Check for updates

컴퓨터교육학회 논문지 2026년 제29권 제9호
https://doi.org/10.32431/kace.2026.29.9.010

디지털포렌식 기술 역량평가를 위한 숙련도 시험 콘텐츠 개발에 관한 연구*

A Study on the Development of Proficiency Test Contents for Evaluating Digital Forensics Technical Competency

신수민† · 김은진†† · 나보미††† · 오정훈†††† · 황현욱††††† · 김기범†††††††
Su-Min Shin† · Eun-Jin Kim†† · Bo-Mi Na††† · Jung-Hoon Oh†††† ·
Hyun-Wook Hwang††††† · Gi-Bum Kim†††††††

요약

본 연구는 국내 환경에 적합한 디지털포렌식 증거 분석 숙련도시험 콘텐츠를 설계하고, 문항의 신뢰도와 타당도 일부 측면을 탐색적으로 검토하는 것을 목적으로 한다. 산업기밀 유출 시나리오를 기반으로 가상환경 증거 이미지와 20문항을 개발하였으며, 실무자 17명의 응답 자료를 활용해 고전검사이론(CTT) 중심의 문항분석과 난이도·변별도 보조지표를 산출하였다. 분석 결과, 전체 검사의 내적일관성 신뢰도(Cronbach's α)는 0.857로 우수한 수준이었다. 다만 일부 문항에서 모호성 증가와 역변별(음의 문항-총점 상관)이 확인되어 문항 서술 및 문항 표현(서술)의 개선 필요성이 제기되었다. 이에 본 연구는 다차원적 평가 기준의 정교화, 부분점수(partial credit) 체계 도입, 증거 이미지 용량 최적화, 문항 표현의 명확성 강화 방안을 제안한다. 본 연구 결과는 디지털포렌식 분야에서 객관적 역량평가 체계를 고도화하기 위한 기초 자료로 활용될 수 있다.

주제어 디지털포렌식, 숙련도시험, 기술 역량 평가, 디지털 증거 분석, 문항 품질 평가

ABSTRACT

This study set out to build a proficiency test for digital forensics that reflects how practitioners in Korea actually work. We designed a virtual evidence image set around an industrial secret leakage scenario and wrote 20 items to go with it. Seventeen working practitioners sat the test, and we analyzed their responses mainly through Classical Test Theory. Internal consistency was high, with Cronbach's α reaching 0.857. A handful of items drew higher ambiguity ratings and showed negative discrimination indices, pointing to problems in how those items were worded and how the item stems were constructed. Items that accept more than one correct answer need better-defined scoring rules. Multistep tasks should be scored in a way that allows partial credit. Evidence-image files need to be small enough for the test to run stably across different environments. Where participants struggled with the wording of an item stem, that wording should be revised. The study offers an initial empirical basis for strengthening competency assessment in digital forensics.

Keywords Digital Forensics, Proficiency Test, Technical Competency Evaluation, Digital Evidence Analysis, Item Quality Assessment

†정회원 한국국방연구원 보안팀
††정회원 성균관대학교 일반대학원 과학수사학과 박사과정
†††정회원 한국저작권보호원 정보기술부
††††정회원 ETRI부설연구소 선임연구원
†††††정회원 ETRI부설연구소 책임연구원
††††††정회원 성균관대학교 일반대학원 과학수사학과 교수(교신저자)

논문투고 2026년 04월 20일
심사완료 2026년 07월 28일
게재확정 2026년 08월 12일
발행일자 2026년 09월 30일

* 본 논문은 2025년 국가보안기술연구소의 위탁으로 성균관대학교에서 수행된 연구보고서(2025-0017)의 내용을 기반으로 수정·보완하여 작성되었음

1. Introduction

As digital devices become more widespread, digital evidence is relevant to a growing number of investigations [1]. Because digital evidence is easily copied, generated, forged, and deleted, it causes difficulties in proving the integrity and reliability of the evidence in court [2]. Therefore, the technical competency of digital forensic examiners who collect and analyze such evidence is being highlighted as a crucial factor [3, 4]. The evaluation of the technical competency of digital forensic examiners is conducted in various forms, such as certification exams, CTF (Capture The Flag) challenges, and proficiency tests. Regarding certifications, the Korean Institute of Forensic Sciences operates the Digital Forensic Expert certification domestically. It consists of written and practical exams, and the practical exam evaluates the participants' analysis competency by providing images and presenting questions [5]. Internationally, certifications such as the EnCase Certified Examiner (EnCE) hosted by OpenText, the AccessData Certified Examiner (ACE) by AccessData, the Magnet Certified Forensics Examiner (MCFE) by Magnet, and the X-Ways Professional in Evidence Recovery Techniques (X-PERT) by X-Ways Software Technology AG also provide practical exams through images [6, 7]. For CTFs and challenges, HM Company, a domestic internal audit firm, independently operates a CTF site and utilizes it to evaluate the analysis capabilities of experts and job applicants [8].

Furthermore, organizations like the Korea Institute of Information Security and Cryptology create images for competitions and conduct challenges to evaluate the capabilities of digital forensic-related experts and practitioners [9, 10]. Overseas, the Digital Forensic Research Workshop (DFRWS) holds an annual challenge and provides images related to each topic, such as data carving, memory dump files, PC images, network packets, memory forensics, and GPU malware analysis [11]. In addition, The Honeynet Project Forensic provides scenarios for analyzing network traffic server attacks and malicious PDF file attacks by providing TCP dump images and log files through digital forensic challenges [12]. Meanwhile, the digital forensics Proficiency Test (PT) is an accredited system that evaluates the technical competency of digital forensic testing institutions [13]. In South Korea, recognized by the Korea Laboratory Accreditation Scheme (KOLAS), institutions like the National Forensic Service, the National Police Agency's Digital Forensic Center, and the Supreme Prosecutors' Office regularly conduct proficiency tests in the digital forensics field. The proficiency test operating institution is an organization that evaluates and certifies the development and operation of schemes (scenarios) for proficiency testing operations in accordance with international standards. Currently, the National Security Research Institute is the sole domestic proficiency test operating institution in the digital forensics field, and most proficiency tests rely on overseas proficiency test operating agencies such as ISFCE, CTS, and Forensic Foundations. Overseas proficiency tests are conducted in a format where participants are provided with storage media or image files containing digital evidence to solve problems within a certain period, and then submit the correct answers. The digital forensics proficiency test can verify participants' capabilities by evaluating digital forensic skills, including the generation of digital images, hash value assignment, and the identification of active or deleted files.

Internationally, too, the assessment of competency in digital forensics is becoming more systematic, in both standards and the scholarly literature. The 2023 revision of ISO/IEC 17043 substantially strengthened the general requirements for proficiency-testing providers, introducing requirements on impartiality, risk-based information management, and alignment with ISO/IEC 17025:2017 [14]. In the United Kingdom, the Forensic Science Regulator's Code of Practice — the first statutory code of practice for forensic science activities in England and Wales, and the first anywhere in the world — came into force on 2 October 2023; its second version, published in June 2025, took effect on 2 October 2025 [15]. In the United States, the Scientific Working Group on Digital Evidence (SWGDE) released the second edition of its Core Competencies for Digital Forensics in December 2024, updating the knowledge, skills, and abilities required of digital forensic practitioners and providing an explicit basis for training, certification, and proficiency-testing programs [16]. Within this institutional

trajectory, academic research has also begun to examine competence at both the laboratory and the practitioner level [17]. In parallel with these formal standards frameworks, industry-led capture-the-flag (CTF) events have been expanding as informal competency benchmarks. For example, registrations for the Magnet Virtual Summit CTF, an internationally open digital-forensics competition, rose from more than 1,350 in 2024 (491 of them solving at least one challenge) to over 1,742 in 2025 (578 solvers) and a record of more than 3,950 in 2026 (1,041 solvers) [18-20]. These figures point to the growing role of informal, practice-oriented benchmarks in the digital-forensics community.

