

Assessing Undergraduate Digital Forensic Competencies Through Incident Investigation Assignments: Evidence from Log-Based Forensic Analysis

Rizqi Ramadhan, Adzanil Rachmadi Putra, Purnama Anaking, Dudung Adijono

Abstract—Assessing complex investigative competence in a way that is both authentic and reliable remains an open measurement problem in forensic and cybersecurity education, and not only in resource-constrained settings. This paper reports a single-cohort pilot of a log-based incident-investigation assignment, built from multiple case variants, that was used to explore this problem in an undergraduate Computer Forensics course. Ten student groups each received their own synthetic evidence package describing a different multi-stage intrusion, with one or two red herrings planted in each case. Each group submitted a forensic report, which was scored with a six-dimension competency rubric aligned to NIST SP 800-86, the NICE Cybersecurity Workforce Framework, and Bloom's revised taxonomy. The analytic sample consists of the ten group submissions (about 40 students were enrolled). Since each case variant was completed by only one group, case difficulty and group ability are confounded by design, so case-level patterns are treated as preliminary and hypothesis-generating, not as proof that some scenarios are harder than others. With this caveat in mind, the groups performed strongly and consistently on evidence identification, multi-source correlation, and analytical reasoning. Professional reporting was the weakest dimension. In four of the ten groups, the core analysis was clearly stronger than the report that communicated it, a pattern we call the synthesis-and-reporting gap. Critical evaluation was high on average but bimodal, as two groups missed their red herring entirely. We place the instrument within Kane's argument-based validity framework, present it as a pilot assessment design rather than a validated measurement instrument, and describe the multi-rater, multi-cohort work needed before stronger claims about validity or reliability can be made.

Index Terms—authentic assessment, competency-based education, digital forensics education, incident response, log analysis, validity argument

I. INTRODUCTION

Assessing complex professional competence is a long-standing measurement problem. Tasks that are realistic enough to draw out authentic reasoning tend to be open-ended,

and open-ended tasks are hard to score in the standardized way that reliable, comparable measurement needs. Tasks that are structured enough to score reliably, on the other hand, often turn into recall exercises and lose the authenticity they were supposed to capture. This tension between authenticity and reliability is well documented in the validity literature. Kane's argument-based framework does not treat validity as a fixed property of an instrument. Instead, it describes a chain of inferences (scoring, generalization, extrapolation, and implication), each of which needs its own supporting evidence. Under this view, a task can be highly authentic and still lack the evidence required to support broader claims about what its scores mean [9].

Digital forensic investigation is a clear example of this tension. Reconstructing what happened on a system from the traces it leaves behind is an authentic and valuable professional task. It is also open-ended, cognitively demanding, and hard to score consistently. The investigator has to separate signal from operational noise, correlate events across different sources, and infer intent from incomplete evidence, and no two investigations unfold in the same way. This is not a problem specific to one country's curriculum. A recent UK-wide survey of forensic science practitioners found that fewer than half of respondents reported using observed competency testing, in which a qualified assessor watches someone actually carry out the task. This method lagged well behind simpler assessment strategies, even though there is growing agreement that authentic, competency-based assessment is what the workplace needs [10]. Hands-on cybersecurity assessment has the same gap. Capture-the-Flag (CTF) exercises are now a common way to teach security skills and have been mapped explicitly onto the NICE Cybersecurity Workforce Framework [11], and comparative studies across institutions report gains in self-efficacy and applied skill [12]. However, these studies usually evaluate engagement and learning gain. They do not show that CTF scores are a valid and reliable measure of competence in Kane's sense. Digital forensics education inherits the same open problem. Authentic, log-based tasks are attractive for teaching, but the evidence needed to treat their scores as validated measures of competence is rarely collected.

¹Manuscript received 26 June 2026. R. Ramadhan, A. R. Putra, and P. Anaking are with the Department of Information Systems, Telkom University, Surabaya, Indonesia (corresponding author: R. Ramadhan; e-mail: rizqir@telkomuniversity.ac.id). D. Adijono is with the Intermediate Staff and Leadership School (Sespimmen), Sespim Lemdiklat Polri, Lembang, West Bandung Regency, Indonesia.

This paper examines that problem in the context of undergraduate Information Systems education, where it is felt strongly. Computer Forensics is often a young elective course, and the students who take it usually have little prior exposure to the command line or to security operations. Instruction therefore has to build procedural fluency from a low starting point, while assessment still has to show whether students can actually carry out an investigation, not just define terms such as *chain of custody* or *indicator of compromise* (IoC). The course reported here is Computer Forensics (*Forensika Komputer*) at Telkom University, Surabaya, Indonesia. The Indonesian context provides the setting for the pilot but is not the central problem of the paper.

