・國家菁英・ No.65
第 19 卷第 1 期
【專題】
2026 年 6 月 頁 1-38

Implications of Computerized Adaptive Testing in the Medical Field for National Professional and Technical Examinations

Jyun-Hong Chen*

Abstract

National professional and technical examinations are the gatekeepers of public safety and the foundation of society's trust. Historically, such examinations have relied on linear testing formats—whether paper-based tests (PBTs) or computer-based tests (CBT)—consisting of fixed-form, multiple-choice questions. Although such tests are efficient, they primarily evaluate the recall of knowledge, the "knows" and "knows how" tiers of Miller's (1990) pyramid of clinical competence. Accordingly, it may lead to construct underrepresentation, failing to capture the dynamic problem-solving skills and reasoning demanded of modern professionals. The study examines how licensure examinations can mitigate this threat and modernize through computer-native testing.

Specifically, this study examined three computer-native benchmarks of medical licensing that assess the professional competence represented in the "shows how" tier of Miller's (1990) pyramid. First, the National Council Licensure Examination for Registered Nurses in the United States was examined as an exemplar of test efficiency because it employs computerized adaptive testing (CAT) to substantially reduce test length without compromising test precision. Second, the United States Medical Licensing Examination Step 3 was examined as an exemplar of environmental fidelity because it employs computer-based case simulations to evaluate candidates' dynamic responses and time management in evolving patient models. Third, the American Board of Radiology Core Exam was examined as an exemplar of instrumental authenticity because it embeds real-world imaging software directly into testing interfaces, mitigating the construct-irrelevant variance associated with static testing modes.

The psychometric viability of these testing approaches depends on item response

theory (IRT) and evidence-centered design (ECD). IRT addresses the limitations of classical test theory (CTT) (e.g., test-dependent scoring) and provides the mathematical foundation for CAT. By leveraging item information functions, CAT dynamically maximizes test efficiency and achieves equal precision across the ability spectrum to ensure fairness. Conversely, ECD structures high-fidelity simulations through a rigorous framework that links observable candidate interactions in simulated tasks to latent proficiencies, mitigating threats to test validity.

The present study proposes a phased roadmap for transitioning high-volume national licensure examinations to CAT and integrates sequential dynamic modules to assess specialty-specific competencies. This roadmap calls for examination authorities to adopt control procedures for test constraints (e.g., item exposure control) to maintain validity under year-round testing. Furthermore, the roadmap advocates for integrating domain-specific tools into testing interfaces to ensure instrumental authenticity, aligning assessment environments with professional practice to accurately measure the complex problem-solving skills required for licensure.

To ensure long-term test sustainability and diagnostic depth, this study recommends two advanced approaches. First, automated item generation can be used to mitigate item bank attrition caused by frequent administration and potential leaks. Second, cognitive diagnostic assessments can enhance psychometric testing by providing fine-grained formative profiles of skill mastery for candidates. These intelligent assessment technologies are vital to securing the integrity of professional credentials and ensuring public safety.

Keywords: national licensure examinations, professional and technical examinations, modern test theory, computerized adaptive testing, evidence-centered design

* Associate Professor, Department of Psychology, National Cheng Kung University

電腦適性測驗對醫事領域國家專技考試之意涵

陳俊宏*

摘要

國家專門職業及技術人員考試是維護公共安全與社會信賴的守門員。傳統上，此類考試採線性測驗（linear testing）模式，不論是紙筆測驗（paper-based testing, PBT）或電腦化測驗（computer-based testing, CBT），皆以固定題本搭配選擇題為主。這類測驗雖然效率高，但主要針對知識記憶，即米勒（Mille, 1990）臨床能力金字塔中的「知其然」（knows）與「知其所以然」（knows how）層次。因此，這樣的測驗形式可能導致測驗構念代表性不足（construct underrepresentation），無法充分評量現代專業人員所需具備之動態問題解決能力與推理能力。本研究探討如何透過電腦原生測驗（computer-native testing），降低專技考試面臨之構念代表性不足問題，並推動考試現代化。

