence layer because it retains observable enactment that may be difficult to capture in a single end-of-session score. Student-generated and instructor-reviewed videos have been used to document laboratory techniques, support peer review, enable self-assessment, and help students critique correct and incorrect procedures [6, 10–13]. However, a video does not automatically constitute valid measurement. Camera position, visibility, lighting, event timing, evidence selection, and the route by which observations are interpreted can introduce construct-irrelevant variation. Recent work on digital and automated assessment likewise emphasizes that technology-supported scores require explicit validation of what the assessment captures before they are used for educational decisions [14, 15]. Quantitative chemistry education research therefore requires an explicit validity argument connecting observable evidence to the interpretation and intended use of scores [16, 17]. The relevant question is not whether software can output a number, but whether the evidence represented by that number is observable, traceable, and sufficiently stable for the intended educational decision. Video-supported preparation, student-produced technique videos, digital-badge submissions, and video reports have likewise been used to make laboratory action reviewable for instruction, critique, and assessment [23, 24, 32–34, 36].

A second problem arises after evidence has been converted to a rubric score. A weighted total is administratively useful, but it is a many-to-one transformation: different multidimensional profiles can map to the same scalar result. Strength in one domain can compensate for weakness in another, so two students with the same grade need not have the same underlying performance configuration. This matters for formative assessment because equal totals can imply unequal instructional needs. Broader chemistry education research has similarly emphasized that assessment should serve learning and feedback functions rather than merely rank students [18, 19]. A practical assessment that preserves domain-level structure can therefore be more educationally informative than one that reports only an overall mark.

The present study examines these two linked issues using a video-derived dataset from an undergraduate titration practical. The educational contribution is deliberately separated from software novelty: automated processing is treated as an evidence-location and organization layer, while the analysis focuses on what practical scores reveal and conceal. Three research questions guided the study. RQ1 asked how evidence-derived titration performance varied across laboratory-skill domains and

critical micro-skills. RQ2 asked whether students with identical aggregate practical scores could nevertheless exhibit different underlying domain profiles. RQ3 asked which measurement-integrity constraints determine the defensible interpretation of scores produced from a non-uniform video-evidence workflow. Together, these questions test whether a practical-assessment reporting architecture preserves information that is useful for diagnosis and feedback without overstating the validity of automated scoring.

2. Method

2.1 Research design and educational context

This study used a cross-sectional observational performance-assessment design based on existing laboratory video evidence. It did not evaluate a teaching intervention, nor did it test whether an automated system improved learning outcomes. Instead, the study examined the structure, diagnostic information, and measurement constraints of scores generated from an undergraduate strong acid-strong base titration task. The student was the independent sampling unit. Domain scores and micro-skill ratings were repeated observations nested within students and were never treated as additional participants. This distinction was maintained in all descriptive and inferential analyses.

The educational implementation took place in an undergraduate chemistry-education context at a public university in Central Java, Indonesia. The recorded task represented a standard acid-base titration sequence and required students to prepare the sample and apparatus, use volumetric glassware, read and operate the buret, regulate titrant delivery, coordinate buret and flask actions, observe and verify the endpoint, record the final reading, and demonstrate procedural repetition. The construct of interest was enacted titration performance visible in the video record. Consequently, the analysis did not infer unobserved competence from actions that fell outside the camera view and did not use written-report quality or self-report as substitutes for observed practical behavior.

2.2 Participants, task records, and unit of analysis

The verified educational implementation dataset contained 26 undergraduate students. Each participant contributed one practical-performance video, producing 26 participant-level video records. Student identifiers were replaced with analytic codes for pairwise profile analysis. The current workbook contained 182 domain-score records (26 students × 7 domains), 650 micro-skill records (26 × 25), and 2,600 criterion-level records (four criterion traces per micro-skill). These repeated records were not treated as independent participants.

Each participant contributed one continuous practical-performance video. The records captured the sequence from preparation through titration, endpoint work, reporting, and closing/repetition phases, but they were not produced under a fully standardized cinematographic protocol. The source files did not establish a common camera model, fixed camera-to-bench distance, or identical lighting for every participant. Recording hardware was therefore not modeled as a student characteristic. Instead, visibility

limitations were handled analytically by auditing evidence route, confidence/support information, and manual-review flags as possible sources of construct-irrelevant score variation.

2.3 Development corpus, technical quality assurance, and workflow freezing

The evidence-processing workflow was developed independently of the educational implementation using a corpus of 300 titration-related records. A development record represented a video/evidence-processing case rather than a participant in the present educational sample. The corpus was used to establish the frame-sampling routine, region-of-interest structure, activity-label vocabulary, mapping between observable evidence and rubric items, evidence-packet format, and structured output schema. None of the 300 development records contributed to the participant-level statistics reported in this article, and implementation videos were not added retrospectively to the development corpus.

Technical readiness was evaluated before implementation through a five-part quality-assurance gate: source integrity, temporal traceability, rubric coverage, evidence recoverability, and leakage control. A development case was considered technically usable only when the source video could be opened, relevant frames or intervals could be linked to timestamps, major apparatus/action regions could be represented, evidence packets could be produced for the applicable rubric families, and the structured output could be traced back to its source evidence. Cases that failed these conditions were not used to define interpretation rules. The gate was designed to establish whether the workflow could recover auditable evidence, not to inflate the development corpus by retaining unusable records.

The quality-assurance stage established operational readiness rather than educational validity. The development corpus was not an exhaustive expert-labelled gold standard for all 25 micro-skills, so no accuracy, sensitivity, or specificity statistic is claimed for it. After the quality gate, region definitions, feature extraction, activity mappings, evidence-packet structure, and output fields were frozen before the implementation dataset was analyzed. The defensible development claim is therefore limited and reproducible: the 300-record corpus supported a stable, traceable evidence-extraction architecture. Whether a located action constituted competent laboratory performance remained a rubric-based educational judgment rather than a property inferred from technical processing alone.

2.4 Titration performance rubric and score construction

The analytic rubric contained seven functional domains and 25 micro-skills. The domains were designed to retain distinctions that are instructionally relevant within a titration rather than to claim seven independent psychological traits. Table 1 summarizes the framework. The existing scoring architecture assigned weights of A = .20, B = .10, C = .15, D = .20, E = .20, F = .10, and G = .05. Domain profiles were treated as the primary diagnostic representation, while the weighted total was treated as a secondary summary used to investigate score compression.