We report the design of an assignment-based instrument that addresses this problem, along with its first application in a single cohort. In the assignment, each student group acts as a Computer Security Incident Response Team (CSIRT). The team receives a package of evidence from a fictitious client and has to investigate, reconstruct, and report on a security incident. Three design features are meant to address the validity problem directly. First, every group receives a different evidence package built around a different attack scenario. This discourages collusion and creates a natural spread of case difficulty that can be studied in future work with larger samples. Second, each package contains deliberate red herrings, that is, benign activity that looks like an attack on the surface, so that critical evaluation can be measured directly instead of being inferred. Third, scoring follows an explicit rubric anchored to ground truth and aligned with recognized forensic and workforce frameworks. This allows performance to be broken down by competency dimension, with a clear account of the evidence behind each score.

This paper makes three contributions. The first is a reusable design for a variant-based, log-grounded forensic assessment instrument, including how it is built to resist collusion and how its scoring is anchored to ground truth. The second is a competency model that maps the rubric criteria to NIST SP 800-86 activities, NICE framework tasks, and Bloom's revised taxonomy, together with explicit performance-level descriptors and a rule for partial credit. The third is a first empirical application in a single cohort, reported as a pilot case study that reveals preliminary, hypothesis-generating patterns in group performance. In line with Kane's framework [9], we treat this pilot as evidence toward the instrument's validity, not as proof of it. We also state what further evidence the argument still needs, namely additional cohorts, additional independent raters, and more than one group per case.

II. RELATED WORK AND BACKGROUND

A. Digital forensic process models and competencies

Modern digital forensics takes its process vocabulary from early consensus models. The DFRWS road map [2] and the event-based investigation framework of Carrier and Spafford [3] divide an investigation into phases (identification,

preservation, examination, analysis, and presentation), and these phases still form the backbone of both practice and teaching. For incident response in particular, NIST Special Publication 800-86 [4] brings forensic technique into the response lifecycle and is widely used as a reference for what an investigator should be able to do with data sources such as logs. At the workforce level, the NICE Cybersecurity Workforce Framework [5] describes these abilities as discrete task, knowledge, and skill statements, which makes it a convenient anchor for designing competency-based assessment.

B. Validity theory and authentic, competency-based assessment

Authentic assessment asks students to perform tasks that resemble real professional work and judges them on observable performance [1]. Competency-based assessment goes a step further by breaking performance into named abilities, each scored against explicit criteria. Both approaches rely on constructive alignment [6], meaning that learning outcomes, learning activities, and assessment should all target the same cognitive operations. Bloom's revised taxonomy [7] provides the vocabulary to check this alignment. Kane's argument-based approach to validity [9] changes the goal of instrument design. Instead of asking whether a task is authentic in the abstract, it asks what chain of inferences connects an observed score to the claim that score is supposed to support, and what evidence backs each link. We use this framing explicitly in Section VI to state which inferences the present pilot supports and which it does not yet support. Hamnett and Matthews' UK-wide practitioner survey [10] gives an empirical picture of the resulting gap in the field. Even in an established discipline with professional accreditation bodies, observed and authentic competency testing is still a minority practice.

C. Hands-on and Capture-the-Flag-style cybersecurity assessment

Digital forensics education is part of a wider shift toward hands-on cybersecurity assessment. Capture-the-Flag (CTF) exercises have been mapped explicitly onto NICE task and skill statements, which gives educators a competency vocabulary similar to the one used in this paper [11]. Comparative deployments across institutions and countries report that CTF-style exercises improve self-reported skill, confidence, and career interest [12], and problem-based learning offers the broader pedagogical rationale for open, ill-structured team tasks [8]. These studies show that hands-on, scenario-based exercises engage students and are seen as effective. In general, though, they do not establish validity or reliability at the level of scores in Kane's sense. This is the gap the present instrument tries to narrow by combining scenario-based tasks with scoring anchored to ground truth.

D. Log analysis and synthetic evidence as a teaching and assessment substrate

Logs are a practical basis for forensic teaching and assessment because they are realistic, cheap to distribute, and can be analyzed with free and widely available tools. A credible investigation can be carried out using command-line text utilities alone, which lowers the barrier for beginners while still exercising authentic skills. Using synthetic, deliberately constructed security logs and telemetry for training and evaluation is an active and growing area, covering generative and hybrid methods for producing realistic attack traces at scale. The shared motivation behind this work is that synthetic construction gives the designer full control over ground truth and makes it possible to plant evidence and distractors, which real-world capture corpora cannot offer. The instrument described here relies directly on this property, although at classroom scale rather than dataset-generation scale.