具體而言，本研究分析三項醫事執照考試中，能評估米勒（1990）臨床能力金字塔中「展現」（shows how）層次專業能力之電腦原生測驗。首先，以美國註冊護理師執照考試（National Council Licensure Examination for Registered Nurses, NCLEX-RN）作為高測驗效率之代表案例，討論電腦適性測驗（computerized adaptive testing, CAT）如何在不損失測驗精確度情況下，大幅縮短測驗長度。再以美國醫師執照考試第三階段（United States Medical Licensing Examination Step 3）作為高環境擬真度（environmental fidelity）之代表案例，討論利用電腦化臨床模擬，評量應考人在動態病患情境下的應變與時間管理能力。最後，以美國放射學委員會核心考試（American Board of Radiology Core Exam）為高工具真實性（instrumental authenticity）代表案例，探討其將實務用影像分析軟體，直接整合測驗介面，以降低靜態測驗形式造成的構念無關變異（construct-irrelevant variance）。

上述測驗模式之心理計量可行性（psychometric viability）主要建立在試題反應理論（item response theory, IRT）與證據中心設計（evidence-centered design,

ECD）兩項理論基礎。IRT 克服了古典測驗理論（classical test theory, CTT）的限制，如考生計分的測驗依賴問題，並提供 CAT 所需數學算則基礎。利用試題訊息量函數（item information function），CAT 能動態選取最具測量效能之試題，並在整體能力區間內達到相近的測量精確度，確保測驗公平性。另一方面，ECD 透過嚴謹架構，建立高擬真模擬，將模擬任務中可觀察之應考者行為與潛在能力連結，確保測驗效度。

本研究提出分階段推動計畫，將 CAT 逐步引入大規模國家證照考試，並搭配連續性動態模組（sequential dynamic modules）評量各專業領域核心能力。同時，建議考試主管機關針對實務所需測驗限制建置管控程序，如試題曝光控管（item exposure control），以在全年辦理考試的情境下，維持測驗效度。此外，並應將專業領域實務工具整合至測驗介面，提升工具真實性，使測驗環境更貼近實際工作情境，以準確評量執照取得所需之複雜問題解決能力。

為確保測驗永續發展與診斷深度，本研究進一步提出兩個發展方向。首先，透過自動化試題生成（automated item generation），減緩因頻繁施測及試題外洩風險造成之題庫耗損問題。其次，透過認知診斷評量（cognitive diagnostic assessment），更細緻診斷應考人各項技能的精熟情形，以強化心理測驗的功能。上述智慧化評量技術，對確保專業證照公信力及公共安全具關鍵意義。

關鍵詞：國家專門職業及技術人員考試、專技人員考試、現代測驗理論、電腦適性測驗、證據中心設計

* 國立成功大學心理學系副教授

I. Introduction

Professional and technical examinations serve as gatekeepers of public safety and societal trust. Historically, the purpose of these high-stakes assessments was to determine whether a candidate possessed the domain knowledge required to practice safely. However, as modern professions ranging from health care and law to engineering and architecture have become increasingly complex, the definition of readiness to practice has extended beyond recalling subject knowledge to include applying such knowledge under pressure, solving problems under conditions of uncertainty, and executing complex procedures with professionalism and efficiency (Gunderman et al., 2001). Accordingly, to remain relevant, national licensure examinations must evaluate both procedural competence and subject knowledge (Epstein & Hundert, 2002).

Miller's pyramid of clinical competence provides a robust framework for evaluating clinician competence (Miller, 1990). Originally developed for medical education but also applicable to assessments of specialized competencies, the pyramid divides cognitive and behavioral mastery across four hierarchical tiers: "knows" (factual knowledge), "knows how" (applied knowledge and clinical reasoning), "shows how" (performance in a simulated setting), and "does" (action in real-world practice). For decades, national licensure examinations have relied on linear testing formats, including both paper-based tests (PBT) and fixed-form computer-based tests (CBT), with multiple-choice questions (MCQs) serving as the predominant format. MCQs are efficient, objective, and broad enough to ensure high internal consistency and reliability (Crocker & Algina, 1986). However, they have limited ecological validity and can only measure the lowest two tiers of Miller's pyramid (knows and knows how) (Haladyna & Downing, 2004). Furthermore, MCQs only require candidates to select an answer from a given list of options, a receptive cognitive process based on recognition rather than generation (Schuwirth et al., 1996; Schuwirth & van der Vleuten, 2003).