Table 1
Analytical architecture of titration performance

Domain	Competence construct	Micro-skills	Functional interpretation
A	Pipetting	A1-A5	Volumetric handling and transfer
B	Flask preparation	B1-B3	Procedural preparation
C	Buret operation	C1-C4	Measurement and flow regulation
D	Titration execution	D1-D4	Dynamic manipulation and coordination
E	Endpoint observation	E1-E4	Visual recognition and stopping decision
F	Reading/reporting/closing	F1-F3	Measurement completion and documentation
G	Repetition/consistency	G1-G2	Procedural reproducibility

Micro-skill ratings were represented on an ordinal 0–4 scale. The intended anchors were 0 = observable but not correctly demonstrated, 1 = minimally demonstrated, 2 = partially demonstrated, 3 = substantially demonstrated, and 4 = fully and clearly demonstrated. In a human-rated implementation, NA is reserved for insufficiently observable evidence and is conceptually distinct from 0. The current workbook, however, encoded all 650 micro-skill records as scored; therefore, the present analysis did not retrospectively manufacture NA values or reinterpret uncertain observations as failures. Domain scores were expressed on a 0-100 scale as 25 times the mean item rating within the domain. Ratings of 3 or 4 were used only as a descriptive proficiency indicator in selected micro-skill comparisons and are not presented as an externally validated mastery standard.

Domain score = 25 × mean item rating; Overall score = 0.20A + 0.10B + 0.15C + 0.20D + 0.20E + 0.10F + 0.05G

2.5 Video-evidence extraction and decision-support workflow

The implementation workflow linked raw videos to rubric evidence through seven stages: raw-video ingestion, frame sampling, region-of-interest detection, OpenCV feature extraction, contextual interpretation, human-review routing, and structured output. Motion information was used to locate periods of apparatus manipulation, spatial regions were used to distinguish pipette, buret, flask, and reporting areas, color features supported endpoint-related evidence search, and timestamps preserved temporal traceability. For every retained micro-skill record, the structured output stored the source identifier, rubric code, evidence timestamp or interval, activity label, evidence image/path, confidence or support information where available, scoring rationale, and manual-review flag. Figure 1 summarizes this auditable evidence chain.

The role of automated processing was deliberately bounded. OpenCV and contextual AI outputs were treated as evidence-location and decision-support signals rather than autonomous educational judgments. This distinction is especially important because the implementation workflow did not use the same contextual-model route for every item or every student. The purpose of retaining timestamps and evidence images was therefore auditability: a human reviewer should be able to return from a numerical score to the observable source evidence that motivated it.

2.6 Measurement-integrity audit and human-validation boundary

Before substantive interpretation, the implementation dataset was audited for participant counts, completeness of the A-G score structure, ordinal-score distribution, manual-review recommendations, evidence-processing route, repeated item vectors, overall-score range, and associations between processing route and total score. The audit was treated as part of the validity argument rather than as a secondary technical check because workflow defaults can create apparent student differences that are partly measurement-generated. The purpose was therefore to determine the level of interpretation that the available scores could legitimately support.

The audit identified two indicators of route dependence that were carried forward into interpretation: manual-review count and contextual-model count were each associated with overall score. These associations were not interpreted causally and were not used to correct, smooth, or reweight individual student scores. Instead, they defined an explicit use boundary: the present ratings were analyzed as evidence-derived diagnostic measures suitable for examining score structure, bottlenecks, and profile compression, but not as validated summative grades. A blinded two-rater validation protocol, including 0–4/NA anchors, adjudication rules, coverage requirements, weighted kappa, and ICC reporting, is provided in Electronic Supplementary Material 1 (ESM 1) for any confirmatory or high-stakes reuse of the workflow.

2.7 Statistical analysis

Domain-level distributions were summarized using mean, standard deviation, median, interquartile range, and range. Because the same students contributed scores to all seven domains and the current domain distributions were discrete and tied, a Friedman repeated-measures test was used for the omnibus comparison, with Kendall's W as an effect-size measure. Pairwise comparisons used paired Wilcoxon signed-rank tests with Holm adjustment across the 21 domain contrasts. Rank-biserial effect sizes were reported for contrasts that remained significant after multiplicity correction.

For micro-skill analysis, the number and percentage of students with a current rating of 3 or 4 were reported together with Wilson 95% confidence intervals. Four pre-specified functional comparisons were examined descriptively: endpoint recognition to timely stopping, endpoint recognition to endpoint verification, flow regulation to coordinated execution, and repetition to procedural consistency. Because

the current evidence timestamps are evidence-selection locations rather than a validated event log, these comparisons were not interpreted as temporal transition probabilities or causal pathways.

To evaluate aggregate-score compression, students with exactly the same overall score were grouped and all within-group pairs were compared across the seven domains. Weighted absolute profile distance was calculated as Σwk|Dik - Djk| using the scoring weights above, and mean absolute domain difference (MADD) was calculated as the unweighted mean of the seven absolute domain differences. A distance of zero represented an identical A-G profile; any non-zero distance represented configurational non-equivalence despite identical total score. For exact-score pairs, the signed weighted domain differences necessarily sum to zero; therefore, one-half of the weighted absolute profile distance represents the amount of weighted score contribution on one side of the profile that is offset by opposing differences in other domains. This derived quantity was used only to interpret the degree of compensation and was not treated as a separate inferential statistic. A secondary near-score sensitivity analysis examined pairs within five overall-score points, but exact-score groups were retained as the main analysis because they do not depend on an arbitrary tolerance threshold.

All tests were two-sided. Statistical significance was evaluated at α = .05 after the specified multiplicity adjustment. Given the small participant-level sample and the measurement-status findings, effect sizes, confidence intervals, and pattern consistency were emphasized over isolated p-values.

3. Results

3.1 Measurement integrity and interpretive boundary of the evidence-derived scores

The first result concerned measurement integrity rather than student ability. The dataset was structurally complete at the participant and rubric-record levels, with 26 participants, 182 domain-score records, 650 micro-skill ratings, and 2,600 criterion-level records. The evidence-generation route, however, was highly non-uniform: 627 of 650 micro-skill ratings (96.46%) were flagged for manual review, 623 (95.85%) were produced through the fallback/no-model route, and only 27 (4.15%) contained contextual-model output. This corresponds to a fallback-to-contextual routing ratio of approximately 23:1. The ordinal distribution was also concentrated around score 2, which accounted for 463 ratings (71.2%), compared with 179 score-3 ratings (27.5%), two score-1 ratings (0.3%), six score-4 ratings (0.9%), and no score-0 ratings. Table 2 condenses these measurement-status indicators. These indicators are interpreted as measurement-quality diagnostics rather than achievement indicators, consistent with validity guidance that instrument behavior and evidence quality should be separated from substantive score interpretation [16, 17].

The routing imbalance was not statistically neutral with respect to the resulting total score. Manual-review count correlated negatively with overall score (Spearman $r_s = -.482$, $p = .0127$), whereas contextual-model count correlated positively with overall score ($r_s = .483$, $p = .0125$). The two

correlations point in opposite directions but to the same measurement concern: participants whose evidence passed through different processing routes did not receive scores that were independent of route exposure. This does not establish that the route caused the score difference, but it means that workflow characteristics are entangled with the observed outcome and must be considered when interpreting between-student differences. Consequently, the current totals are more defensible as evidence-derived diagnostic summaries than as autonomous summative grades. This treatment follows validity-oriented quantitative practice in chemistry education, in which construct-irrelevant variation should be identified before score differences are given substantive meaning [16].