E. Rater reliability in rubric-based performance assessment

Performance assessment scored with a rubric is open to rater subjectivity, and this weakness is not the same across all criteria. Studies of rubric-based scoring in higher education report that inter-rater reliability, measured with statistics such as Cohen's kappa or the intraclass correlation coefficient (ICC), tends to be higher for criteria with clear, checkable referents and lower for criteria that require qualitative judgment of writing or reasoning quality [13]. This distinction matters for the instrument reported here. Its six dimensions combine criteria that can be checked against ground truth (evidence identification, timeline correlation, and evidence integrity) with criteria that require judgment (analytical reasoning and professional reporting). We come back to this difference when discussing single-rater risk in Section IV-B and again as a limitation in Section VI-C.

III. COURSE CONTEXT AND INSTRUMENT DESIGN

The instrument was developed in three stages, following a process that is typical for building assessment instruments. The first stage specified the target competencies and their theoretical anchors, as described in Section IV. The second stage built the task materials, namely the case variants and evidence packages, so that they would draw out those competencies under controlled and verifiable ground truth. The third stage was a first application of the instrument in a single cohort, reported as a pilot case study in Section V. This section documents the second stage, that is, the fixed technical template applied in the same way to all ten case variants, regardless of the specific scenario or student group.

A. Course context

The instrument was deployed in Computer Forensics (*Forensika Komputer*), an undergraduate course in the Information Systems program at Telkom University,

Surabaya. The course starts with metadata forensics, moves on to mobile and server-side evidence, and ends with server log analysis and network traffic forensics. The investigation assignment is placed after the server log-analysis unit, so by that point students have already been taught about log sources, the investigation lifecycle, chain-of-custody concepts, and basic command-line parsing. Most students enter the course with no prior experience analyzing a server log.

B. Task structure and deliverables

The assignment is a take-home team investigation that runs over fourteen days. Each group receives an evidence package and a one-line complaint from the client saying only that a security incident is suspected. The nature, vector, and impact of the incident are not revealed and must be discovered by the group. Groups submit a forensic report with a professional structure: an executive summary, case background, methodology, an attack timeline with a diagram, technical analysis for each source, an IoC table, root-cause analysis, remediation recommendations, a conclusion, and references. Along with the report, they submit a technical appendix containing the commands and queries used, a CSV version of the IoC table, the timeline diagram, and a disclosure of any AI tool use (Section VI-D). Each member also writes a one-page individual reflection. Three checkpoints, on days three, seven, and ten, keep groups making steady progress and give formative feedback, but they do not count toward the summative score.

Tooling is kept minimal and free on purpose. Students use command-line text utilities (`grep`, `awk`, `sort`, `uniq`, `wc`), a spreadsheet for statistics and timeline building, and Wireshark's stream-following feature for the packet-capture evidence. More advanced platforms such as the ELK stack are allowed but never required, so that authentic analysis does not depend on expensive infrastructure.

C. Variant-based evidence packages

To assess several groups while discouraging collusion, we built ten distinct evidence packages. Each package encodes a different attack scenario, with its own actor names, attacker and server addresses, incident dates, and independently randomized benign log content, so no two packages share identical log lines. Each package includes the relevant parts of an authentication log, a web-access log, a system log, and a shell history. It also contains a packet capture with planted evidence, a case-narrative card, and a hash manifest for checking integrity. Each log file is roughly 150–280 lines long and mixes baseline traffic, operational noise, the embedded attack sequence, and one or two red herrings. Table I summarizes the ten variants and their ground-truth characteristics.

TABLE I
GROUND-TRUTH SUMMARY OF THE TEN CASE VARIANTS

ID	Scenario	Primary attack vector / technique	Principal IoC class	A priori difficulty
C1	Web defacement	SSH brute force + web shell upload	Failed-auth burst, web-shell GET/POST	Low–Med

ID	Scenario	Primary attack vector / technique	Principal IoC class	A priori difficulty
C2	Database breach	SQL injection → database dump	Anomalous query strings, large response	Med
C3	Lateral movement	Spear-phishing macro → internal pivot	Macro process spawn, new auth sessions	Med–High
C4	Insider data theft	Privileged USB exfiltration	USB mount, bulk file reads, log clear	Low–Med
C5	Resource abuse	Cryptominer via vulnerable plugin	Outbound pool conn., CPU/cron anomaly	Med
C6	Ransomware	RDP brute force → encryption	RDP auth burst, mass file rename	Med
C7	Supply-chain	Stolen vendor VPN credentials	Off-hours VPN login, unusual subnet	High
C8	Covert exfiltration	DNS tunneling	High-entropy DNS queries, volume spike	High
C9	Account takeover	Credential stuffing + reset abuse	Distributed login attempts, reset chain	Med
C10	CMS compromise	Joomla defacement + persistent backdoor	Admin path probing, backdoor callbacks	Med–High

D. Embedded red herrings

Every package contains at least one red herring. This is a benign sequence of events designed to look like malicious activity, for example a backup job that reads many files at once, an automated scanner running from an internal monitoring host, or an administrator doing legitimate maintenance outside working hours. A group that labels every anomaly as an attack has not shown forensic judgment. Competence is shown by correctly ruling out the benign event and explaining why. The rubric gives a bonus for identifying a red herring and justifying its dismissal, which turns an otherwise hidden discriminator into a scored competency (Section IV-C).