In real-world practice, practitioners are rarely, if ever, presented with a well-structured scenario with only four or five possible responses, one of which is guaranteed to be correct. Instead, they must independently gather information, generate

diagnostic hypotheses, determine solutions, and adapt strategies based on environmental feedback. When licensure examiners rely heavily on MCQs to test a practitioner, they make the unsupported assumption that the ability to recognize a correct answer on a piece of paper translates to the ability to correctly execute a procedure in practice (Kane, 2013). From a psychometric perspective, reliance on MCQs leads to construct underrepresentation—the failure of an assessment to capture essential dimensions of the traits it intends to measure (Haladyna & Downing, 2004; Messick, 1989).

The fundamental challenge, however, lies not only in the item formats but in the inherent physical constraints of both PBT and fixed-form CBT. Traditional paper formats—including both MCQs and conventional constructed-response questions—are inadequate for eliciting generative behaviors, capturing multistep problem-solving, or providing high-fidelity simulations of the consequences of candidate decisions. To address this need for realism and precision, there is a necessary transition toward computer-based assessment (Parshall et al., 2002; Zenisky & Sireci, 2002). This study further conceptualizes this shift as "computer-native assessment." Aligning with Bennett's (2005) framework, this study operationalizes computer-native assessment not merely as the digitalization of test content, but as the transformation of assessments from static administrative events into dynamic, algorithmic interactions between the system and the examinee. However, the practical implementation of such complex systems at a national scale requires a robust, evidence-based transition strategy. The present study, therefore, proposes a strategic roadmap for modernizing our country's national licensure examinations by drawing on advanced empirical models adopted in the United States medical professions, which have long served as the forefront of high-stakes assessment innovation because of their rigorous demands for both test precision and clinical authenticity (Swanson et al., 1995).

In 1994, the National Council of State Boards of Nursing (NCSBN) transitioned the National Council Licensure Examination for Registered Nurses (NCLEX-RN) to a computerized adaptive testing (CAT) format, demonstrating that item response theory (IRT) algorithms could be implemented reliably at a national scale (Zara, 1999). Similarly, in 1999, the United States Medical Licensing Examination (USMLE) Step 3

introduced computer-based case simulations (CCS) and demonstrated that dynamic, interactive testing could be scored objectively and used to effectively determine whether candidates were qualified for licensure (Clauser et al., 2002; Dillon et al., 2004). Subsequently, organizations such as the American Board of Radiology (ABR) have substantially enhanced the instrumental authenticity of examinations by integrating authentic, domain-specific tools (e.g., volumetric image manipulation software) directly into licensure assessments (Becker & Dunnick, 2008; Hollingsworth et al., 2010).

Although the content of these medical examinations is highly specialized, the underlying computer-native architectures are content-agnostic. This study argues that by examining the successes and challenges of the medical field's transition to computer-native assessment, national examination authorities can avoid decades of trial-and-error, adopting instead mature, evidence-based frameworks to modernize licensure programs. Accordingly, this study proposes a structured multi-phase roadmap for adopting such frameworks.

The remainder of this study is organized as follows. Chapter II, Benchmarks in Medical Licensure, explores three medical examinations—the NCLEX-RN, USMLE Step 3, and ABR Core Exam—to illustrate advances in the computational efficiency, environmental fidelity, and instrumental authenticity of computerized licensure examinations. Chapter III, Implication I: Modern Test Theory and Adaptive Testing, examines the mathematical foundations of IRT and CAT, exploring how adaptive testing increases efficiency. Chapter IV, Implication II: Validity Arguments in Evidence-Centered Design (ECD), investigates the conceptual framework of ECD, focusing on how this model enhances the validity of computer-native assessments. Chapter V, Strategic Recommendations for National Licensure Examinations, translates these implications into practical strategies for national examination authorities, proposing an integrated CAT framework for high-volume general examinations and tool-integrated simulations for specialized professions. Finally, Chapter VI, Future Directions and Conclusions, explores the future of intelligent assessment by examining how automated item generation (AIG) and cognitive diagnostic assessments (CDA) can be used to modernize professional licensure examinations.