As such, activities for evaluating digital forensic technical competency are gradually increasing. However, the unfortunate aspect is that, excluding certification courses and proficiency tests, objective education and systems that analyze data utilizing digital forensic images are insufficient. In addition, the conditions for providing a practical environment for digital forensics are also not sufficient. Although a basic curriculum for human resource development is prepared, the focus is on theory and tool utilization techniques rather than analysis considering the actual environment. Furthermore, because images produced overseas are utilized, difficulties exist during practical analysis due to scenarios, artifacts, language, and environmental differences that are not suitable for the domestic environment. Images produced domestically are mostly utilized for certifications, digital forensic tool verification, or challenges [21, 22]. In addition, the items provided by domestic and international challenges and proficiency tests utilizing images do not present their rationale, and it is difficult to determine whether the items are practically appropriate for evaluating the technical competency of digital forensic examiners. Therefore, this study aims to develop proficiency test contents for evaluating the technical competency of digital forensic examiners in the domestic context and to provide an initial empirical examination of their assessment item quality. The structure of this paper is as follows: Chapter 2 reviews digital forensic technical competency, proficiency tests, and item evaluation indices (difficulty, discrimination, and validity) through previous studies. Chapter 3 proceeds with the design and execution of proficiency test contents through the creation of digital forensic images based on an industrial espionage scenario. Chapter 4 presents the results of an exploratory CTT-based analysis of item quality, including difficulty, discrimination, and reliability, together with selected validity-related evidence based on participants' responses. Chapter 5 derives practical implications in terms of proficiency test evaluation criteria, partial score calculation, image file capacity optimization, and the enhancement of item clarity by synthesizing the statistical analysis results and participant feedback. Finally, Chapter 6 summarizes the research results and discusses the limitations and future research directions.

2. Theoretical Background

2.1 Digital Forensic Technical Competency

Competency is defined variously depending on scholars and research. Woodruffe (1993) described the concept of competency as a behavioral dimension that enables a person to competently perform a specific action or task [23]. Kim et al. (2017) explained competency as a concept comprising knowledge, skill, and attitude [24]. Yoon et al. (2017) investigated the priorities of competencies required for digital forensic experts utilizing the Analytic Hierarchy Process (AHP) technique [3]. A survey was conducted targeting 16 experts, and "utilization skills for digital forensic equipment/tool programs" was derived as the most important competency among 20 core competencies. In addition, that study categorized competency groups into three areas: knowledge, skills, and attitudes as shown in Table 1.

Table 1. Digital forensics expert competencies along knowledge, skills, and attitudes dimensions

Dimension 1	Dimension 2
Knowledge	Understanding of storage devices, file systems, operating systems, applications, networks, and databases. Knowledge of live forensics, anti-forensics, passwords, and hash values
Skills	Skills in utilizing digital forensic equipment, tools, and software. Programming language proficiency and network equipment operation / Data extraction and imaging from digital devices. Proficiency in OA software and forensic report writing
Attitude	Communication skills, integrity, morality, critical judgment, concentration, and presentation

Oh et al. (2023) proposed 25 competency evaluation factors, organized into nine competency groups, based on expert FGIs and a Delphi survey [25]. From a practical perspective, they selected 9 competency groups, including digital forensic theory, evidence collection and management, and disk forensics, and established corresponding competencies to secure validity. Park (2014) conducted a comparative analysis of internationally recognized cyber forensic expert certifications and examined the domains required by domestic CCPP, Forensics Level 2, EnCE, and CHFI (Computer Hacking Forensic Investigator) [26]. These domains include not only data analysis aspects such as computer architecture, email, and operating systems, but also tool utilization abilities, an introduction to forensics, and legal knowledge. As a result of reviewing previous studies, the competencies related to digital forensic examiners are background knowledge concerning digital evidence, data extraction and analysis, and report writing abilities. Consequently, these can be grouped into a total of 4 categories and 8 competency groups (Table 2), and each competency category and its details can be diversely presented in fields such as disk, embedded, and mobile forensics.

Prior studies have largely stopped at describing the domains and competency frameworks in terms of what practitioners should know and be able to perform. By contrast, this study derives a classification of the observable artifacts and analytical tasks that individual proficiency-test items can sample from. Accordingly, Chapter 3 develops an item classification framework for assessment, and its derivation process and expert-validation procedure are described in detail in Section 3.2.2.

Table 2. Competency by digital forensics procedure

Category	Competency
Basic Knowledge	Identification of technical forensic resources
	Knowledge (Digital forensic theory and legal principles)
Data Extraction	On-site digital evidence collection
	Image generation
	Data acquisition from storage devices or other evidence
Data Analysis	Digital forensic tool utilization skills
	Advanced data analysis (Deleted file recovery, email, and artifact analysis)
Reporting	Digital forensic finding report writing

2.2 Proficiency Test

Proficiency testing evaluates participants according to criteria established in advance, and in Korea it is run by organizations such as KOLAS (the Korea Laboratory Accreditation Scheme) [13, 27]. It is used in many different fields, including mechanics, chemistry, electricity, and forensic science, and each field has its own procedures and test materials that reflect its particular characteristics [28]. In digital forensics, organizations such as ISFCE, CTS, and Forensic Foundations operate proficiency testing programs outside Korea. ISFCE (the International Society of Forensic Computer Examiners) is a U.S. organization specializing in digital forensic proficiency testing, and it administers these tests using the same materials as those used in the CCE certification process [29]. They design and distribute tests that can measure the competencies of digital forensic examiners, including the use of write-blockers, media imaging, hash value assignment, file list generation, and the recovery of active and deleted files. Specifically, it requires a summary of the scenario and an analysis of all media, files, and folders related to the incident. Collaborative Testing Services (CTS) is a proficiency test operating institution for forensics, and regarding digital forensics, it covers mobile (Android, iOS), computer drives, USBs, and multimedia (video, audio) [30]. Participants download image files through links provided on the site or receive physical shipments of removable disks requiring evaluation of the acquisition and extraction process, after which they solve problems and write answers in a short-answer or descriptive format. Simultaneously with image production and distribution, participants' proficiency is evaluated through test items, which are classified into hash, system and partition, registry and system configuration, web and communication, and file and metadata analysis. Forensic Foundations is a proficiency testing institution in Australia, which began providing proficiency tests in the digital forensics field in 2022. It is conducted in the form of inter-laboratory comparison tests. There are a total of two fields: small media and mobile devices, and samples (such as USBs and mobile devices) are shipped directly to the participants. Participants search for relevant evidence and write the analysis results in a report

format. Unlike ISFCE and CTS, it also provides problems in the legal domain; therefore, it is characterized by requiring participants to possess legal background knowledge regarding digital evidence and by providing comprehensive questions and items. Verifying the validity of tools, checking for damage to proficiency test items such as image generation, analyzing the scenario's timeline, identifying the owner and user of storage devices, securing related evidence documents, solving problems through the analysis of shared artifacts, and furthermore, drafting legal opinions such as case analysis (law) represent a novel type not seen in previous proficiency tests. Table 3 summarizes the operating procedures of the three providers.

2.3. Item Evaluation

2.3.1 Difficulty

Item difficulty is an index indicating the degree of easiness or difficulty, representing the proportion of respondents who answered correctly. Generally, it can be interpreted and classified as follows: a correct answer rate of less than 20% indicates a very difficult item, 20% to less than 40% is a difficult item, 40% to less than 60% is an item of medium difficulty, 60% to less than 80% is an easy item, and 80% or more indicates a very easy item [31]. Alternatively, an item difficulty of 0.25 or below is classified as a difficult item, 0.25 to less than 0.75 as an appropriate item, and 0.75 or above as an easy item [32].

$$P = \frac{R}{N}$$

N : *Total Number of Participants*
R : *Number of Correct Respondents by item*

2.3.2 Discrimination

Discrimination is an index indicating the extent to which each item can discriminate the participants' ability levels. It is distributed from -1 to +1, and the closer it is to 1, the higher the discriminatory power. If the item discrimination is 0.3 or higher, it can be evaluated as a good item [33]. Methods for calculating discrimination include the "upper and lower group differentiation method, " which compares the difficulty difference between the high-scoring and low-scoring groups, and the "item-total correlation method, " which calculates the correlation coefficient between an item and the total score [34]. The upper and lower group differentiation method divides respondents into upper and lower groups based on their scores and measures discrimination by analyzing the difference in item responses within each group. The item-total correlation coefficient measures discrimination by

Table 3. Comparison of major international digital-forensics proficiency-test providers

Dimension	ISFCE	CTS	Forensic Foundation
Country / Region	USA	USA	Australia
Domain coverage	Digital evidence handling: write-blocker use, media imaging, hash assignment, file-list generation, recovery of active/deleted files, unallocated-space recovery	Digital and multimedia evidence: computer drives, mobile, external storage; covers acquisition, preservation, analysis (Windows registry, file system, location, DB, carving, physical-media extraction), audio, video	Digital media (USB, thumb drives, SD cards); mobile devices
Test-item format / Delivery	Image file, online distribution	USB and image file, online distribution	Physical sample mailed (USB, SD card)
Item homogeneity / integrity	MD5 hash provided	MD5 hash provided	No formal integrity-verification element
Registration process	Order via ISFCE website; payment link sent by email	Order and pay via CTS website	Order and pay via Forensic Foundations website
Response format	Plain-text submission, including the participant's own analytical report	Per-item answer submitted through CTS portal (short-answer + descriptive)	Report written in the provided template
Accreditation	ANAB-accredited; ISO/IEC 17043:2023	ISO/IEC 17043 accredited; long-standing forensics PT provider	ISO/IEC 17043:2023 accredited; entered the digital-forensics PT field in 2022
Inclusion of legal knowledge	Not separately assessed	Not separately assessed	Explicitly assessed (legal-opinion items)
Position relative to this study	Single-image, item-based; submission format requires non-Korean reporting conventions	Broad device coverage but limited Korean-context relevance; web-portal answer format closest to the platform built in this study	Adds legal-opinion items, but materials in English and physical-shipment dependency creates time-zone and logistics overhead

analyzing the correlation between the score of each item and the total score.

$$r = \frac{x\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

x : *Item Score*, y : *Total Score*, n : *Number of Respondents*

2.3.3 Validity

Validity refers to the extent to which an evaluation tool or test reflects its intended evaluation goals [35]. Types of validity are classified based on evidence based on test content, internal structure, relations to other variables, response processes, and testing consequences. Evidence based on test content (content validity) is defined as the degree to which items, questions, and purposes represent the content intended to be measured. It is verified by the subjective judgment of content experts with professional expertise. Evidence based on internal structure (construct validity) analyzes abstract human psychological characteristics or traits as psychological constructs and, after operational definition, verifies whether they were measured properly [31].