IV. COMPETENCY ASSESSMENT FRAMEWORK

The scoring criteria were grouped into six competency dimensions. Each dimension is linked to one or more NIST SP 800-86 activities and NICE task statements and is placed on Bloom's revised taxonomy. Table II shows this mapping. The five core dimensions are worth 95 base points in total, and a critical-evaluation bonus of up to 5 points is added on top. The maximum possible score is therefore 100 (95 base points plus 5 bonus points), not 105. The rest of this section covers the unit of analysis (IV-A), the scoring procedure and the uneven risk of rater bias across dimensions (IV-B), the performance-level descriptors and partial-credit rule (IV-C), and the position on ethics and data use (IV-D).

TABLE II
RUBRIC CRITERIA MAPPED TO COMPETENCY DIMENSIONS, FRAMEWORKS, AND COGNITIVE LEVEL

Competency dimension	Points	Rubric criteria (constituent)	Framework anchor	Bloom level
Evidence identification (IoC extraction)	20	Completeness of IoCs: IPs, users, files, hashes, timestamps	SP 800-86 examination, NICE collect/analyze	Apply
Timeline & multi-source correlation	20	Timeline accuracy, correct cross-source ordering	SP 800-86 analysis, NICE analyze	Analyze
Analytical reasoning (vector & root cause)	20	Depth of analysis, vector and root-cause inference	SP 800-86 analysis	Analyze–Evaluate
Evidence integrity & chain of custody	10	Hash verification, traceability of each finding to source	SP 800-86 preservation	Apply
Professional reporting & communication	25	Report structure, recommendations, technical appendix	SP 800-86 reporting, NICE oversee/communicate	Create
Critical evaluation (red-herring discrimination)	+5 bonus	Correct dismissal of benign look-alike with justification	SP 800-86 analysis	Evaluate

A. Unit of analysis and participants

The instrument was used in one section of the Computer Forensics course. About 40 students were enrolled, and they formed ten investigation groups of three to four members each. Each group was assigned one of the ten case variants listed in Table I. The analytic sample for this study is the ten group-level submissions (N = 10), consisting of the forensic report, technical appendix, and individual reflections produced by each group. Following the recommendation that group assignments be analyzed at the level at which they were produced, all results in Section V are reported and interpreted as properties of group submissions. This paper makes no claim

about the competence of any individual student. The three checkpoint submissions gave visibility into the process but were not graded summatively and are not part of the analytic sample. Because every group worked on a different case with a known ground truth, each report could be scored not only on the quality of its presentation but also on its correctness against the planted evidence. This is what allows competence to be measured rather than simply judged.

B. Scoring procedure and rater-bias risk

All ten reports were scored by the course instructor, who acted as the only rater, using the rubric in Table II. Scoring by a single rater carries a known risk of subjective bias, and that

risk differs across the six dimensions. Three dimensions (evidence identification, timeline correlation, and evidence integrity) can be checked against ground truth. For these, the score reflects the share of ground-truth elements, such as specific IoCs or timeline events, that a group recovered and attributed to the correct source, which leaves much less room for subjective judgment. The other two scored dimensions, analytical reasoning and professional reporting, require qualitative judgment of argument quality and writing. Here, single-rater bias is a real and more serious concern. The literature on rubric-based rater reliability shows this same pattern, with higher inter-rater agreement on checkable criteria than on judged ones [13]. No second, independent scoring pass was carried out in this pilot, so we cannot report any inter-rater reliability statistic such as Cohen's kappa or ICC. We state this as a limitation in Section VI-C. Adding a second independent rater and reporting an agreement statistic is a

necessary next step before the analytical-reasoning and professional-reporting scores can be treated as reliable measurements and not just the impressions of one rater.

C. Performance-level descriptors and partial credit

Table III shows the generic four-band descriptors used to award partial credit. For the three dimensions that can be checked against ground truth, the score equals the fraction of ground-truth elements that were recovered and attributed to the correct source, multiplied by the points allocated to that dimension. For the two judged dimensions, the report is compared holistically with the four-band descriptors and receives the matching fraction of the dimension's points. The critical-evaluation bonus is binary. A group receives the full 5 points if its report identifies the embedded red herring and clearly explains why it is benign, and 0 points otherwise. No partial credit is given for this criterion.