A second diagnostic came from the internal structure of the 25 micro-skills. Five sets of items shared exactly the same participant-level score vector: A1-A5, B1-B3, C2/C4, D1/D2, and D3/D4. Fourteen nominal items therefore participated in duplicated-vector clusters, leaving at most 16 distinct participant-level score patterns across the 25 rubric items. The redundancy is too systematic to be interpreted as ordinary similarity among related skills and instead indicates that the fallback logic compressed several distinct rubric behaviors into common numerical outputs. Taken together, routing imbalance, central score compression, and duplicated item vectors converge on the same conclusion: the participant-level profile structure can be analyzed, but the nominal 25-item granularity should not be treated as 25 independent validated measurements. The subsequent analyses therefore emphasize robust domain and configurational patterns rather than fine-grained claims about every individual item. The decision to privilege domain-level patterns over fine item-level distinctions is therefore consistent with rubric-based practical assessment, in which observable criteria need to remain behaviorally distinguishable if they are to support diagnostic inference [8, 9].

Table 2
Compact measurement-status audit for the implementation dataset

N	Domain n	Micro-skill n	Manual review	Fallback route	Contextual AI	Score = 2	Overall M ± SD
26	182	650	627/650 (96.46%)	623/650 (95.85%)	27/650 (4.15%)	463/650 (71.2%)	57.50 ± 4.67

Note. Overall score is mean ± SD. The audit supports cautious, evidence-derived interpretation rather than autonomous summative grading.

3.2 Domain-level structure of titration performance

Despite the measurement constraints, the domain scores showed a structured pattern rather than uniform performance. The overall participant score averaged 57.50 ± 4.67, with a median of 57.40, an interquartile range of 55.00-59.61, and a range of 49.38–70.62. Endpoint observation (E) had the highest mean score (64.90 ± 10.62), followed by reading/reporting/closing (F; 61.86 ± 11.59), whereas titration execution (D; 51.92 ± 5.80) and repetition/consistency (G; 50.96 ± 3.40) were lowest. The distributions were also different in shape: E ranged from 37.50 to 81.25, showing substantial heterogeneity around its higher center, while G occupied a narrow 50.00-62.50 band, indicating a compressed low-scoring pattern.

Thus, the domain result was not simply a ranking of means; the strongest domain was also one of the most heterogeneous, whereas the weakest reproducibility domain showed relatively little spread. Interpreting these domains as distinct but interdependent components is consistent with multidimensional accounts of laboratory learning rather than a single undifferentiated practical construct [3, 20].

The Friedman test indicated an overall difference among the seven domain distributions, $\chi^2(6) = 41.472$, $p < .001$, Kendall's $W = .266$. The multiplicity-controlled pattern was selective rather than global. Only three of the 21 paired contrasts remained significant after Holm adjustment: endpoint observation was higher than titration execution (Holm $p = .0036$, |rank-biserial| = .880) and repetition/consistency (Holm $p = .0013$, |rank-biserial| = .960), while reading/reporting/closing was higher than repetition/consistency (Holm $p = .0148$, |rank-biserial| = 1.000). All three surviving effects were large, but they were concentrated around the contrast between endpoint/reporting performance and execution/reproducibility rather than distributed across all possible domain pairs. Table 3 presents the descriptive and inferential results. The selective inferential pattern is therefore better read as evidence of a structured competence profile than as a simple ranking of seven independent skills, an interpretation aligned with contemporary views of laboratory learning as multidimensional and practice-based [1, 20].

This selective contrast structure is analytically important. The data do not support seven clearly separated levels of competence, because most pairwise comparisons - including many comparisons among pipetting, flask preparation, and buret operation - did not survive multiplicity control. Instead, the strongest separation occurred along a narrower performance axis: observable endpoint and reporting behaviors were more favorable than dynamic execution and repeated consistency. The result therefore localizes the domain-level difference rather than merely declaring that domains are unequal. In practical terms, the class profile was characterized by a relative advantage in recognizing or documenting titration states and a relative disadvantage in sustaining controlled and reproducible enactment. This emphasis on enacted laboratory processes rather than only end products parallels authentic practical-skill assessment approaches that retain information about specific components of performance [7, 9]. Recent arguments for assessing science practices in undergraduate chemistry laboratories similarly emphasize direct evidence of process and alignment between intended competencies and what assessments actually reward [22, 29].

Table 3
Domain-level titration-performance structure and multiplicity-controlled contrasts

Domain	Mean ± SD	Median [Q1-Q3]	Range	≥ 75 n/N (%)
A. Pipetting	56.73 ± 11.31	50.00 [50.00-62.50]	50.00–75.00	7/26 (26.9%)
B. Flask prep	55.77 ± 10.74	50.00 [50.00–50.00]	50.00–75.00	6/26 (23.1%)
C. Buret operation	56.49 ± 10.07	50.00 [50.00-62.50]	50.00–75.00	5/26 (19.2%)
D. Titration execution	51.92 ± 5.80	50.00 [50.00–50.00]	50.00–75.00	1/26 (3.8%)
E. Endpoint observation	64.90 ± 10.62	65.62 [62.50-71.88]	37.50-81.25	7/26 (26.9%)
F. Reading/reporting	61.86 ± 11.59	66.67 [50.00–75.00]	50.00–75.00	9/26 (34.6%)
G. Repetition/consistency	50.96 ± 3.40	50.00 [50.00–50.00]	50.00-62.50	0/26 (0.0%)

Omnibus: Friedman $\chi^2(6) = 41.472$, $p < .001$, Kendall's $W = .266$. Holm-significant paired contrasts: E > D ($p = .0036$, $|r_{rb}| = .880$), E > G ($p = .0013$, $|r_{rb}| = .960$), and F > G ($p = .0148$, $|r_{rb}| = 1.000$).

3.3 Functional discontinuities across critical titration micro-skills

The micro-skill analysis identified a pronounced asymmetry between recognizing an endpoint and acting on it. Twenty-one of 26 students (80.8%, Wilson 95% CI 62.1–91.5) had a current E2 rating of at least 3 for endpoint recognition, whereas only 11 of 26 (42.3%, 25.5–61.1) met the same threshold for E3 timely stopping. The absolute difference was − 38.5 percentage points, and the descriptive retention ratio was 0.52. By contrast, endpoint verification remained close to recognition: E4 was at least 3 for 20 of 26 students (76.9%, 57.9–89.0), only 3.9 percentage points below E2, with a retention ratio of 0.95. The large difference in retention between E2→E3 and E2→E4 shows that the major loss in the current output is specific to the conversion of recognition into stopping behavior rather than to endpoint-related evidence as a whole. The distinction between noticing an endpoint and executing the associated action is also compatible with video-based titration research showing that observation, critique, and technique execution are separable assessment targets [10, 13]. Related video-based laboratory studies also indicate that observing or critiquing a technique and enacting that technique are distinguishable learning and assessment targets [23, 36].