2.3.4 Reliability

For reliability, Cronbach's alpha (α), which is an internal consistency reliability, is utilized. A Cronbach's α coefficient closer to 1 implies higher reliability [31].

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_i^2}{\sigma_T^2}\right)$$

N : *Number of items*
σ_i^2 : *item Variances*
σ_T^2 : *Variance of Total Scores*

3. Design of Proficiency Test Content Based on Technical Competency

3.1 Scope of the Study

Digital forensics and digital evidence are exceptionally broad fields. Therefore, this study limits the design scope of the proficiency test contents to computer devices (e.g., PCs, laptops) and focuses on web browsers, file systems, and so forth. Fundamentally, it assumes a situation where a user's PC and USB have been seized. In addition, the participants of the proficiency test are limited to those with an understanding of digital evidence, such as digital forensic examiners and investigators.

3.2 Research Methodology

3.2.1. Research Procedure

The production of the proficiency test contents was largely divided into five stages. First, a proficiency test item classification framework was established. Subsequently, a digital forensic scenario was created, and corresponding actions were performed within a virtual environment to generate digital evidence. At this stage, proficiency test item was drafted according to the item classification framework established previously. Furthermore, proficiency test item was produced through the imaging of the virtual environment where the digital evidence was stored. The overall five-stage development process—from establishing the classification framework, through scenario creation and evidence generation, to item drafting and imaging—is summarized in Figure 1, which serves as the roadmap for the procedures detailed in the following subsections (3.2.2–3.2.7).

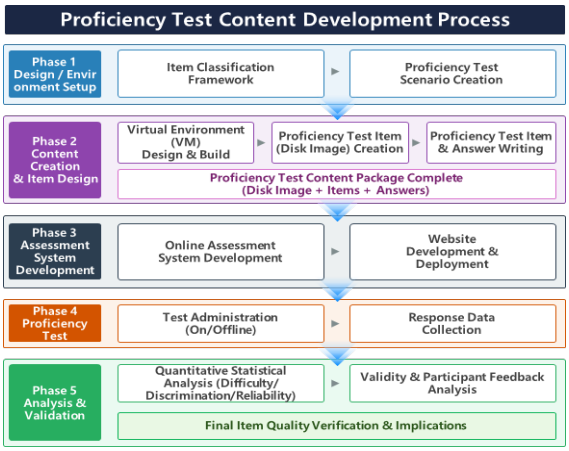

Figure 1. Proficiency Test Content Development Procedure

Figure 1 presents the five stages in their logical dependency order rather than as a strict chronological sequence. In practice the stages

overlapped: refinement of the classification framework (Section 3.2.2), scenario-based evidence generation and imaging (Section 3.2.4), item drafting (Section 3.2.5), and portal development (Section 3.2.6) proceeded concurrently in late September 2025, which is why the reported date ranges overlap. The proficiency test in Section 3.2.7 was administered on the finalized image, item set, and portal.

3.2.2. Establishment of Proficiency Test Item Classification Framework

The classification framework consists of three hierarchical levels. The top level was organized into eight major categories that group artifacts by the activities that can occur in a Windows environment: (1) System, (2) User behavior and processes, (3) File-based, (4) Network-based, (5) Security policy-based, (6) Web browser, (7) Data concealment and manipulation, and (8) Purpose-specific behavior. These eight major categories are further refined into 52 sub-categories and, at the most granular level, into approximately 170 specific artifacts. For example, the 'File-based' major category comprises sub-categories such as File Metadata, File System, Volume & boot information, Deleted Data, Cache & preview data, and External Storage device connection. Within File Metadata, the specific artifacts include File/Folder Metadata, File Creation/Modification/Deletion Logs, and NTFS Object ID. This serves as the basis for deriving the individual assessment items. Table 4 summarizes the major and sub-categories.

The framework was developed and validated in two stages. First, a structure-appropriateness survey was administered by email over one week, from September 26 to October 2, 2025, to 15 digital-forensics experts recruited through purposive sampling. The selection criteria were (i) affiliation with an investigative agency (police, prosecution, or special judicial police) or a digital-forensics/investigation institution, (ii) current practice in the relevant field, and (iii) prior experience analyzing OS-based artifacts. The 15 participants comprised nine from the police (including one from the Korea Coast Guard), one from the prosecution,

Table 4. Item classification framework: 8 major categories and 52 Sub-categories

Major Category	Sub Category	Major Category	Sub Category
(1) System	1. Installation & update management 2. Power state & boot events 3. Memory analysis 4. OS information 5. Restore & residual data 6. Service & account policies 7. Event logs 8. Error & optimization records 9. Virtualization settings	(5) Security policy-based	32. Security program activity 33. Isolation environment 34. Firewall logs 35. Audit scheduling
(2) User behavior & process	10. Windows session 11. Program execution records 12. Program execution detail 13. Program execution frequency & activity 14. Command input 15. Recent-use traces 16. Document/file use traces 17. Explorer use history 18. Windows settings 19. User notes & notifications 20. Processes	(6) Web browser	36. Visit history 37. Cache & session 38. User input traces 39. Account information 40. Local data 41. Downloads 42. Crashes
(3) File-based	21. File metadata 22. File system 23. Volume & boot information 24. Deleted data 25. Cache & preview data 26. External storage device connection	(7) Data concealment & manipulation	43. File-system-level concealment 44. Volume encryption (BitLocker) 45. Multimedia-based concealment 46. Document-based concealment 47. Network-based concealment
(4) Network-based	27. Network device connection 28. Network configuration 29. Network session & server logs 30. RDP logs 31. Other remote-access services	(8) Purpose-specific behavior	48. Crypto-assets 49. Server operation 50. Finance/payment 51. Remote work & collaboration 52. Social/content

one from a military investigative body, and four from special judicial police and administrative-investigation agencies (the Financial Supervisory Service, the Korea Fair Trade Commission, the National Election Commission, and the Gyeonggi Provincial special judicial police), spanning the breadth of investigative and inspection practice. Their experience ranged from 3 to 25 years (mean ≈ 9.5, median 9), with 13 having five or more years. The survey primarily addressed the appropriateness of the major-, sub-, and specific-artifact divisions of Windows OS artifacts, along with queries about missing artifacts.

Appropriateness was determined by aggregating expert opinion rather than by a numerical rating. An item was revised when two or more experts raised the same change (relocation, consolidation, or removal) and was otherwise retained as drafted, and a missing artifact was added only when proposed in common by two or more experts. The items