TABLE III
GENERIC PERFORMANCE-LEVEL DESCRIPTORS USED FOR PARTIAL CREDIT

Band	Ground-truth-checkable dimensions (IoC, timeline, integrity)	Judged dimensions (analytical reasoning, reporting)
76–100% (Exemplary)	Recovers nearly all ground-truth elements, correctly attributed to source.	Vector and root cause fully and correctly inferred. Report is well structured, evidence-traceable, and actionable throughout.
51–75% (Proficient)	Recovers the majority of ground-truth elements, with minor omissions or ordering errors.	Vector correctly identified, but root-cause reasoning is partial. Report is structured, but some sections are thin or under-evidenced.
26–50% (Developing)	Recovers roughly half of ground-truth elements, with some misattribution or sequencing errors.	Analysis is largely descriptive rather than causal. Report is missing key sections or recommendations.
0–25% (Beginning)	Recovers few ground-truth elements, or reports significant false positives.	No coherent vector/root-cause argument. Report lacks structure or traceability to evidence.

D. Ethics and data use

The assignment is a regular, credit-bearing part of the Computer Forensics course. The reports, technical appendices, and individual reflections analyzed here are routine assessment artifacts and were not collected through a separate research intervention. Students were not asked to do any additional task for research purposes. At the start of the semester, the course syllabus informed students that anonymized and aggregated assignment results might be analyzed and reported for teaching improvement and scholarly purposes. It also stated that any group could ask for its submission to be excluded from such reporting without any effect on its grade. No group made this request. Because one of the authors was both the course instructor and a study author, all grading was completed and finalized in the university's academic system before any material was selected for analysis. The analysis used only anonymized group codes (C1–C10), and student names and identification numbers were removed. No student or group can be identified from any figure or table in this paper. Since the study uses secondary, anonymized assessment data rather than identifiable human-subjects data, it counts as routine course evaluation carried out under Telkom University's academic-integrity and personal-data-protection policies [14], [15]. Formal ethics-committee approval was therefore not required.

V. RESULTS

All results below describe the ten group submissions (N = 10). With a sample this small, results are reported

descriptively using means, standard deviations, and exact frequencies. No inferential statistics are computed or implied.

A. Performance by competency dimension

Fig. 1 shows the mean normalized scores for the ten group reports across the six competency dimensions. Table IV gives the same data for each group, so the averages can be checked against the individual results. Overall, the reports perform strongly and fairly evenly on the analytical dimensions. Analytical reasoning (87.5%), multi-source correlation (86.0%), evidence identification (82.0%), and evidence integrity (82.0%) are close together and well above the pass threshold. The weakest dimension is professional reporting (73.2%). Table IV shows that this weakness is not spread evenly. In four groups (C2, C4, C6, C8), the reporting score is at least 10 percentage points lower than the group's own average on the three ground-truth-checkable dimensions. For these groups, the problem lies in synthesis and communication rather than in analysis. Group C2 is the clearest example. Its ground-truth-checkable scores average 85% and its analytical-reasoning score is 95%, yet its professional-reporting score is only 40%, the largest within-group gap in the cohort. Critical evaluation shows a different pattern. Its mean (80.0%) is high, but its standard deviation (42.2 percentage points) is the largest of all dimensions. This reflects an all-or-nothing, bimodal outcome. Eight of the ten groups correctly identified and dismissed their embedded red herring, while two groups (C3, C6) missed it completely even though their other analytical scores were strong. This suggests that telling a red

herring apart from a real attack is a separate competency that analytical strength does not guarantee.

60% pass threshold. Professional reporting is the lowest-mean dimension, and critical evaluation has the highest variance.

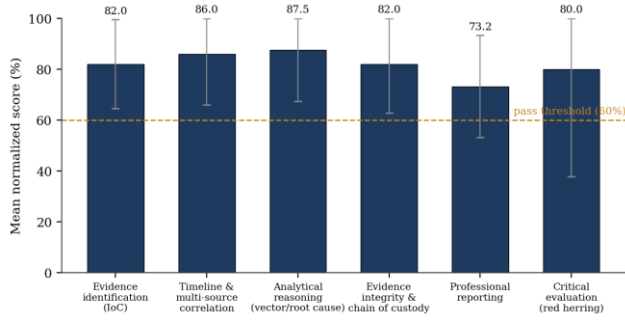


Fig. 1. Mean normalized competency scores across the ten group reports, with ± 1 SD whiskers clipped to the 0–100 range. The dashed line marks the

TABLE IV

Group / case	IoC	Timeline	Analysis	Integrity	Reporting	Crit. eval.	Overall*
C1 Defacement	90	100	100	80	88	100	92.6
C2 SQLi dump	85	80	95	80	40	100	73.7
C3 Spear-phish	70	100	100	100	88	0	90.5
C4 USB exfil	95	100	100	100	88	100	95.8
C5 Cryptominer	55	50	50	50	60	100	53.7
C6 Ransomware	95	100	100	100	72	0	91.6
C7 Supply-chain	90	90	90	80	80	100	86.3
C8 DNS tunneling	50	50	50	50	40	100	47.4
C9 Cred. stuffing	90	90	90	80	80	100	86.3
C10 Joomla	100	100	100	100	96	100	98.9
Mean (SD)	82.0 (17.5)	86.0 (20.1)	87.5 (20.2)	82.0 (19.3)	73.2 (20.1)	80.0 (42.2)	81.7 (17.8)