II. Benchmarks in Medical Licensure

To establish a strategy for modernizing national licensure examination systems, policymakers must investigate successful examples of applying modern test theory at scale. This chapter examines three such examples of computer-native medical licensure examinations: the NCLEX-RN, USMLE Step 3, and ABR Core Exam. Each example is selected to illustrate a key dimension of modern computer-native assessment: the algorithmic efficiency of adaptive testing, the environmental fidelity of dynamic clinical simulations, and the instrumental authenticity of testing interfaces.

A. NCLEX-RN: Adaptive Testing

Administered by the NCSBN, the NCLEX-RN represents a benchmark in assessment efficiency. In 1994, the NCSBN transitioned its national nursing licensure examination from a traditional linear testing format to a CAT format (Wendt & Harmes, 2009; Zara, 1999). The rationale for this transition was the inherent inefficiency of traditional fixed-form, linear testing. Conventional linear testing involves the administration of an identical set of items to all candidates regardless of their latent ability. To maintain discrimination across the entire ability spectrum, items in a linear test must necessarily vary in difficulty (Lord, 2012). Consequently, for many candidates, the majority of items are misaligned with their latent trait levels: High-ability candidates expend time and cognitive effort answering elementary items, whereas low-ability candidates facing highly difficult items may engage in random or disengaged responding, introducing measurement error into their scores and increasing test anxiety.

To resolve this misalignment between candidate ability and item difficulty, the NCLEX-RN implemented a variable-length CAT design (85-150 items) that terminates when the precision of the ability estimate reaches a predefined threshold. Typically, a variable-length CAT format begins by administering an item of moderate difficulty. Based on the candidate's response, the adaptive algorithm dynamically adjusts the difficulty of subsequent items to correspond with a continually updated estimate of the candidate's ability. After a minimum number of items have been administered, the CAT algorithm assesses the precision of his or her ability estimates by using the accumulated

test information or the confidence interval of the latent ability to determine if the predefined stopping criterion has been met. On licensure examinations requiring binary classifications, CAT algorithms can determine the outcome by assessing whether the confidence interval intersects the passing standard (cutoff score); in other words, the test terminates with a "pass" or "fail" decision if the entire confidence interval falls above or below, respectively, the cutoff score. By transitioning to CAT, the NCSBN significantly reduced test length, thereby substantially increasing efficiency (Zara, 1999).

The NCSBN expanded upon the success of the NCLEX-RN with the Next Generation NCLEX (NGN) in 2023 (National Council of State Boards of Nursing, 2023a). Because clinical judgment involves complex cognitive processing, the NGN introduced novel, interactive item types, such as unfolding case studies (longitudinal scenarios that evolve sequentially), bow-tie questions (visual representations of clinical relationships), and matrix grids (multidimensional decision-making tasks), grounded in the clinical judgment measurement model (Dickison et al., 2019) to measure a candidate's clinical reasoning and dynamic decision-making skills in realistic, evolving scenarios that require the application of knowledge rather than mere recall (National Council of State Boards of Nursing, 2023b). Despite these content and format innovations, the NGN retains the underlying adaptive algorithm used by its predecessor. The success of the NCLEX-RN reveals the value of adaptive algorithms that tailor assessments to a candidate's latent trait level in real time in high-stakes computerized testing. The NCLEX-RN further demonstrates the capacity of CAT to enhance test efficiency without compromising precision—a primary objective of national examination authorities managing high-volume, knowledge-intensive licensure programs in disciplines such as nursing, pharmacy, or social work.