A second discontinuity appeared between flow regulation and coordinated execution. C3 flow regulation was at least 3 for six students (23.1%, 11.0-42.1), whereas D4 buret-flask coordination was at least 3 for three students (11.5%, 4.0–29.0), an 11.6-point difference and a descriptive retention ratio of 0.50. Unlike the endpoint comparison, however, both skills started from a low proficiency base, indicating a broader difficulty with dynamic manipulation rather than a single sharp loss from a strong upstream skill. Reproducibility showed an even lower floor: G1 repetition reached the threshold for two students (7.7%, 2.1–24.1), and G2 procedural consistency for none (0.0%, 0.0-12.9), yielding a retention ratio of 0.00. Table 4 and Fig. 2 show these four pre-specified comparisons. The low repetition and consistency

outputs also support treating repeated enactment as a distinct performance dimension rather than inferring reproducibility from one apparently successful trial [2, 21].

The four comparisons therefore form two distinct analytical patterns. Endpoint recognition to stopping is a high-base discontinuity: the upstream behavior is common, but the downstream action is much less common. Flow regulation to coordination and repetition to consistency are low-base bottlenecks: both sides of the comparison are already weak, so the issue is not only retention but limited evidence of proficiency throughout the action sequence. Endpoint recognition to verification is the counter-pattern, with little loss between the two measures. These contrasts help identify where the observed performance structure changes most sharply without implying chronology or causality. The timestamps identify selected evidence locations rather than validated event onset and offset, and some item scores share fallback-derived structure; accordingly, the appropriate conclusion is that recognition and action are imperfectly coupled in the current assessment output, not that endpoint recognition causes stopping failure. Accordingly, the comparisons are used diagnostically, consistent with rubric-based direct assessment that links observable actions to specific feedback targets rather than treating a total score as sufficient evidence [8].

Table 4
Descriptive functional discontinuities across critical titration micro-skills

Functional comparison	Upstream proficiency n/N (%) [95% CI]	Downstream proficiency n/N (%) [95% CI]	Absolute change	Descriptive retention
Endpoint recognition → timely stopping	E2: 21/26 (80.8%) [62.1–91.5]	E3: 11/26 (42.3%) [25.5–61.1]	-38.5 pp	0.52
Endpoint recognition → endpoint verification	E2: 21/26 (80.8%) [62.1–91.5]	E4: 20/26 (76.9%) [57.9–89.0]	-3.9 pp	0.95
Flow regulation → coordinated execution	C3: 6/26 (23.1%) [11.0-42.1]	D4: 3/26 (11.5%) [4.0–29.0]	-11.6 pp	0.50
Repetition → procedural consistency	G1: 2/26 (7.7%) [2.1–24.1]	G2: 0/26 (0.0%) [0.0-12.9]	-7.7 pp	0.00

3.4 Aggregate-score compression and configurational non-equivalence

The central profile result emerged from students who shared exactly the same weighted overall score. Fourteen of the 26 participants (53.8%) belonged to a repeated exact-score group, creating 15 within-score pairs. Thirteen of these 15 pairs (86.7%) had a non-zero A-G profile distance. Equal totals therefore commonly represented unequal underlying performance configurations, and this was not limited to a single duplicated score group. The result demonstrates a many-to-one mapping from multidimensional

performance to the weighted total: different combinations of domain strengths and weaknesses can collapse to the same scalar outcome. Such compression is consequential in laboratory education because practical competence spans multiple forms of knowledge and action that may not be represented adequately by one compensatory total [1, 3].

The contrast was particularly clear at overall scores of 55.00 and 57.50. Each score occurred for four students, producing six within-score pairs, and all six pairs in each group were configurationally non-equivalent. At 55.00, the varying domains were B, D, E, and F, with a median MADD of 6.25 and a median weighted profile distance of 5.00. At 57.50, variation extended across A, B, C, D, and E, with a median MADD of 8.04 and a median weighted profile distance of 10.00. The 61.25 group contained two students with different B and E scores despite an identical total, whereas the 52.50 and 56.25 pairs had identical domain profiles. Thus, exact-score equivalence was sometimes genuine profile equivalence, but more often it was produced by compensation across domains. The presence of multiple pathways to the same total complements direct-assessment approaches that retain criterion-level performance information rather than relying only on holistic grades [6, 7].

The compensation can be quantified directly from the weighted profile distance. For an exact-score pair, positive and negative weighted domain differences cancel in the total, so one-half of the weighted L1 distance represents the amount of weighted advantage that is offset by disadvantage elsewhere. The median distance of 5.00 in the 55.00 group therefore corresponds to approximately 2.50 weighted score points of opposing contribution on each side of the profile. In the 57.50 group, the median distance of 10.00 corresponds to approximately 5.00 weighted points being offset in each direction. Across all 15 exact-score pairs, the median weighted distance of 7.50 implies a median compensation magnitude of 3.75 weighted points. This makes the compression effect interpretable: identical totals were not merely associated with small profile noise, but often required meaningful cross-domain offsets.

The near-score sensitivity analysis showed that the configurational result was not an artifact of exact ties. Of all 325 possible participant pairs, 200 (61.5%) were within five overall-score points. Among these near-score pairs, the median weighted profile distance was 6.51 [IQR 4.38–9.84] and the median MADD was 6.70 [4.46–8.93], indicating appreciable domain-level separation even when the overall scores were close. Table 5 summarizes the exact-score groups, and Fig. 3 visualizes how identical totals can map to different numbers of unique A-G profiles. Together, the exact-score and near-score analyses show that the aggregate total loses information in two ways: it can make distinct profiles numerically identical, and it can make substantively different profiles appear only slightly different. This pattern supports profile-preserving reporting for formative interpretation, consistent with work emphasizing authentic practical assessment and explicit feedback on laboratory technique [9, 13].

Table 5
Exact-score groups and hidden profile divergence

Exact score	n	Pairs	Diff. pairs	Med MADD	Max MADD	Med wL1	Varying domains
52.50	2	1	0/1	0.00	0.00	0.00	None
55.00	4	6	6/6	6.25	10.71	5.00	B, D, E, F
56.25	2	1	0/1	0.00	0.00	0.00	None
57.50	4	6	6/6	8.04	8.93	10.00	A, B, C, D, E
61.25	2	1	1/1	5.36	5.36	5.00	B, E

Cohort summary: 14/26 students belonged to repeated exact-score groups; 13/15 exact-score pairs (86.7%) had non-identical profiles; median weighted L1 distance among exact-score pairs = 7.50. For exact-score pairs, one-half of weighted L1 is the magnitude of weighted domain contribution that is offset by opposing differences elsewhere in the profile.

4. Discussion

4.1 Aggregate practical scores compress instructionally meaningful differences

The profile analysis changes the interpretation of practical assessment from score ranking to competence configuration. Because different domain patterns repeatedly collapsed to the same weighted total, the aggregate score functioned as a lossy summary rather than a complete representation of performance. This is consistent with the broader view that laboratory learning spans multiple forms of knowledge and action and should not be reduced to one undifferentiated outcome [1, 3]. The contribution here is more specific: the compensation analysis shows how multidimensional strengths and weaknesses can numerically cancel, producing apparent equivalence even when the underlying profiles remain educationally distinct. The practical consequence is that the same grade can correspond to different feedback priorities.