B. Performance across case variants

Each group worked on a different scenario, and no case variant was completed by more than one group. Case difficulty and group ability are therefore confounded in this design. A low score on a case could reflect a harder case, a weaker group, or both, and this pilot cannot tell them apart. With that confound in mind, Fig. 2 and Table IV show that scores cover a wide range, from 47.4% to 98.9%. The pattern fits signal visibility as one possible explanation among others. The intrusions with clear signatures are at the top: the Joomla backdoor (C10, 98.9%), the insider USB exfiltration (C4, 95.8%), the SSH-driven defacement (C1, 92.6%), the RDP ransomware (C6, 91.6%), and the spear-phishing case (C3, 90.5%). In these cases the attack signature becomes directly visible once the right log is filtered. The two lowest scores belong to the covert, low-signal cases, DNS tunneling (C8, 47.4%) and cryptomining (C5, 53.7%). Both require inferring malicious activity from query entropy, traffic volume, or behavioral anomalies instead of from an explicit signature. We report this as a preliminary, hypothesis-generating pattern that is consistent with a signal-visibility explanation, not as evidence that these scenarios are objectively harder. Separating case difficulty from group ability would require a design with more than one group per case or repeated use across cohorts. Observed results also differed from the a priori difficulty ratings in Table I for two cases. The supply-chain scenario (C7) scored higher than expected, and cryptomining

(C5) scored lower. This shows why a priori difficulty ratings need to be checked empirically instead of being accepted at face value. The two missed red herrings (C3, C6) both came from groups with otherwise high core scores, which supports the view that critical evaluation is a distinct competency (Section V-A) and not simply a side effect of case difficulty.

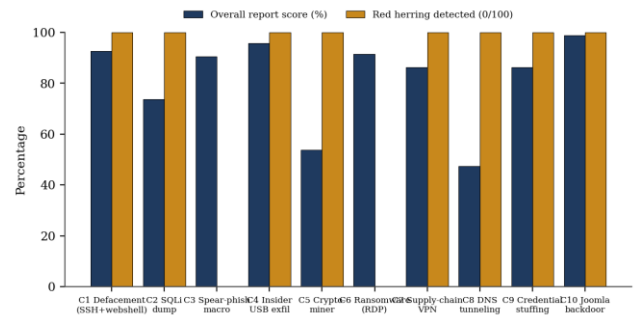


Fig. 2. Overall report score and red-herring outcome by case variant (one group per case, $N = 10$). Because case and group ability are confounded by design, patterns here are preliminary. The two missed red herrings (C3, C6) occurred in otherwise high-scoring groups.

C. Taxonomy of recurrent patterns, with exact frequencies

Table V reports the recurring patterns found in the per-group data, using exact frequencies instead of qualitative labels. A synthesis-and-reporting gap, defined as a reporting score at least 10 percentage points below the group's own average on the three ground-truth-checkable dimensions, appeared in 4 of 10 groups (C2, C4, C6, C8). This is the most

consistent pattern in the data. Difficulty in detecting covert signals was, by design, limited to the 2 of 10 groups that received the low-signal cases (C5, C8). Both of these groups scored 55% or lower on every ground-truth-checkable dimension. Accepting a red herring as real affected 2 of 10 groups (C3, C6). Apart from these three measurable patterns,

the qualitative review during grading also found several reports that pasted raw command output in place of interpreted analysis. Since this was not scored separately or coded systematically, we describe it here in narrative form rather than as a frequency. Systematic coding of such report-writing sub-patterns is left for future work.

TABLE V
RECURRENT PATTERNS AND THEIR COMPETENCY LOCUS, WITH EXACT FREQUENCIES

Pattern	Operational Definition	Competency Locus	Frequency
Synthesis / reporting gap	Reporting score ≥ 10 points below the group's own mean on IoC, timeline, and integrity.	Professional reporting	4 / 10 (C2, C4, C6, C8)
Covert-signal detection difficulty	Case required inference from entropy/volume rather than an explicit signature.	Evidence id. / analysis	2 / 10 (C5, C8)
Red-herring acceptance	Embedded red herring not identified or not dismissed with justification.	Critical evaluation	2 / 10 (C3, C6)