B. USMLE Step 3: Environmental Fidelity

To assess candidates' abilities to solve complex, sequential problems, national examination authorities can adopt measures similar to those used in the USMLE Step 3. The final examination in the USMLE sequence, Step 3, assesses whether a physician possesses the medical knowledge and clinical competence essential for unsupervised practice. To fulfill this objective, the National Board of Medical Examiners introduced

CCS to USMLE Step 3 in 1999 (Dillon et al., 2004). CCS creates an interactive virtual environment that requires candidates to assume the role of an attending physician and manage a simulated patient's care across clinical settings (e.g., urgent care, intensive care units, and outpatient clinics) over durations ranging from hours to months (Dillon et al., 2004).

The environmental fidelity of CCS depends on three key innovations. The first innovation is a requirement to generate scenarios without prompting the candidate with answers. In contrast to linear testing with traditional item types such as MCQs, CCS interfaces do not provide a list of answers. Instead, candidates must use a free-text order sheet connected to an extensive database of diagnostic tests, medications, procedures, and consultation requests. This interface mitigates the "cueing effect" of MCQs by demanding high-level information synthesis and performance rather than recall and recognition, ensuring that candidates who pass the test know what to do in a medical emergency without being prompted (Clauser et al., 1997). The second innovation is a virtual clock that incorporates time as a dynamic, endogenous clinical variable that candidates must proactively manage. For instance, once a candidate orders a diagnostic test or administers a medication in a simulation, they must account for the simulated interval required for the results of the test or to observe the effects of their intervention. This simulation of time in medical emergencies compels candidates to demonstrate that they know not only what to do but also when to do it, mirroring the temporal pressures of real-world clinical practice (Clauser et al., 1997; Dillon et al., 2004).

The third innovation is dynamic physiological modeling. The virtual patients in CCS do not follow predetermined branching narratives; rather, they are generated by complex mathematical models that simulate human physiology and pathophysiology (Swanson et al., 1995). The patient's state vector (e.g., blood pressure, heart rate, and oxygen saturation) dynamically changes based on the interaction between the disease trajectory and the sequence and timing of the candidate's interventions. This dynamism ensures that the assessment environment authentically examines a candidate's unique problem-solving abilities (Clauser et al., 1997).

The successful implementation of CCS in the USMLE Step 3 demonstrates that

complex, open-ended, and highly interactive clinical tasks can be reliably simulated and scored in high-stakes licensure examinations. By capturing sequence and timing in a comprehensive transaction log, automated scoring algorithms map complex behavioral trajectories to expert-defined management rubrics (Dillon et al., 2004). In summary, the USMLE Step 3 provides compelling evidence that environmental fidelity is achievable in licensure examinations.

C. ABR Core Exam: Instrumental Authenticity

Modern clinical practice relies heavily on the expert manipulation of specialized digital software and information contained in complex databases. The ABR Core Exam provides a benchmark for measuring proficiency involving such techniques. Introduced in 2008, the ABR Core Exam is a comprehensive, fully computerized medical licensure examination that emphasizes interaction with actual imaging software rather than static images (American Board of Radiology, n.d.; Becker & Dunnick, 2008; Hollingsworth et al., 2010).

In routine clinical practice, radiologists typically engage with volumetric imaging data sets (e.g., computed tomography and magnetic resonance imaging) by dynamically scrolling through axial, coronal, and sagittal planes to isolate and characterize findings in three-dimensional space (Becker & Dunnick, 2008; Gunderman et al., 2001). This procedural competence extends beyond recognizing individual key slices; it involves the strategic navigation of image stacks and the optimization of display parameters to reveal subtle pathologies (Becker & Dunnick, 2008; Hollingsworth et al., 2010). Assessments that present only static views may be subject to construct underrepresentation and construct-irrelevant variance because they fail to engage clinicians with the procedural aspects of image interpretation essential to professional practice (Haladyna & Downing, 2004). The ABR Core Exam addresses these limitations by presenting candidates with complete, multislice volumetric image sets embedded within an interactive viewer. Candidates are required to scroll through these data sets and manipulate display settings (e.g., making “window and level” adjustments), which correspond to the real-world tasks of optimizing contrast and brightness to differentiate tissue densities (e.g., distinguishing subtle pulmonary parenchymal opacities from adjacent mediastinal soft tissue).