Aggregate scores remain useful for administrative reporting, but their interpretation should be bounded by the profile from which they were produced. Direct-assessment approaches such as digital badges, station-based practical examinations, rubric-based observation, and authentic laboratory evaluation have sought to make hands-on performance more visible [4–9]. The present findings add a reporting principle to that literature: preserving multidimensional measurement is not enough if the reporting layer immediately collapses those dimensions back into a single total. A more instructionally useful report would pair the total with a compact A-G profile and explicitly identify the domains that compensate for one another. This would allow feedback to distinguish between students who arrive at similar totals through different performance pathways.

4.2 Titration performance couples perception, action, and reproducibility

The domain and micro-skill results further indicate that titration competence is organized around coordination, not simply the possession of isolated procedural knowledge. The strongest asymmetry occurred when a relatively common perceptual behavior - endpoint recognition - was followed by a much less common action - timely stopping. The counter-pattern is equally informative: endpoint verification remained close to endpoint recognition, suggesting that endpoint-related visual evidence itself was not uniformly weak. What distinguished the stopping step was the need to translate perception into motor control at the same moment that titrant flow, flask movement, and apparatus handling had to be coordinated. This provides an empirically grounded rationale for treating perception-action coupling as a distinct target of practical instruction.

Reproducibility represents a second instructional target. The very low G1 and G2 outputs indicate that one successful-looking trial cannot automatically be interpreted as stable laboratory competence. Experimental practice requires students to repeat, compare, and refine procedures while controlling sources of variation, and laboratory-skills research similarly emphasizes deliberate practice, explicit criteria, and repeated enactment [2, 13, 21]. The current low-base bottleneck pattern therefore suggests that consistency should be assessed as its own performance dimension rather than inferred from first-pass completion. For titration teaching, this implies that feedback should distinguish what students can perceive, what they can control in the moment, and what they can reproduce across attempts. Research on students' laboratory goals and independent audits of basic laboratory technique likewise cautions against treating successful completion as evidence of stable or reproducible skill [28, 30, 31].

4.3 Implications for video-based assessment and limitations

The findings support a restrained role for automated video analytics. The practical value of the workflow is that it can organize long video records, locate candidate evidence, link evidence to rubric codes, and retain an audit trail, while the analytical value lies in exposing domain structure, bottlenecks, and compensating profiles that are difficult to see from a total score alone. That function is compatible with formative assessment because it can help instructors inspect domain profiles and identify feedback priorities. It is not equivalent to autonomous grading. The current measurement audit shows why: 96.46% of ratings were flagged for manual review, the fallback route dominated the implementation data, contextual-model use was associated with overall score, and several micro-skill vectors were duplicated by the scoring logic. Reporting these properties rather than hiding them is necessary for a defensible validity argument.

The study has important limitations. First, the participant-level sample comprised 26 students from one institutional context, so profile frequencies should not be generalized as population estimates. Second, the scores are evidence-derived outputs rather than uniformly human-validated summative ratings; this

limitation is handled by restricting the claims to diagnostic score structure and by providing a prespecified blinded double-rating protocol in Electronic Supplementary Material 1 (ESM 1) for any confirmatory or high-stakes reuse. Third, the timestamps identify selected frames or intervals rather than validated behavioral event onsets and offsets, so the functional comparisons are descriptive rather than chronological transitions. Fourth, the domain weights belong to the operational scoring architecture and should not be interpreted as empirically optimized importance weights. Finally, video evidence is sensitive to camera placement and visibility; in a human-rated implementation, genuinely non-observable evidence should be coded as NA rather than converted to low performance.

Within these limits, the study identifies a clear assessment problem with direct implications for chemistry teaching: practical scores that look equivalent can be educationally non-equivalent. A defensible reporting system should therefore retain the total when required for grading, but pair it with a compact domain profile and a small number of evidence-linked improvement priorities. This approach uses video and automation to increase auditability and feedback specificity while keeping the educational judgment anchored to interpretable evidence and human review. The same reporting principle is transferable to other performance-based STEM and professional-learning settings in which digital evidence is used to support judgments about complex practical competence [14, 15].

5. Conclusion

This study examined what aggregate practical scores reveal and conceal in a video-derived assessment of undergraduate titration performance. The evidence-derived dataset showed a structured domain pattern, a pronounced descriptive gap between endpoint recognition and timely stopping, and substantial configurational non-equivalence among students with identical overall scores. The central implication is that the same practical total can represent different combinations of laboratory strengths and weaknesses and therefore different learning needs. Video evidence is valuable when it preserves the connection between rubric decisions and observable action, but automated processing should remain an evidence-support layer rather than an autonomous assessor. The study therefore contributes a practical assessment argument rather than an automation claim: multidimensional performance profiles should be retained alongside aggregate scores, and measurement-integrity information should determine the level of interpretation attached to any video-derived rating. For summative grading or cross-context generalization, independent human double-rating and replication remain necessary. More broadly, the study illustrates how digital assessment can support practical learning when technology increases evidence traceability without replacing the interpretive role of educators.

Declarations

Ethics approval and consent to participate

This study was reviewed by the Research Ethics Committee (Komisi Etik Penelitian), Universitas Negeri Semarang, Indonesia, under reference No. B/8282/UN37.3.1/KEP.00/2026, and was classified by the committee as Exempted Review. All procedures involving human participants were conducted in accordance with the applicable ethical guidelines and regulations of the Research Ethics Committee (Komisi Etik Penelitian), Universitas Negeri Semarang, Indonesia. Participation was voluntary, and informed consent to participate was obtained from all participants before their practical-performance video records and derived data were used for research purposes.

Consent for publication

Informed consent to participate was obtained from all participants. Informed consent for publication of identifiable images and other visual material was obtained from all participants whose identifiable visual material appears in the manuscript.

Competing interests

The authors declare that they have no competing financial or non-financial interests relevant to the work reported in this manuscript.

Use of artificial intelligence and automated tools

OpenCV-based computer-vision procedures were used to support frame sampling, region-of-interest processing, motion/color feature extraction, and localization of candidate evidence within the titration videos. Google Gemini 3.1 Flash-Lite was used selectively during the 2026 video-analysis workflow to assist contextual interpretation of selected evidence for rubric alignment.

Funding

This research was funded by Universitas Negeri Semarang (UNNES) through DPA FMIPA UNNES research funding (96.7.5/UN37/PPK.04/2026), under the Research Assignment Agreement No. 139.03.2.693449/2026. The funder had no role in the interpretation of findings or the decision to submit the manuscript for publication.

Author Contribution

ATP: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Validation, Writing – review and editing.DL: Investigation, Data curation, Validation, Writing – original draft, Writing – review and editing.SHD: Investigation, Data curation, Validation, Interpretation of findings,

Writing – review and editing.DGR: Methodology, Software, Formal analysis, Visualization, Data curation, Writing – original draft, Writing – review and editing.