VI. DISCUSSION

A. What the pilot suggests about competence

Read as a pilot and not as a confirmatory study, the breakdown of scores suggests that the reports are not weak at the higher cognitive levels of the forensic process. The dimensions tied to Bloom's *analyze* level, multi-source correlation and analytical reasoning, have the highest scores. This is an encouraging, though preliminary, sign that the scaffolded instruction before the assignment carried over into authentic analysis. The two possible weaknesses are found elsewhere. The first sits at the *create* level. Professional reporting has the lowest mean, and the per-group gap analysis (Section V-A) suggests that for a meaningful subset of groups the bottleneck is synthesis and communication, not the investigation itself. The second concerns the consistency of *evaluation*, not its average. Critical evaluation is high on average but bimodal. The concern, then, is not that the cohort is generally weak at this skill, but that some otherwise capable groups fail to tell benign activity from malicious activity at all. Both observations should be treated as hypotheses for future studies with larger samples, not as established findings from this single cohort.

B. Toward a validity argument, not a validated instrument

We use Kane's argument-based framework [9] to state clearly what this pilot supports and what it does not. Evidence for the scoring inference is relatively strong for the three ground-truth-checkable dimensions, where marks are tied to a verifiable answer key rather than a holistic impression. It is relatively weak for the two judged dimensions, where single-rater subjectivity has not been addressed (Section IV-B). Evidence for the generalization inference, meaning whether these ten cases and this one rater represent the wider range of investigations and judges, has not yet been established. One case per variant, one rater, and one cohort are each insufficient on their own to support generalization, and we do not claim otherwise. Evidence for the extrapolation inference, which would link assignment scores to real investigative performance in the workplace, was not collected in this study. Evidence for the implications inference, concerning the

consequences of using this instrument for grading or placement decisions, is also outside the scope of this paper. For these reasons, we present the empirical content of the paper as a pilot assessment design with a first descriptive application, not as a validated measurement instrument. The argument still needs specific additional evidence before broader claims are justified: a second independent rater with a reported agreement statistic, more than one group per case, and replication across cohorts.

C. Limitations

Several limitations restrict the claims made in this paper. The study covers a single cohort of ten groups, with one group per case variant, so case difficulty and group ability are confounded (Section V-B). The quantitative results should be read as a within-course pilot, not as a generalizable estimate. Scoring was done by a single instructor acting as rater. Ground-truth answer keys limit subjectivity on three of the six dimensions, but the two judged dimensions (analytical reasoning and professional reporting) still carry an unaddressed risk of single-rater bias, and no inter-rater reliability statistic (Cohen's kappa or ICC) is available for this deployment. The evidence packages are synthetic. This makes ground truth possible, but it cannot reproduce the scale and messiness of real production telemetry. Assessing groups can also hide variation between individuals, and the individual reflections only partly address this. Finally, with $N = 10$, all reported figures are descriptive, and no inferential statistical claims are made or implied anywhere in the paper.

D. Disclosed Use of AI assistance

Students were allowed to use large language models as analysis assistants, provided that any use was disclosed in the technical appendix. Undisclosed AI-generated analysis was treated as plagiarism. Because disclosure was self-reported and not independently checked, the present data cannot show how much AI assistance affected the quality of the analysis as opposed to the quality of the report writing. We note this as an additional limitation, alongside those listed in Section VI-C.

VII. CONCLUSION AND FUTURE WORK

We presented a variant-based, log-grounded incident-investigation assignment, designed as a pilot response to a general problem in forensic and cybersecurity education, namely that authentic tasks are hard to score with the reliability needed for validated measurement. Because the instrument is built to resist collusion, uses scoring anchored to ground truth, and includes embedded red herrings, several competencies can be measured with a clear account of the evidence behind each score. Read descriptively, the ten group reports in this pilot performed strongly and consistently on evidence identification, correlation, and analytical reasoning. The two possible weaknesses were a synthesis-and-reporting gap in four groups and bimodal performance on critical evaluation. Viewed through Kane's argument-based validity framework, the pilot offers initial evidence for the scoring inference but does not yet provide the generalization, extrapolation, or implications evidence that a validated instrument would need. In future work, we plan to repeat the deployment across several cohorts, assign more than one group to each case variant so that case difficulty can be separated from group ability, add a second independent rater and report an inter-rater reliability statistic for the judged dimensions, and include a pre/post competency measure to estimate learning gain. Only after this work will we make broader claims about the instrument's validity or about case difficulty.

ACKNOWLEDGMENT

The authors thank the students of the Computer Forensics course at Telkom University, Surabaya, whose investigation reports formed the basis of this analysis.