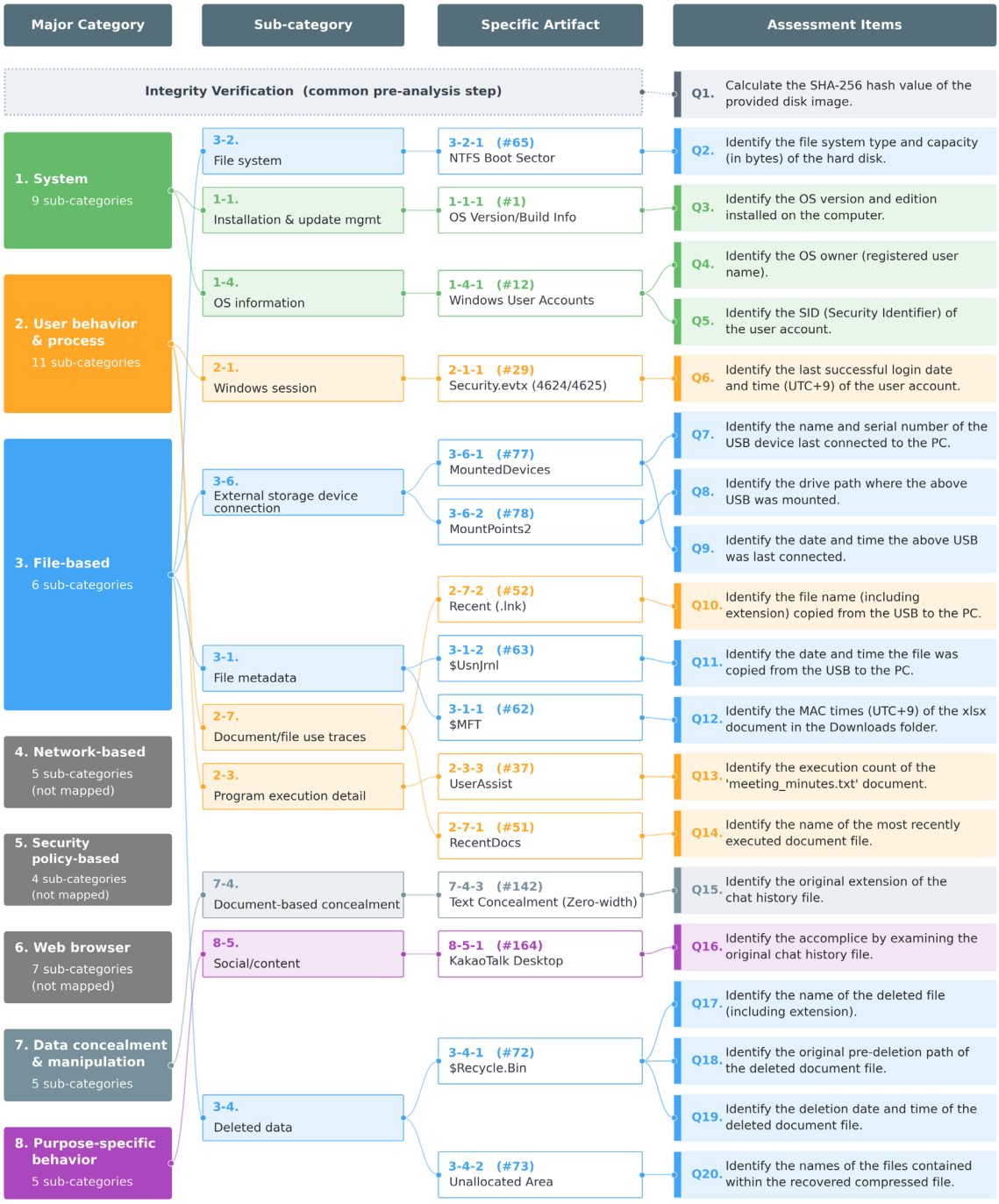

Figure 2. Developing PT Assessment Items

thus derived were then subjected to a second-stage difficulty-and-frequency rating survey using a five-point Likert scale, administered to 12 practitioners—a subset of the first-stage panel comprising seven from the police, one from a military investigative body, and four from special judicial police and administrative-investigation agencies, whose experience ranged from 4 to 25 years (mean ≈ 10.3, median 9). Difficulty ranged from 1 (immediately verifiable in the UI) to 5 (requiring advanced skills such as kernel/cryptographic analysis or custom-parser development), and frequency from 1 (never performed) to 5 (performed in virtually every case, ≥ 80%). For each item, the mean, mode, and sample standard deviation were computed and used to prioritize item development: items with a high frequency mean and a low standard deviation—indicating convergent expert judgment—were prioritized for inclusion, while the difficulty mean was used to keep the difficulty distribution across items balanced.

3.2.3. Scenario Development

In this study, a scenario was developed based on the theme of industrial secret leakage. The scenarios are as follows:

"Company X faced management difficulties after a competitor released a similar product to a new product scheduled for release in the second half of the year. Internally, suspicions grew about a leak of industrial secrets, and an internal audit was conducted. The audit revealed that employee Kim Ma-ba had been using a personal USB storage device and copying certain documents to external media. Kim Ma-ba, who was part of the business operations team, had access to confidential industrial documents, and after suspicions were raised that these documents were shared with competitors in external markets, Kim Ma-ba was identified as a key suspect in the leak. The investigative agency seized Kim Ma-ba's work PC in accordance with digital forensics procedures and copied the entire disk image to begin analysis. The main issue is whether Kim Ma-ba actually leaked confidential documents via USB and whether he tried to cover up the leak."

3.2.4. Creation of Proficiency Test Item

The proficiency-test items were developed for personal computers (PCs). First, based on the classification framework defined earlier, digital evidence was generated experimentally in a virtual machine (VM) environment according to the scenarios. To configure an environment similar to an actual work computer, productivity tools such as word processors were installed, and snapshots were taken at each action to verify whether the intended artifacts were properly generated. When artifacts were generated contrary to the intent, the system was rolled back to the previous snapshot and the action was re-performed; this process was repeated to pre-verify whether each scenario action in the Windows environment produced artifacts at the intended time and form. Based on the scenarios finalized through this pre-verification, an actual PC environment was set up to generate digital evidence and perform imaging. Windows 11 Pro was installed on a physical PC equipped with a 512 GB SSD and 32 GB RAM, and a local account "Kim Ma-ba" was created. The scenario actions were then performed sequentially. After evidence generation was complete, the system was powered off and the Logicube Falcon imaging device was connected via the SATA interface to perform a bit-by-bit copy. The integrity of the acquired image was then verified by computing its SHA-1 hash value. Two hash algorithms are used in this study for different stages. At the acquisition stage, image integrity was verified with SHA-1, and this SHA-1 value was also provided to participants on the portal so they could confirm the integrity of the downloaded image. At the analysis stage, assessment Item 1 instead asks participants to compute the SHA-256 value of the disk image as a scored task. The two hashes therefore serve distinct roles—download-integrity verification versus an analysis competency check—and are not interchangeable. The Digital Forensic Toolkit (DFT) was used to analyze the generated artifacts. The technical specifications of this examination environment are summarized in Table 5.

Table 5. Technical Specifications of the Examination Environment

Category	Specification
Operating system	Windows 11 Pro
User account	Single local account ("Kim Ma-ba")
File system	NTFS
Storage / external media	Internal SSD + personal USB removable device

Category	Specification
File transfer to target	Files staged on a separate authoring PC, delivered via OneDrive cloud
Data types staged	Documents (including DOCX/XLSX/PDF/HWP/HWPX/etc), images, compressed archives, and a KakaoTalk-chat screenshot saved as JPG
Imaging device	Logicube Falcon
Acquisition method	Bit-by-bit physical image via direct SATA connection
Integrity verification	SHA-1
Analysis tool	DFT (digital forensic tool)

※ OneDrive was used only to stage the prepared files onto the target PC during evidence creation; the assessment items do not rely on OneDrive artifacts, and any cloud-synchronization traces are incidental to the leakage scenario.

Whether the final image conformed to the scenario design intent was confirmed through two stages of post-verification. First, the researchers who participated in the scenario design repeatedly checked whether the evidence generated at each action step matched the artifacts required by the assessment items, and when inconsistencies were found, the image or the items were revised iteratively. Second, an internal researcher who had not participated in the scenario design directly attempted the test on the completed image, and the final image was confirmed by incorporating the feedback submitted as a result. This verification serves as an internal content-validity check ensuring that every assessment item can be answered solely from the distributed image.

3.2.5. Development of PT Assessment Items

Figure 2 shows the resulting assessment-item structure, which links each of the 20 items to its corresponding category in the classification framework in Table 4 and to the scenario actions. A total of 20 assessment items were constructed on the basis of the proficiency test item developed for this study. These items were intended to assess the identification and analysis of different forms of digital evidence, such as OS information, application execution records, file metadata, file systems, and traces left by the use of external storage devices. Instead of listing the full range of the classification framework established earlier, we focused on those categories that matched the scenario and developed them into specific assessment items. Table 6 summarizes the principal user actions in the scenario and the artifacts analyzed for each.