By situating examinees within an interactive, tool-mediated environment, the ABR Core Exam mitigates the risks associated with basing licensure decisions solely on the recognition of static, artificial stimuli. Instead, the ABR Core Exam captures authentic procedural engagement by ensuring instrumental authenticity. By embedding authentic digital tools into the assessment interface, the ABR Core Exam ensures that cognitive processes engaged during the test mirror those required in clinical practice, thereby enhancing the construct validity of the examination. The ABR Core Exam's tool-first approach provides a robust model for modernizing similar highly specialized licensure assessments across certification domains. For instance, a licensure examination for certified public accountants could incorporate authentic spreadsheet and financial modeling software directly into the testing platform. Adopting such instrumentally authentic examinations enables national licensure authorities to ensure that the credentials they confer accurately and dynamically reflect the technological demands of contemporary professional practice.

III. Implication I: Modern Test Theory and Adaptive Testing

A. From CTT to IRT

For much of the 20th century, CTT served as the psychometric foundation for nearly all large-scale educational and professional testing (Crocker & Algina, 1986). However, several limitations of CTT restrict its utility in advanced applications. To elucidate how computer-native assessments achieve comparable measurement precision using fewer items than traditional linear testing, this study examines the major limitations of CTT.

First, CTT fails to achieve parameter invariance (Hambleton et al., 1991). In examinations based on a CTT framework, item parameters are sample-dependent. Item difficulty is defined simply as the proportion of examinees in a specific sample who answered the item correctly, and a candidate's ability is strictly dependent on the specific subset of items administered. This circular dependency limits the capacity to directly compare the scores of candidates who take different sets of items (Yen & Fitzpatrick, 2006).

Second, CTT assumes a single reliability coefficient and a constant standard error

of measurement (SEM) for all examinees (Embretson & Reise, 2000). Psychometrically, this assumption is limited because tests typically provide more precise estimates for candidates of moderate ability than for those at the extremes of the ability spectrum. Lacking a conditional (individualized) index of measurement precision, CTT cannot dynamically determine when a candidate's ability has been estimated with sufficient certainty—the fundamental stopping rule required for the operational efficiency of CAT. Third, CTT applies strictly to tests as a whole rather than to individual items, thereby lacking the capacity to predict or explain a candidate's performance at the item level (Embretson & Reise, 2000). This limitation prevents CTT from using real-time item responses to algorithmically determine the optimal sequence of subsequent items. Consequently, such tests are inherently static and lack the adaptive flexibility required for modern precision testing.

To overcome the structural limitations of CTT, IRT directly models the probabilistic relationship between a candidate's response and the functional combination of his or her proficiency and item parameters (e.g., difficulty) (Lord, 2012). Within the IRT framework, item parameters are theoretically invariant—independent of the calibration sample—provided that model assumptions are met. Although large samples are necessary for stable parameter estimation, this invariance property is a fundamental feature of the IRT model. Furthermore, IRT calibrates item parameters on a common scale of proficiency and item difficulty. Through equating onto this common scale, a candidate's proficiency can be directly compared with others, even if they were administered different subsets of items.

To address the limitation of a single reliability coefficient, IRT utilizes the item information function (IIF) to calculate the amount of information an item provides at a specific ability level. The sum of the IIFs for all administered items yields the test information function (TIF) that indicates the precision of the test (mathematically, the inverse of the error variance of the candidate's ability estimate) (Embretson & Reise, 2000). Consequently, even if two candidates are administered the same set of items, differences in their response patterns result in varying ability estimates and, thereby, distinct degrees of measurement precision. Furthermore, the IIF serves as the core mechanism for adaptive item selection, enabling CAT to dynamically and efficiently select the most informative subsequent item based on the candidate's current ability

estimate.