Acknowledgement

This research was funded by Universitas Negeri Semarang (UNNES) through DPA FMIPA UNNES research funding (96.7.5/UN37/PPK.04/2026), under the Research Assignment Agreement No. 139.03.2.693449/2026. The funder had no role in the interpretation of findings or the decision to submit the manuscript for publication.

Data Availability

The datasets generated during and/or analysed during the current study are not publicly available due to participant privacy, informed consent, and institutional data protection requirements because the raw student videos contain potentially identifiable information, but are available from the corresponding author on reasonable request.

References

1. Agustian HY, Finne LT, Jørgensen JT, Pedersen MI, Christiansen FV, Gammelgaard B, Nielsen JA. Learning outcomes of university chemistry teaching in laboratories: a systematic review of empirical literature. Rev Educ. 2022;10(2):e3360. https://doi.org/10.1002/rev3.3360.
2. Seery MK. Establishing the laboratory as the place to learn how to do chemistry. J Chem Educ. 2020;97(6):1511–4. https://doi.org/10.1021/acs.jchemed.9b00764.
3. Agustian HY. Considering the hexad of learning domains in the laboratory to address the overlooked aspects of chemistry education and fragmentary approach to assessment of student learning. Chem Educ Res Pract. 2022;23:518–30. https://doi.org/10.1039/D1RP00271F.
4. Towns M, Harwood CJ, Robertshaw MB, Fish J, O'Shea K. The digital pipetting badge: a method to improve student hands-on laboratory skills. J Chem Educ. 2015;92(12):2038–44. https://doi.org/10.1021/acs.jchemed.5b00464.
5. Hensiek S, DeKorver BK, Harwood CJ, Fish J, O'Shea K, Towns M. Improving and assessing student hands-on laboratory skills through digital badging. J Chem Educ. 2016;93(11):1847–54. https://doi.org/10.1021/acs.jchemed.6b00234.
6. Seery MK, Agustian HY, Doidge ED, Kucharski MM, O'Connor HM, Price A. Developing laboratory skills by incorporating peer-review and digital badges. Chem Educ Res Pract. 2017;18(3):403–19. https://doi.org/10.1039/C7RP00003K.
7. Hancock LM, Hollamby MJ. Assessing the practical skills of undergraduates: the evolution of a station-based practical exam. J Chem Educ. 2020;97(4):972–9. https://doi.org/10.1021/acs.jchemed.9b00733.

8. Harwood CJ, Hewett S, Towns MH. Rubrics for assessing hands-on laboratory skills. J Chem Educ. 2020;97(7):2033–5. https://doi.org/10.1021/acs.jchemed.0c00200.
9. Veale CGL, Jeena V, Sithebe S. Prioritizing the development of experimental skills and scientific reasoning: a model for authentic evaluation of laboratory performance in large organic chemistry classes. J Chem Educ. 2020;97(3):675–80. https://doi.org/10.1021/acs.jchemed.9b00703.
10. Lau PN. Enhancing formative and self-assessment with video playback to improve critique skills in a titration laboratory. Chem Educ Res Pract. 2020;21:178–88. https://doi.org/10.1039/C9RP00056A.
11. Loughlin WA, Cresswell SL. Integration of interactive laboratory videos into teaching upper-undergraduate chemical laboratory techniques. J Chem Educ. 2021;98(9):2870–80. https://doi.org/10.1021/acs.jchemed.0c01493.
12. Jacobsen FE. Use of student-generated technique videos to increase laboratory skills in an online general chemistry laboratory. J Chem Educ. 2023;100(4):1460–5. https://doi.org/10.1021/acs.jchemed.2c00867.
13. Tenney SM, Russell AA, Casey JR. Critiquing lab technique videos prior to class: can it improve demonstrated technique? Chem Educ Res Pract. 2024;25:862–76. https://doi.org/10.1039/D4RP00036F.
14. Sweetman S, Sabour A, Eljamal A, Hendawi A, Jones L. Design and early validation of a digital assessment for science perceptions (DigiDAST). Discover Educ. 2026;5:338. https://doi.org/10.1007/s44217-026-01334-y.
15. Kılıç M, Cüvitoğlu G, Polan Ş, et al. A comprehensive item pretesting study to develop an automated short-answer grading platform for Turkish inputs. Discover Educ. 2026;5:755. https://doi.org/10.1007/s44217-026-02033-4.
16. Lewis SE. Considerations on validity for studies using quantitative data in chemistry education research and practice. Chem Educ Res Pract. 2022;23:764–7. https://doi.org/10.1039/D2RP90009B.
17. Lazenby K, Tenney K, Marcroft TA, Komperda R. Practices in instrument use and development in chemistry education research and practice 2010–2021. Chem Educ Res Pract. 2023;24:882–95. https://doi.org/10.1039/D2RP00275B.
18. Kaya ON, Kaya Z. The impact of co-design-based formative assessment practices on preservice science teachers' understanding of chemical concepts in a general chemistry laboratory course. Chem Educ Res Pract. 2024;25:996–1017. https://doi.org/10.1039/D3RP00225J.
19. Shi L, Wang Y, Mitchell-Jones JK, Stains M. Why do we assess students? Investigating general chemistry instructors' conceptions of assessment purposes and their relationships to assessment practices. Chem Educ Res Pract. 2024;25:1159–74. https://doi.org/10.1039/D4RP00143A.
20. Seery MK, Agustian HY, Christiansen FV, Gammelgaard B, Malm RH. 10 guiding principles for learning in the laboratory. Chem Educ Res Pract. 2024;25:383–402. https://doi.org/10.1039/D3RP00245D.
21. Accettone SLW, DeFrancesco C, King CA, Lariviere MK. Laboratory skills assignments as a teaching tool to develop undergraduate chemistry students' conceptual understanding of practical laboratory

skills. J Chem Educ. 2023;100(3):1138–48. https://doi.org/10.1021/acs.jchemed.2c00710.