Table 6. User Actions and Corresponding Artifacts Analyzed

Scenario	Artifact analyzed	Information recovered
USB connection	Registry-SYSTEM hive Enum\USBSTOR, MountedDevices	Device name, serial number, drive letter, connection time
File copy / deletion	NTFS metadata — $MFT, $UsnJrnl, $LogFile	Creation/deletion timestamps; event type via $UsnJrnl Reason Code
Extension tampering	File header signature analysis	Whether the file extension was forged/altered
Document access	Jump List, RecentDocs	Recently accessed file paths and execution counts
File deletion (cover-up)	Recycle Bin — $Recycle.Bin ($I, $R files)	Original path & deletion time ($I); recoverability ($R)

In developing the items, we relied on two pathways already established in the manuscript to support their validity. First, every item was anchored to a specific sub-category of the artifact classification framework presented in Section 3.2.2, which had itself been validated through expert review (a structure-appropriateness survey with 15 practitioners and a difficulty-and-frequency rating survey with 12 practitioners). Second, the items were verified through the iterative scenario-and-image cycle in Section 3.2.4, which confirmed that each item could be answered from the artifacts produced by the scenario in the released image. In this study, the term proficiency test item refers to the digital forensic test object provided to participants (e.g., an image file), whereas assessment items refer to the 20 scored prompts used to evaluate participants' responses. These were supply-type short-answer items in which participants extracted a specific value from the image and entered it as free text (e.g., a file path, device name, serial number, timestamp, or execution count), rather than selection-type items such as multiple choice. Item 1 (disk-image hash) is an integrity-verification step common to all digital-forensic examinations and is therefore shown separately, not mapped to a specific artifact category. A structured content-validation procedure, such as expert Delphi review or CVR analysis, was not conducted in this study and remains a task for future research.

3.2.6. Development of Proficiency Test Guide/ Distribution Website

After producing the items, an information and distribution portal was developed to conduct

the proficiency test. A user login function was implemented to allow only authorized personnel to access the proficiency test page, and tabs such as scenario, test items, and feedback registration were implemented within the portal. In the 'Scenario Tab', participants can review the scenario for problem-solving and download the proficiency test item (image). The 'Test Items Tab' provides the assessment items and allows the submission of image analysis results. Furthermore, the 'Register your opinion Tab' was implemented so participants could raise objections during the problem-solving process or suggest improvements (Figure 3).

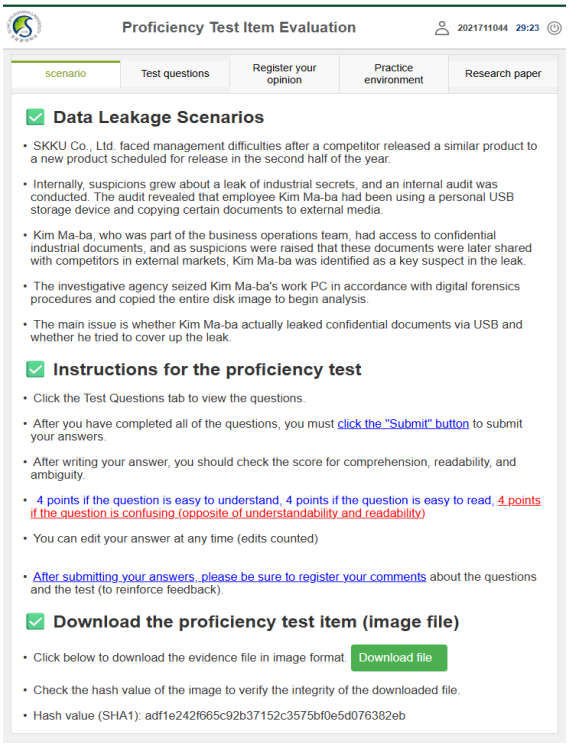

Figure 3. PT Portal Development

3.2.7. Execution of the Proficiency Test

The proficiency test was conducted targeting 17 practitioners in digital forensics-related fields over approximately three weeks, from September 27 to October 17. The 17 participants were recruited from practitioners with digital forensics-related experience. Of these, nine were selected from the private sector (such as companies specializing in digital forensics, internal audit consulting, and information security) and eight from the public sector (such as the military, the police, and intelligence agencies). They were individually contacted by email or phone and took part after confirming their willingness to participate. This recruitment was conducted separately from the expert-validation procedure described in Section 3.2.2, so the proficiency test was administered to a participant pool independent of the expert panel. When each participant performs the proficiency test and submits their results to the portal, the results are stored in the database. This study primarily focused on evaluating the appropriateness of the designed assessment items and identifying issues that emerged from the actual proficiency test results in the domestic context.

4. Research Results

4.1. Proficiency Test Execution Results

In this study, the item difficulty, discrimination, and contribution to reliability (change in α when the item is deleted) were calculated based on the response data of 17 respondents to 20 items (Table 7). Given the small sample size (n = 17), the results reported below should be interpreted as exploratory. This reflects the practical difficulty of recruiting qualified practitioners in the highly specialized field of digital forensics. In addition, the 17 respondents were not a convenience sample but active digital forensic examiners and investigators—the specialized practitioner group the proficiency test is intended to serve—and reliance on such small expert samples is well established in this field (e.g., Yoon et al. (2017) [3] surveyed only 16 experts, and Oh et al. (2023) [25] used expert FGIs and a Delphi panel). Where inferential statistics are presented, they are intended as supplementary information, and the main interpretation is based on the direction and magnitude of correlations as well as item-level patterns. Each item was evaluated focusing on Classical Test Theory (CTT) indices, and for difficulty, the logit-converted difficulty (b*) of the correct answer rate was presented as a supplementary index along with the correct answer rate (p). Furthermore, for discrimination, the CTT item-total correlation coefficient (rpb) and the upper-lower 27% group-based index (UL27) were reported. The analysis was conducted through separate coding utilizing Python 3.10.

Each response was checked against a predefined

answer key by automated string matching, with every item scored dichotomously as correct (1) or incorrect (0). For multi-response items, a response was counted as correct only when all required values were provided, since no partial-credit rule had been established in advance (see Section 5.2). The internal consistency reliability (Cronbach's α) of the entire test was 0.857, indicating a generally excellent level (≥0.8).

4.2. Evaluation of PT Assessment Items

4.2.1. Assessment Item Difficulty

Under CTT the average proportion-correct across items came to 0.600, which puts the test just on the easy side of moderate difficulty. When those proportion-correct values were rescaled to logits, the mean difficulty estimate (b*) was -0.482—again suggesting that participants, on the whole, did not find the items especially demanding. This means that participants did not perceive the assessment items as particularly difficult. It represents an appropriate difficulty distribution for measuring general competency or basic conceptual understanding, which is the purpose of the PT. At the same time, some revision will be necessary if the test is to distinguish more effectively between higher- and lower-performing groups. In particular, the very difficult items (Items 6 and 13), whose near-floor success rates limit their measurement value, are the primary candidates for revision or replacement. The very easy items (≥80% correct: Items 4, 12, 14, 16, 19, and 20) are less of a concern: although they leave little headroom at the top of the ability range, most still discriminate adequately (rpb = 0.48 to 0.83) and were therefore retained, flagged only for possible difficulty rebalancing in future iterations.

Table 7. Summary of Validation Results for PT Assessment Items

No	Success Rate (%)	Difficulty		Discrimination	UL27	UL27_p	α if Item Deleted	Item Clarity Ratings			Remarks (Decision)
		CTT(p)	b* (logit from p)	CTT (rpb)				C	R	A	
1	76.5	0.765	−1.179	0.762	0.6	0.1667	0.837	3.412	3.412	1.000	ACC
2	64.7	0.647	−0.606	0.090	0	1	0.866	3.471	3.588	1.059	ACC / PD
3	23.5	0.235	1.179	0.263	0	1	0.858	3.235	3.294	1.059	MOD
4	94.1	0.941	−2.773	0.479	0.2	1	0.852	3.353	3.353	1.118	VE
5	70.6	0.706	−0.875	0.390	0.2	1	0.853	3.235	3.353	1.118	ACC
6	5.9	0.059	2.773	0.120	0	1	0.860	3.000	3.118	1.176	VD / PD
7	35.3	0.353	0.606	0.424	0.4	0.5238	0.852	3.059	3.176	1.353	ACC / RR
8	58.8	0.588	−0.357	0.404	0.8	0.0476	0.853	3.000	3.176	1.294	ACC
9	41.2	0.412	0.357	−0.003	−0.4	0.5238	0.871	3.176	3.118	1.176	SND / MR
10	52.9	0.529	−0.118	0.612	0.8	0.0476	0.843	2.588	2.824	1.765	ACC / IR
11	41.2	0.412	0.357	0.560	0.8	0.0476	0.846	2.647	2.824	1.706	ACC / IR
12	88.2	0.882	−2.015	0.585	0.4	0.4444	0.847	3.235	3.176	1.294	VE
13	11.8	0.118	2.015	−0.280	−0.4	0.4444	0.872	2.588	2.824	1.529	MOD / SND / MR / RR
14	88.2	0.882	−2.015	0.821	0.4	0.4444	0.839	3.294	3.294	1.118	VE
15	52.9	0.529	−0.118	0.612	0.8	0.0476	0.843	3.294	3.235	1.176	ACC
16	82.4	0.824	−1.540	0.830	0.6	0.1667	0.836	3.353	3.353	1.118	VE
17	64.7	0.647	−0.606	0.392	0.4	0.4444	0.853	3.118	3.176	1.294	ACC
18	76.5	0.765	−1.179	0.651	0.6	0.1667	0.842	3.353	3.353	1.176	ACC
19	82.4	0.824	−1.540	0.706	0.4	0.4444	0.841	3.353	3.353	1.118	VE
20	88.2	0.882	−2.015	0.821	0.4	0.4444	0.839	3.235	3.235	1.176	VE

※ The overall internal consistency reliability (Cronbach's alpha) of the test is .857

※ $b^* = \ln\left(\frac{1-p}{p}\right)$ is the logit transformation of the proportion correct (p), reported as a supplementary descriptive index for comparing item difficulty. It should not be interpreted as an IRT difficulty parameter.