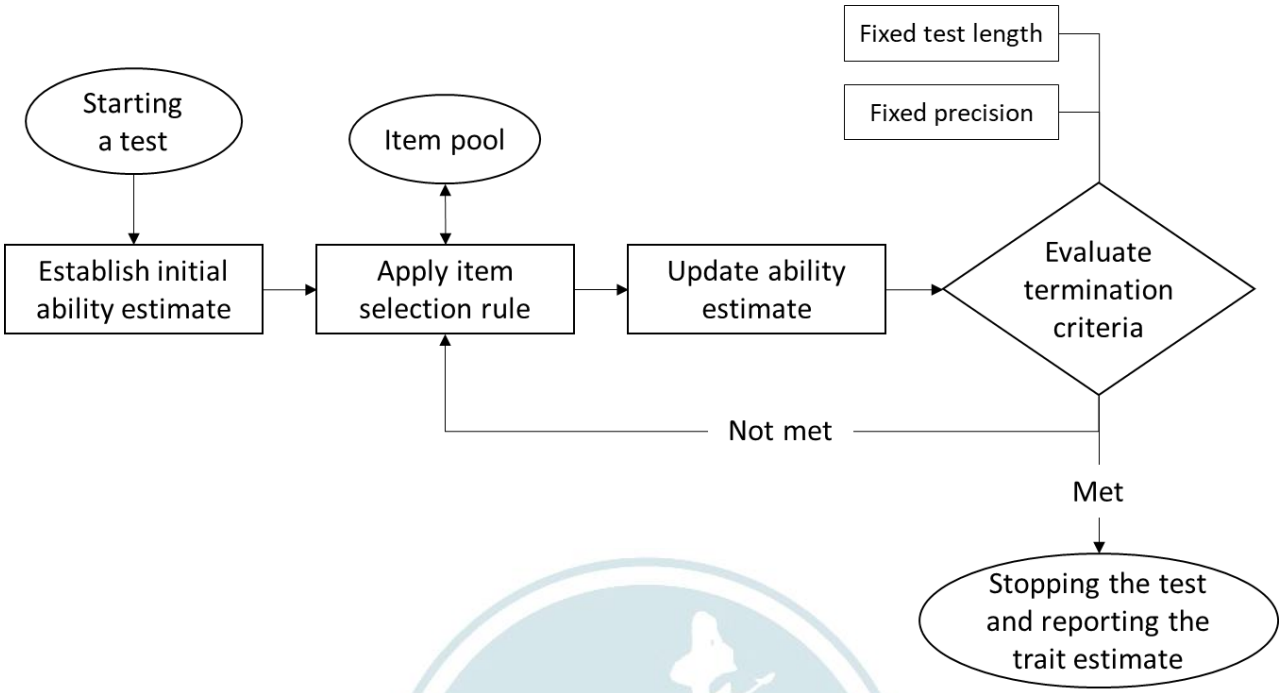
By addressing these structural limitations, IRT has facilitated significant advancements across the field of measurement, including differential item functioning assessment, test equating procedures, the development of advanced measurement models, and the modeling of aberrant testing behavior (e.g., Hori et al., 2022; Huang, 2023; Meijer & Sijtsma, 2001; Thissen et al., 1993). Among these notable applications, CAT—the assessment modality operationalized by the NCLEX-RN—emerges as one of the most impactful, substantially enhancing test efficiency (Chen & Chao, 2025). The following section delineates the fundamental principles detailing how CAT leverages IRT to execute this dynamic assessment process.

B. Applications of CAT

As established in the introduction, computer-native assessments are characterized by the transformation of static administrative events into dynamic algorithmic interactions between the system and the examinee. CAT represents one of the most efficient frameworks within this domain, as it leverages psychometric information derived from these dynamic interactions to continually update a candidate's ability estimate, thereby enhancing test efficiency. Under optimal operational conditions—such as a high-quality item bank, accurate parameter calibration, and a well-aligned distribution of item difficulties and examinee abilities—CAT can achieve similar measurement precision with fewer items than traditional linear testing (van der Linden, 2005; Wainer, 2000). Specifically, the CAT procedure comprises four primary steps: (1) establishing an initial ability estimate, (2) applying an item selection rule, (3) updating a candidate's ability, and (4) evaluating the termination criteria (Figure 1).