22. Ramesh VB, Seery MK, Cole R. Assessing science practices in undergraduate chemistry laboratories: why we need to do better—and how we should. J Chem Educ. 2026;103(1):27–35. https://doi.org/10.1021/acs.jchemed.5c00988.
23. Altowaiji S, Haddadin R, Campos P, Sorn S, Gonzalez L, Villafañe SM, Groves MN. Measuring the effectiveness of online preparation videos and questions in the second semester general chemistry laboratory. Chem Educ Res Pract. 2021;22:616–25. https://doi.org/10.1039/D0RP00240B.
24. Hennah N, Newton S, Seery MK. A holistic framework for developing purposeful practical work. Chem Educ Res Pract. 2022;23(3):582–98. https://doi.org/10.1039/D1RP00168J.
25. Keen C, Sevian H. Qualifying domains of student struggle in undergraduate general chemistry laboratory. Chem Educ Res Pract. 2022;23(1):12–37. https://doi.org/10.1039/D1RP00051A.
26. DeKorver BK, Towns MH. General chemistry students' goals for chemistry laboratory coursework. J Chem Educ. 2015;92(12):2031–7. https://doi.org/10.1021/acs.jchemed.5b00463.
27. Bretz SL, Fay M, Bruck LB, Towns MH. What faculty interviews reveal about meaningful learning in the undergraduate chemistry laboratory. J Chem Educ. 2013;90(3):281–8. https://doi.org/10.1021/ed300384r.
28. Sanchez JM. Are basic laboratory skills adequately acquired by undergraduate science students? How control quality methodologies applied to laboratory lessons may help us to find the answer. Anal Bioanal Chem. 2022;414(12):3551–9. https://doi.org/10.1007/s00216-022-03992-x.
29. Stowe RL, Scharlott LJ, Ralph VR, Becker NM, Cooper MM. You are what you assess: the case for emphasizing chemistry on chemistry assessments. J Chem Educ. 2021;98(8):2490–5. https://doi.org/10.1021/acs.jchemed.1c00532.
30. Santos-Díaz S, Hensiek S, Owings T, Towns MH. Survey of undergraduate students' goals and achievement strategies for laboratory coursework. J Chem Educ. 2019;96(5):850–6. https://doi.org/10.1021/acs.jchemed.8b00984.
31. DeKorver BK, Towns MH. Upper-level undergraduate chemistry students' goals for their laboratory coursework. J Res Sci Teach. 2016;53(8):1198–215. https://doi.org/10.1002/tea.21326.
32. Govindarajoo G, Lee JY, Emenike ME. Proof of concept for a thin-layer chromatography digital badge assignment within a laboratory practical exam for a nonchemistry majors' organic chemistry lab. J Chem Educ. 2021;98(9):2775–85. https://doi.org/10.1021/acs.jchemed.1c00598.
33. Box MC, Dunnagan CL, Hirsh LAS, Cherry CR, Christianson KA, Gibson RJ, Wolfe MI, Gallardo-Williams MT. Qualitative and quantitative evaluation of three types of student-generated videos as instructional support in organic chemistry laboratories. J Chem Educ. 2017;94(2):164–70. https://doi.org/10.1021/acs.jchemed.6b00451.
34. Blackburn RAR, Villa-Marcos B, Williams DP. Preparing students for practical sessions using laboratory simulation software. J Chem Educ. 2019;96(1):153–8. https://doi.org/10.1021/acs.jchemed.8b00549.

35. Reid N, Shah I. The role of laboratory work in university chemistry. Chem Educ Res Pract. 2007;8(2):172–85. https://doi.org/10.1039/B5RP90026C.
36. Erdmann MA, March JL. Video reports as a novel alternate assessment in the undergraduate chemistry laboratory. Chem Educ Res Pract. 2014;15(4):650–7. https://doi.org/10.1039/C4RP00107A.

Figures

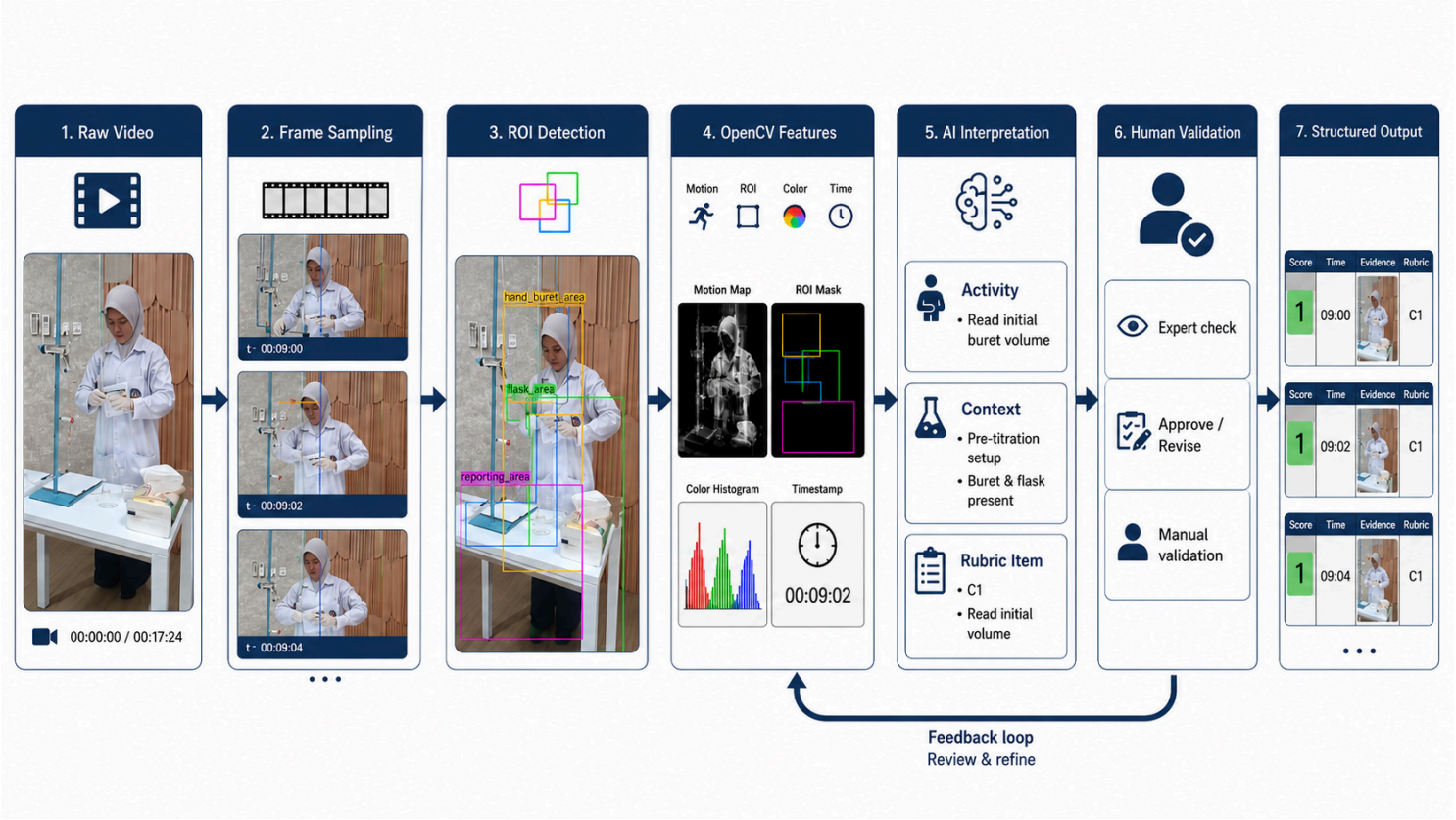


Figure 1

AI-assisted titration video workflow from raw video, frame sampling, and visual evidence extraction to human validation and structured rubric output