※ UL27: upper−lower 27% discrimination index; rpb: item−total correlation

※ CTT: Classical Test Theory, C: Comprehensibility, R: Readability, A: Ambiguity

※ Given the small sample size (n = 17), IRT parameter estimates were not reported, and all item decisions were based on CTT indices.

※ Decision Codes: ACC = Acceptable; MOD = Moderate; VE = Very Easy; PD = Low/Poor Discrimination; VD = Very Difficult; RR = Revision Required (Ambiguity); IR = Improve Readability/Comprehension; SND = Suspected Negative Discrimination; MR = Major Revision

4.2.2. Assessment Item Discrimination

Under classical test theory, item-total correlations ranged from -0.28 to 0.83, with an average of about 0.45. Taken together, this suggests that the items were doing a reasonably good job of distinguishing higher-performing participants from lower-performing ones. Decisions about whether to keep, revise, or remove items were therefore based mainly on rpb and UL27. Items 2 and 6 yielded very low discrimination and were identified as items with insufficient discriminatory power. In particular, Item 6 had only one correct answerer, and its item-total correlation was close to 0, indicating that the item's function practically did not operate properly. In addition, Item 13 showed a clearly negative item–total correlation (rpb = −0.280), while Item 9 showed a near-zero correlation (rpb = −0.003) together with a negative upper–lower index (UL27 = −0.40); in neither case did the higher-scoring group outperform the lower-scoring group. For Item 9, this signal is attributable to its structural dependence on Item 7 (see Section 4.3) rather than to a defect in the item itself. Because the item conditions or wording may be ambiguous, the wording should be revised and the correct answers re-examined; a detailed cause analysis is provided in Section 4.3.

4.2.3. Exploratory Evidence Based on Response Processes (Comprehensibility, Readability, Ambiguity)

This section examines exploratory evidence based on response processes by analyzing participants' ratings of comprehensibility, readability, and ambiguity for each assessment item. To this end, participants rated each assessment item in terms of comprehensibility, readability, and ambiguity. Most items received scores of 3 or above for comprehensibility and readability, indicating that participants generally understood the assessment items without much difficulty. On the whole, the wording appears to have been reasonably clear and the language easy to follow. At the same time, several items received lower ratings on one or more of these dimensions, suggesting that their wording should be improved before the test is used again. Nevertheless, Items 10, 11, and 13 showed relatively low comprehensibility and readability, along with comparatively high ambiguity, suggesting that participants had difficulty grasping the meaning of these items. In addition, Items 7 and 8 demonstrated a medium level of comprehensibility and readability, but their ambiguity was slightly high. Therefore, it is necessary to improve these items by simplifying their sentence expressions and clearly adjusting the semantic differences in the item stems. Next, the relationships between item comprehensibility, readability, ambiguity, and the correct answer rate were examined. Overall, a positive correlation was observed where higher comprehensibility and readability were associated with a higher correct answer rate. The average correct answer rate of all items was approximately 60.0% (consistent with the mean proportion-correct of 0.600 reported in Section 4.2.1), the average comprehensibility was 3.150 points, the average readability was 3.212 points, and the average ambiguity was 1.241 points. Item-level correlations among proportion correct, comprehensibility, readability, and ambiguity were examined. The correlation analysis results illustrated that item comprehensibility ($r = .608^{**}$, $p < .01$) and readability ($r = .546^{*}$, $p < .05$) exhibited a significant positive correlation with the correct answer rate. Conversely, ambiguity ($r = -.364$) had a negative correlation with the correct answer rate and was not statistically significant ($p = .112$), as shown in Table 8.

Table 8. Correlation Matrix: Proportion Correct and Item Clarity Ratings (C/R/A)

Variable	Proportion Correct	Comprehensibility	Readability	Ambiguity
Proportion Correct	1.000			
Comprehensibility	0.608**	1.000		
Readability	0.546*	0.956***	1.000	
Ambiguity	−0.364	−0.922***	−0.904***	1.000

※ * p < .05, ** p < .01, *** p < .001

※ The correlation between the proportion correct and ambiguity was not statistically significant (p = .112)

※ Since the unit of analysis is items (n = 20), statistical power is limited.

※ Significance levels are provided for reference, and interpretations focus on the direction and magnitude of the correlation coefficients.

※ Comprehensibility, readability, and ambiguity are near-collinear (|r| = 0.90–0.96) and are interpreted jointly as one "item clarity" dimension.

This suggests a descriptive pattern in which more ambiguous items tended to be accompanied by lower comprehensibility and readability, rather than directly affecting the correct answer rate.

In fact, ambiguity demonstrated a very strong negative correlation with comprehensibility ($r = -.922$, $p < .001$) and readability ($r = -.904$, $p < .001$), indicating a tendency where more ambiguous items simultaneously led to lower comprehensibility and readability. Accordingly, the items should be revised to improve clarity, reduce potential misunderstanding, and enhance comprehensibility and readability.

4.3 Analysis Results of Proficiency Test Feedback

Participant feedback was collected through the proficiency-test web portal in two structured forms. First, on the per-item answer page, participants assigned three integer ratings to each item—comprehensibility, readability, and ambiguity—on a 0-to-4 scale. Second, after completing all twenty items, participants could submit free-text comments on the test platform, image, or item wording through a dedicated "Opinion Submission" tab. Free-text submission was voluntary, and the volume and scope of submitted opinions varied across respondents. From these submissions, the research team selected the comments directly relevant to item quality or test administration and grouped them by target into two categories: comments on items and comments on the test image (Table 9). The item comments were further distinguished by whether they addressed item wording or artifact interpretation. The grouped comments were then cross-checked against the quantitative findings reported in Section 4.2, particularly the items with negative discrimination (Items 9 and 13) and the items with high ambiguity scores (Items 10, 11, and 13). This qualitative procedure was conducted not as a formal content analysis based on a pre-defined codebook but as a complementary analysis intended to support and contextualize the quantitative results. To protect participant identity, the respondents whose comments are quoted in this section were each assigned a single-letter pseudonym (A through F), with the mapping between pseudonyms and the underlying participant identifiers held only by the research team; no institutional, demographic, or other identifying information appears in the quoted material. Overall, participants saw the test as clearly structured and easy enough to follow, but a few items still caused confusion. Those items were also the ones that showed lower readability and comprehensibility and higher ambiguity in the earlier analysis. Participants A and B both referred to uncertainty about the technical assumptions built into some questions, especially ones involving LNK files, Jump Lists, and Prefetch records. That may mean the technical explanation was too condensed, leaving the purpose of the question unclear. Participant C raised much the same issue in Item 15. Even though the test was evaluated positively overall, this participant felt that the phrase “file extension” left room for more than one interpretation. The wording was technically correct, but not precise enough to prevent different readings. In other words, the item may have been

Table 9. Participant Feedback on Proficiency Test (Summary)

Div	Item No	Participant Feedback
A	10, 13	· Regarding the file copy items, since LNK and Jump Lists are artifacts generated upon file execution, it would be more appropriate to rephrase the questions to focus on "files executed from external storage media." · For the TXT file execution count, Prefetch tracks the execution frequency of 'notepad.exe' rather than a specific file, which leads to some ambiguity.
B	10, 11, 13	· Items 10 and 11 asked for time information regarding files copied from a USB to a PC; however, LNK and Jump Lists only provide the execution dates from the USB. I solved the tasks based on this interpretation of the question's intent. · Regarding Item 13 (TXT execution count), Prefetch shows the application execution count. Therefore, I interpreted and solved it as the execution count for 'notepad.exe'.
C	15	· The test items are well-structured, allowing for a comprehensive analysis of various Windows OS artifacts using forensic tools. · However, there was some confusion regarding the question about the "original extension"—specifically, whether it referred to the extension in the filename or the actual file type (signature).
D	7, 13	· I was confused between the USB records in setupapi.dev and USBSTOR. · It was difficult to distinguish between the 'SanDisk Cruzer Blade' and 'SanDisk 3.2 Gen 1' models, so I relied on setupapi as the reference. · Additionally, I could not find the .txt file execution history in UserAssist, so I was unsure where to look for that information.
E	Image File	· It would be preferable if the PT image files were slightly smaller in size
F	Image File	· The file size was so large that I had to procure additional hardware, which caused a delay in completing the test