REFERENCES

- [1] G. Wiggins, "The case for authentic assessment," *Practical Assessment, Research & Evaluation*, vol. 2, no. 2, pp. 1–3, 1990.
- [2] G. Palmer, "A road map for digital forensic research," in *Proc. 1st Digital Forensic Research Workshop (DFRWS)*, Utica, NY, 2001, pp. 27–30.
- [3] B. Carrier and E. H. Spafford, "Getting physical with the digital investigation process," *Int. J. Digital Evidence*, vol. 2, no. 2, pp. 1–20, 2003.
- [4] K. Kent, S. Chevalier, T. Grance, and H. Dang, *Guide to Integrating Forensic Techniques into Incident Response*, NIST Special Publication 800-86. Gaithersburg, MD: NIST, 2006.
- [5] W. Newhouse, S. Keith, B. Scribner, and G. Witte, *National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework*, NIST Special Publication 800-181. Gaithersburg, MD: NIST, 2017.
- [6] J. Biggs, "Enhancing teaching through constructive alignment," *Higher Education*, vol. 32, no. 3, pp. 347–364, 1996.
- [7] L. W. Anderson and D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman, 2001.
- [8] J. R. Savery, "Overview of problem-based learning: Definitions and distinctions," *Interdisciplinary J. Problem-Based Learning*, vol. 1, no. 1, pp. 9–20, 2006.
- [9] M. T. Kane, "Validating the interpretations and uses of test scores," *J. Educational Measurement*, vol. 50, no. 1, pp. 1–73, 2013.
- [10] H. J. Hamnett and D. S. Matthews, "Authentic assessment in forensic science programmes: A UK-based survey of practitioners," *Science & Justice*, vol. 65, pp. 35–42, 2025.
- [11] V. Švábenský, P. Čeleda, J. Vykopal, and S. Brišáková, "Cybersecurity knowledge and skills taught in capture the flag challenges," *Computers & Security*, vol. 102, art. 102154, 2021.

- [12] T. Schafeitel-Tähtinen and W. Lazarov, "Teaching and learning cybersecurity using capture the flag: Effectiveness comparison between university students in Finland and Czechia," *Computer Applications in Engineering Education*, vol. 33, no. 5, art. e70082, 2025.
- [13] E. Ekvall Hansson, P. J. Svensson, E. L. Strandberg, M. Troein, and A. Beckman, "Inter-rater reliability and agreement of rubrics for assessment of scientific writing," *Education*, vol. 4, no. 1, pp. 32–36, 2014.
- [14] Republic of Indonesia, *Undang-Undang No. 19 Tahun 2016 tentang Informasi dan Transaksi Elektronik (UU ITE)*. Jakarta, 2016.
- [15] Republic of Indonesia, *Undang-Undang No. 27 Tahun 2022 tentang Pelindungan Data Pribadi (UU PDP)*. Jakarta, 2022.

Rizqi Ramadhan received the S.T. degree in electrical engineering from Universitas Jember, Jember, East Java, Indonesia, in 2021. He is currently pursuing the Ph.D. degree in information systems at Institut Teknologi Sepuluh Nopember (ITS), Surabaya, East Java, Indonesia, under the supervision of Dr.Eng. Febriliyan Samopa and Izzat Aulia Akbar, Ph.D. He began his career with an internship at Angkasa Pura, followed by positions as a Social Media Specialist at PT Berdunia Digital and a Project Manager at PT Antika Raya, where his responsibilities included content creation for Indonesian e-commerce marketplace platforms.

He is currently a Lecturer in the Information Systems program at Universitas Telkom Kampus Surabaya, Surabaya, Indonesia, teaching Computer Forensics, Information Systems Security, Computer Networks, Introduction to Information Systems, and Operating Systems. He also serves as Director of Insan Kreatif Labs. His publications include "Triase Forensik URL Phishing pada Kasus Penipuan Online di Indonesia: Pendekatan Random Forest Menggunakan Fitur Leksikal" (*Jurnal Cyber Security dan Forensik Digital*, 2026). His research interests include wearable photoplethysmography (PPG) signal processing, TinyML embedded deployment, digital health, and digital forensics.

Adzanil Rachmadi Putra received the S.Kom degree in information systems from Universitas Brawijaya, Malang, East Java, Indonesia, in 2016. He is currently Lecturer with the Department of Information Systems, Telkom University, Surabaya, Indonesia. His research interests include Geoinformatics.

Purnama Anaking received the S.Kom, in Institut Teknologi Sepuluh Nopember. He is currently Lecturer with the Department of Information Systems, Telkom University, Surabaya, Indonesia. His research interests include Software Engineer.

Dudung Adijono received the Police Academy, Indonesia, in 1994. He is currently Leader with the Intermediate Staff and Leadership School (Sespimmen), Sespim Lemdiklat Polri, Lembang, West Bandung Regency, Indonesia. His publications include "Triase Forensik URL Phishing pada Kasus Penipuan Online di Indonesia: Pendekatan Random Forest Menggunakan Fitur Leksikal" (*Jurnal Cyber Security dan Forensik Digital*, 2026). His research interests include Forensics.