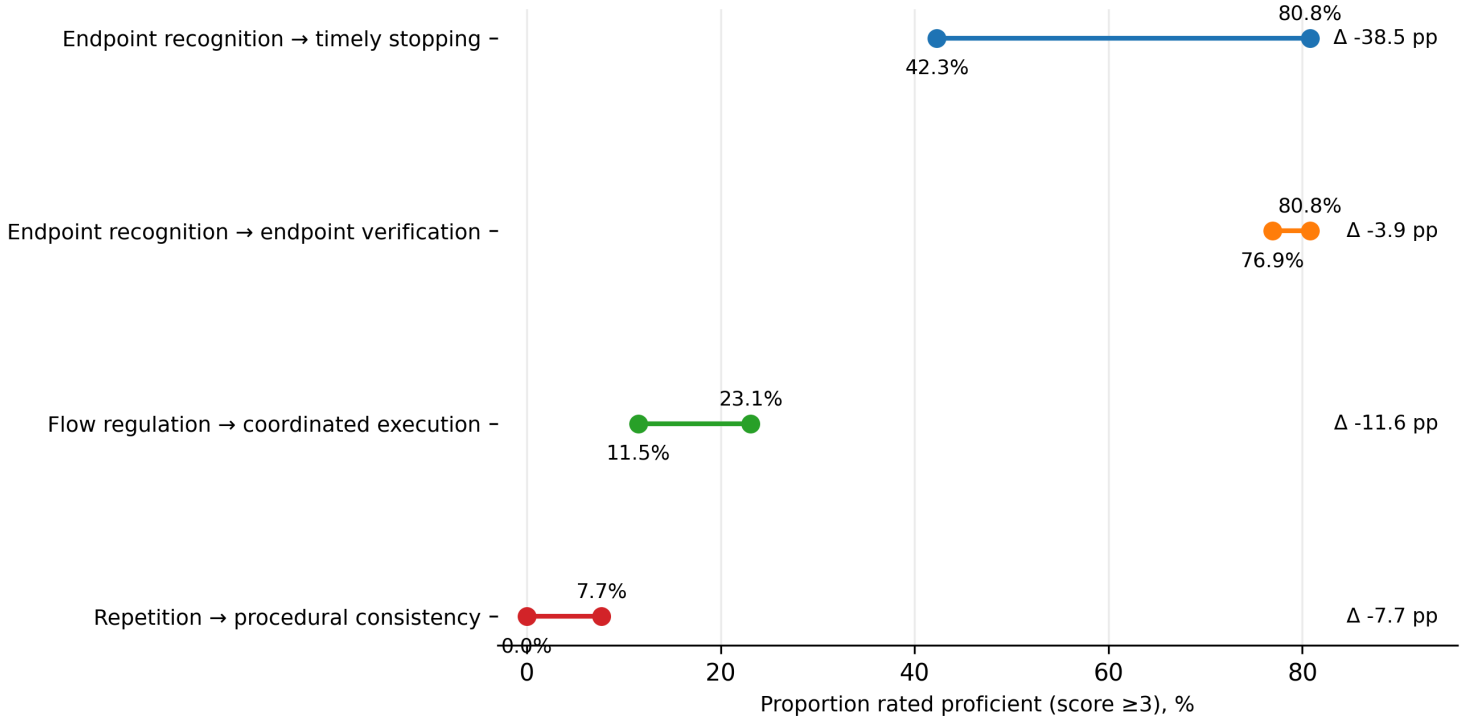


Figure 2

Functional decoupling across selected titration micro-skills; the largest descriptive gap was between endpoint recognition and timely stopping

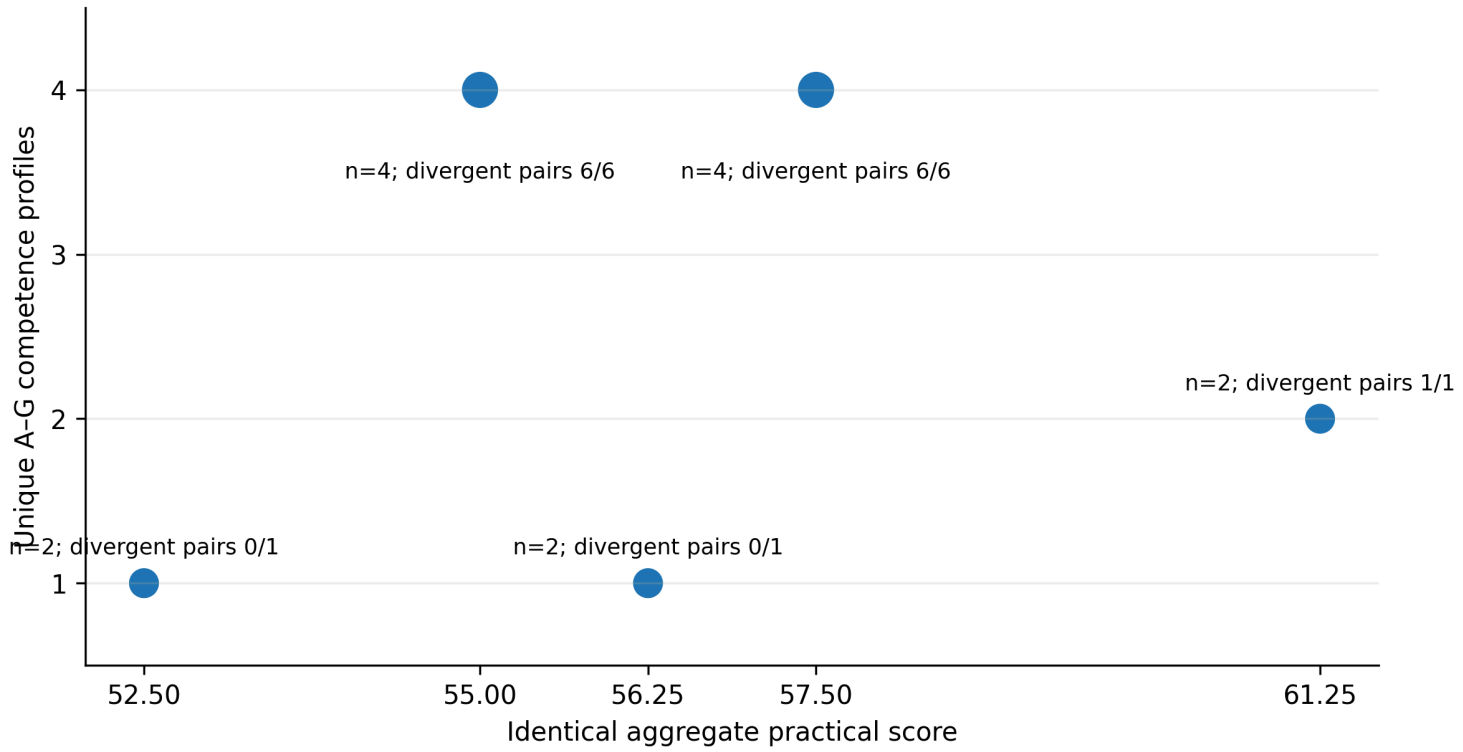


Figure 3

Identical aggregate practical scores can conceal multiple distinct A–G performance configurations

Supplementary Files

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

- Table1AnalyticalArchitectureofTitrationPerformance.csv
- Table2MeasurementStatusAudit.csv
- Table3DomainLevelTitrationPerformance.csv
- Table4FunctionalDiscontinuities.csv
- Table5ExactScoreProfileDivergence.csv