readable at the surface level, but the focus of the question was not sufficiently clear, allowing multiple interpretations. For that reason, both word choice and the scope of the item description need to be made more explicit. Comments from Participants E and F also suggest that the time required to download the image files and the limitations of analytical tools or hardware when handling large image files may affect performance during the test itself. Item-by-item analysis of the proficiency-test data showed that the point-biserial correlations between item score and total score ranged from −0.28 to 0.83, with an overall mean of approximately 0.45. Among these, two items (Item 9 and Item 13) showed negative correlations (−0.003 and −0.280, respectively), confirming them as malfunctioning items in which the higher-scoring group had a lower correct response rate than the lower-scoring group. Such negative discrimination can generally arise from an error in the answer key, confusion in stem interpretation, or a mismatch in item difficulty. We therefore analyzed the cause by jointly reviewing each item's stem, the artifacts from which the correct answer is derived, and the participants' feedback. Item 13 asked participants to report the execution count of the file [CONFIDENTIAL] meeting_minutes.txt; the correct answer is derived from the UserAssist key in NTUSER.DAT, and the correct response rate was 11.76% (2 of 17 participants). This item is a representative case in which participants were confused between the execution count of the document and the execution count of the program: comprehension itself was normal, but the high ambiguity produced a low correct response rate. Among the participants who submitted feedback, A and B answered with the execution count of notepad.exe based on Prefetch, while D reported being unable to find a .txt-file execution entry in UserAssist. In other words, the higher-scoring group, in the course of broadly examining artifacts such as Prefetch and UserAssist, interpreted the stem's intended target as 'the execution of notepad.exe' and thus arrived at an incorrect answer. The negative discrimination is judged to stem from terminological ambiguity in the item stem. Item 9 asked participants to report the timestamp at which 'the USB above' was last connected, where 'the USB above' is an anaphoric reference to the USB device that participants were asked to identify in Item 7. The correct answer is derived from SetupAPI.dev.log and the System event log, and the correct response rate was 41.18% (7 of 17 participants). Participant D reported, for Item 7, that 'I was confused whether it was SanDisk Cruzer Blade or SanDisk 3.2 Gen 1, so I went with SetupAPI,' and for Item 9 that 'USBSTOR and SetupAPI.dev records were confused.' Because Item 9 is structurally dependent on the identification made in Item 7, a respondent who identified the USB device differently at Item 7 necessarily mis-specifies the referent at Item 9. Accordingly, even within the higher-scoring group, an identification error at an antecedent item propagates to the dependent item, which we interpret as the cause of the observed negative discrimination.

5. Implications

5.1. Evaluation Criteria for Proficiency Tests

Proficiency testing is a key component in evaluating participants' practical competence in digital forensics, and it is mainly centered on the ability to acquire and analyze evidence. However, the field still lacks clear and unified criteria for measuring proficiency, and for that reason the way these tests are administered differs across institutions. Most proficiency tests are conducted either in the form of problem-solving questions or in the form of analytical report submission. In problem-solving tests, when a participant's answer does not match the answer presented by the testing institution, the result is usually handled by providing statistical scores based on the correct-answer rate or by requiring the participant to take the test again. The limitation of this format is that it simplifies the complex reasoning and interpretive process involved in digital forensics into a binary evaluation of correct or incorrect. By contrast, the report-submission format makes it possible to evaluate the participant's analytical thinking and written explanation, but in the scoring process the standard for determining correctness can become unclear and there is room for the scorer's subjectivity to intervene. In addition, recent proficiency testing has expanded from PC-based testing to mobile, audio, video, and other kinds of digital device analysis, which makes it harder to

standardize analysis methods and submission forms. Therefore, for the validity and reliability of digital forensic proficiency tests, a commonly applicable evaluation framework is needed. It should not be limited to simple answer agreement, but should also include analytical reasoning, the validity of interpretation, and the reproducibility of evidence.

5.2. Partial Scoring Calculation and Answer Submission

The proficiency test used in this study included several items with more than one correct answer. That creates a real scoring problem when a participant provides only part of the expected response. If, for example, a participant identifies only one of three correct answers, the scoring rubric needs to make clear whether that response receives no credit or partial credit. Much the same issue arises in items such as "Write the path of the file." One participant may provide the full path, while another may give only part of the file name. Unless the scoring rule is defined in advance, it is difficult to score such answers consistently. This is more than a routine scoring issue. It means that results may vary not only with forensic competence, but also with differences in reading comprehension and in how participants interpret the item.

Time zones can also affect the answers participants submit. In a test such as the CTS, the same timestamp may be read differently depending on whether the participant works in UTC or in local time. A participant in Korea, for instance, may interpret the evidence in KST (UTC+9). When that happens, the issue is not simply a technical mistake. In many cases, the item itself has left too much room for interpretation by failing to specify the reference time zone. As a result, answers may differ for reasons that have little to do with actual competence, which can undermine the reliability of the test. This can be avoided if time-related items either state clearly that UTC must be used or require a response format such as "YYYY-MM-DD hh:mm:ss UTC±X." A related issue is time precision. Two participants may be referring to the same event but report the time differently, with one giving the value to the minute and the other to the second. The scoring rubric should make clear in advance how those differences will be treated, including whether partial credit will be awarded when a response is correct at one level of precision but not at another. More broadly, proficiency testing would also benefit from standardized response formats and an automated scoring system. That would improve objectivity, reduce the need for repeated item revision, and strengthen the overall reliability of the test by minimizing avoidable errors in interpretation.

5.3. Proficiency Test Item (Image File)

During this research, the majority of participants identified image file capacity as a major inconvenience. The total size of the proficiency test item (image file) was approximately 11GB, causing errors during download and decompression; furthermore, some participants had to secure additional equipment due to insufficient storage space. In particular, after decompression, the total capacity reached approximately 476GB, placing an excessive burden on analysis time and system resources. While full-system imaging is effective for scenarios requiring malware analysis or dynamic behavior tracking, for proficiency tests aimed at behavior-based analysis or data leak detection, a method of selectively imaging only necessary data areas should be prioritized. Therefore, when designing proficiency test items in the future, it is necessary to prepare measures to optimize image file capacity or provide partitioned formats, comprehensively considering test purposes, analysis efficiency, and system resource management.

5.4. PT Assessment Items

The item analysis showed that some assessment items varied quite substantially in terms of comprehensibility, readability, and ambiguity. This seems to have resulted from problems with clarity of wording, weak contextual linkage, and inconsistent use of terms. For example, if an item first refers to "the USB device most recently connected to the PC under examination" and then repeatedly uses the phrase "the above USB," participants may not necessarily understand that reference in the same way, especially depending on the order of the items or the page layout. A similar issue arises with prompts such as "Write the name and serial number." Because responses can be given in several different formats, it becomes

difficult to apply scoring criteria consistently. Item 13 is a good example of this problem. The prompt asked participants to “write the number of times the txt document was executed,” but many were uncertain whether this meant the number of times the document itself had been opened or the number of times the relevant program had been run. This appears to have been caused by imprecise wording and inconsistent use of terms, which is also consistent with the ambiguity results reported earlier. For this reason, item writers need to anticipate, during the drafting stage, the possibility that participants may interpret an item in more than one way, and where necessary provide brief clarification or examples within the item itself. It is also important to establish an item database through which proficiency test assessment items can be reviewed and revised on a regular basis, so that item quality can be improved over time.

6. Conclusion

This study was conducted to establish a proficiency testing model related to 'digital evidence analysis' in the domestic digital forensic field. To achieve this, the entire cycle of a proficiency test-from the item design stage to actual production, distribution, and collection of participant feedback-was empirically implemented. The significance of this study lies in the fact that it specifically derived core elements and practical discussion issues that must be considered when developing and operating digital forensics proficiency test contents in the future. However, this study has several limitations that should be addressed in future research.

First, because digital forensics is a highly specialized field, the pool of qualified practitioners available to participate was limited. As a result, it was difficult to secure a sample size large enough to support robust statistical generalization. Although this specialized practitioner sample lends the findings practical relevance, a larger-scale validation study—drawing on a broader and more representative pool of practitioners across multiple institutions—is needed before the results can be generalized. Second, when developing the classification framework for the proficiency test item, the criteria established during the process were not subsequently validated through additional objective procedures, such as an expert Delphi survey. Third, although the assessment items were improved based on participant feedback, quantitative verification of the changes has not yet been conducted (e.g., in terms of comprehensibility, discrimination and response-process-related evidence). Future studies will require empirical validation using improved assessment items across a broader sample population. Currently, domestic digital forensics proficiency testing is primarily focused on the "evidence acquisition" stage, and assessments of "evidence analysis" competencies have yet to be systematically established. This study is intended to help fill this gap and serve as foundational data for the institutional introduction of an "evidence analysis" category in future domestic digital forensics proficiency testing, as well as to promote qualitative advancement in the related field.

References

[1] Lillis, D., Becker, B., O'Sullivan, T., & Scanlon, M. (2016). Current challenges and future research areas for digital forensic investigation. *arXiv preprint* arXiv:1604.03850. https://doi.org/10.48550/arXiv.1604.03850

[2] Yoo, Y., Song, B., & Park, S. (2009). The necessity and training plans for digital forensic expertise. *Korean Association of Police Science Review, 11*(4), 253-284. UCI: G704-002121.2009.11.4.007

[3] Yun, H., Lee, S., & Lee, J. (2017). Deriving priorities of competences required for digital forensic experts using AHP. *Journal of Society for e-Business Studies, 22*(1), 107-122. https://doi.org/10.7838/jsebs.2017.22.1.107

[4] Na, H., Kim, C., & Lee, N. (2014). A study on designing an undergraduate curriculum in digital forensics per stages for developing human resource. *The Journal of Korean Association of Computer Education, 17*(3), 75-84. https://doi.org/10.32431/kace.2014.17.3.007

[5] Korean Society of Digital Forensics. (n.d.). *Digital forensic expert level 2.* Korean Society of Digital Forensics. https://exam.forensickorea.org/web/exam/items.do

[6] OpenText. (n.d.). *EnCase Certified Examiner (EnCE) certification program.* OpenText. https://www.opentext.com/kr/learning-services/learning-paths-encase-certifications

[7] Exterro. (n.d.). *AccessData certified examiner*. Exterro. https://training.exterro.com/courses/exterro-ace

[8] HM CTF. (n.d.). *HM CTF challenges.* HM CTF. http://www.hmctf.co.kr/challenges

[9] National Intelligence Service. (2025). *Digital forensics challenge 2025*. https://www.dfchallenge.org/
[10] Digital Forensics Research Group. (2025). *Digital forensics challenge 2025*. KIISC. https://kiisc.or.kr/bbs/nn/article/4342
[11] DFRWS. (n.d.). *Forensic challenges*. DFRWS. https://dfrws.org/forensic-challenges/
[12] The Honeynet Project. (n.d.). *Forensic challenges*. The Honeynet Project. https://www.honeynet.org/challenges/
[13] Korea Laboratory Accreditation Scheme. (n.d.). *Proficiency testing*. KOLAS. https://www.knab.go.kr/usr/gud/eri/ErtgTestSumryInfo.do
[14] International Organization for Standardization. (2023). *Conformity assessment — General requirements for proficiency testing* (ISO/IEC 17043:2023). ISO. https://www.iso.org/standard/80864.html
[15] Forensic Science Regulator. (2025). *Forensic science activities: Statutory code of practice — version 2*. UK Government. https://www.gov.uk/government/publications/forensic-science-activities-statutory-code-of-practice-version-2
[16] Scientific Working Group on Digital Evidence. (2024). *Core competencies for digital forensics* (SWGDE 12-F-006-2.0). SWGDE. https://www.swgde.org/documents/published-complete-listing/12-f-006-core-competencies-for-digital-forensics/
[17] Horsman, G., & Dodd, A. (2024). Competence in digital forensics. *Forensic Science International: Digital Investigation, 51*, 301840. https://doi.org/10.1016/j.fsidi.2024.301840
[18] Hyde, J. (2024). *2024 Magnet Virtual Summit CTF winners*. Magnet Forensics. https://www.magnetforensics.com/blog/2024-magnet-virtual-summit-ctf-winners-and-another-chance-to-play/
[19] Hyde, J. (2025). *Announcing the winners of the 2025 Magnet Virtual Summit CTF*. Magnet Forensics. https://www.magnetforensics.com/blog/announcing-the-winners-of-the-2025-magnet-virtual-summit-ctf/
[20] Hyde, J. (2026). *Announcing the winners of the 2026 Magnet Virtual Summit CTF*. Magnet Forensics. https://www.magnetforensics.com/blog/announcing-the-winners-of-the-2026-magnet-virtual-summit-ctf/
[21] Oh, S., Lee, T., & Kim, G. (2021). A study on digital forensic image production model. *Journal of Digital Forensics, 15*(2), 137-150. https://doi.org/10.22798/kdfs.2021.15.2.137
[22] Lee, T., & Shin, S. (2011). Reliability verification of evidence analysis tools for digital forensics. *Journal of the Korea Institute of Information Security and Cryptology, 21*(3), 165-176. https://doi.org/10.13089/JKIISC.2011.21.3.165
[23] Woodruffe, C. (1993). What is meant by a competency? *Leadership & Organization Development Journal, 14*(1), 29-36. https://doi.org/10.1108/eb053651
[24] Kim, D., & Kim, S. (2017). Understanding and issues of core competencies and competency-based education in university education. *Journal of Competency-Based Education, 2*(1), 23-45. https://doi.org/10.52616/JCCER.2017.2.1.23
[25] Oh, S., Jeong, J., & Choi, E. (2023). Development of competency model for police' digital forensic examiner. *Journal of the Korea Institute of Information Security & Cryptology, 33*(4), 647-659. https://doi.org/10.13089/JKIISC.2023.33.4.647
[26] Park, J. (2014). *A study on cyber forensic certification system* [Master's thesis], Dongguk University.
[27] Korea Association of Standards & Testing Organizations. (n.d.). *Proficiency testing dedicated management agency*. KASTO. https://www.kasto.or.kr/kolas/Overview/main.do
[28] Park, K. (2009). *A study on the confidence improvement of testing environment by proficiency test* [Master's thesis], Konkuk University.
[29] ISFCE. (n.d.). *Digital evidence handling proficiency testing*. ISFCE. https://isfce.com/digital-evidence-handling-proficiency-testing.html
[30] CTS. (n.d.). *Digital & multimedia evidence*. CTS. https://cts-forensics.com/program-9.php
[31] Seong, T. (2018). *Foundations of educational evaluation* (2nd ed.). Hakjisa.
[32] Cangelosi, J. (1990). *Designing tests for evaluating student achievement*. Longman.
[33] Ebel, R. (1965). *Measuring educational achievement*. Prentice-Hall.
[34] Johnson, A. (1951). Notes on a suggested index of item validity: The U-L index. Journal of Educational Psychology, 42(8), 499-504. https://doi.org/10.1037/h0060855
[35] Seong, T., Kang, D., Kang, L., Kwak, D., Kim, G., Kim, C., & Hong, H. (2018). *Introduction to modern education. Hakjisa.*

신수민

· 2007년 순천대학교 컴퓨터교육과(이학사)
· 2016년 동국대학교 정보보호전공(정보보호학석사)
· 2025년 성균관대학교 과학수사학과(공학박사)
· 2017년~현재 한국국방연구원 보안팀

관심분야 : 컴퓨터교육, 디지털포렌식, 사이버보안, 숙련도시험, 인공지능 포렌식
shin1121@g.skku.edu

김은진

· 2022년 숭실대학교 글로벌미래교육원 컴퓨터공학과(학사)
· 2024년 성균관대학교 과학수사학과(공학석사)
· 2025년~현재 성균관대학교 과학수사학과(박사과정)

관심분야 : 디지털포렌식, 사이버범죄 수사, 저작권 침해 대응
eunjin810@g.skku.edu

나보미

· 2022년 중앙대학교 심리학과(학사)
· 2025년 성균관대학교 과학수사학과(공학석사)
· 2025년 대검찰청 연구원
· 2026년~현재 한국저작권보호원 정보기술부 주임

관심분야 : 디지털포렌식, 사이버범죄 수사, 범죄심리
forbm507@g.skku.edu

오정훈

· 2024년 고려대학교 정보보호대학원(공학박사)
· 2012년 안랩 A-FIRST 주임연구원
· 2015년~현재 ETRI부설연구소 선임연구원

관심분야 : 디지털포렌식, 파일시스템, 침해사고 대응, 인공지능
blueangel@nsr.re.kr

황현욱

· 2004년 전남대학교 정보보호(이학박사)
· 2015년 Arizona State University 교환연구원
· 2004년~현재 ETRI부설연구소 책임연구원

관심분야 : 디지털포렌식, 취약점 검증, 악성코드, 인공지능
hhu@nsr.re.kr

김기범

· 1997년 경찰대학 행정학과(학사)
· 2009년 고려대학교 정보보호대학원(공학석사)
· 2017년 고려대학교 정보보호대학원(공학박사)
· 2020년~현재 성균관대학교 과학수사학과 교수

관심분야 : 디지털포렌식, 사이버범죄 수사, 정보보호, 국제협력
freekgb02@gmail.com