Figure 1

CAT Procedure



1. Establishing Initial Ability Estimate

At the start of an examination, CAT lacks information regarding a candidate's initial ability level. Typically, CAT assigns a neutral starting point, such as 0 (e.g., representing the population mean) or the mean ability derived from historical cohort data. To improve estimation, van der Linden (1999) proposed incorporating collateral information, such as educational background or demographic variables, to establish a tailored initial ability estimate for each examinee.

2. Applying an Item Selection Rule

CAT typically selects the item that maximizes Fisher information at the candidate's current ability estimate—a criterion known as Maximum Fisher Information (MFI) (Lord, 2012). The advantage of MFI lies in the direct relationship between Fisher information and the standard error of the ability estimate. Specifically, under maximum likelihood estimation (MLE), the standard error is defined as the inverse of the square root of the test information. By always selecting items with the highest Fisher information for a candidate, CAT minimizes standard error and maximizes test efficiency (Weiss, 1982).

3. Updating the Ability Estimate

After a candidate responds to an item, CAT typically updates the ability estimate using either MLE or Bayesian methods, such as expected a posteriori (EAP) estimation (Baker & Kim, 2004). While MLE is asymptotically unbiased, it produces undefined estimates for all-correct or all-incorrect response patterns. To address this limitation, CAT often employs EAP estimation, which incorporates a prior distribution to ensure that the estimation process remains defined and executable, thereby guaranteeing convergence even with extreme response patterns (Bock & Mislevy, 1982).

4. Evaluating Termination Criteria

Although CAT can adopt a fixed-length design similar to traditional linear testing, its methodological superiority is most evident in the fixed-precision (i.e., variable-length) stopping rule. Under this rule, CAT will continue to administer items until a candidate's ability estimate reaches a predefined level of precision. This adaptive procedure flattens the test information function across the ability continuum, yielding uniform precision for all examinees (Lord, 2012; Weiss, 1982).

5. Additional Consideration 1: Classification Decisions

When an assessment requires a dichotomous pass/fail decision rather than a precise score, CAT algorithms must implement classification rules. Two widely adopted rules are the confidence interval rule and the sequential probability ratio test (SPRT) (Spray & Reckase, 1996; Wald, 1947). Under the confidence interval rule, the algorithm constructs a confidence interval (e.g., 95%) around the candidate's ability estimate after each item administration. If this interval lies above the predetermined cutoff score, the examination terminates with a "pass" decision; if it falls below the cutoff score, the examination terminates with a "fail" decision. If the interval straddles the cutoff score, item administration continues until a definitive classification is achieved or the maximum test length is reached (Lin & Spray, 2000). By contrast, the SPRT evaluates two competing statistical hypotheses after each item administration: whether the candidate's demonstrated ability aligns with a "master" or "nonmaster" state. The algorithm calculates the log-likelihood ratio of these two states based on the

accumulated response vector. If this ratio exceeds the upper boundary, the test terminates with a "pass" decision; if it falls below the lower boundary, it terminates with a "fail" decision. Although the SPRT has been characterized as more efficient than the confidence interval rule (Eggen, 1999; Spray & Reckase, 1996), empirical evidence indicates that this advantage is largely contingent on the testing context, including item selection rules, cutoff scores, and the distribution of candidate abilities. Test authorities should therefore select the decision rule that best aligns with their specific classification requirements and operational constraints (Lin & Spray, 2000).

6. Additional Consideration 2: Test Security and Content Constraints

In addition to test efficiency, CAT must address practical test requirements, such as test security and content validity (Chao & Chen, 2023). However, traditional item selection rules such as MFI operate as "greedy algorithms" (Han, 2018). By continually selecting the most informative items based on a candidate's ability estimate, these algorithms inevitably lead to a small subset of items becoming overused (Chang, 2015). In high-stakes national licensure examinations, item overuse threatens test security by facilitating item harvesting and illicit knowledge sharing, which degrades construct validity (Way, 1998). The Sympson-Hetter method (Sympson & Hetter, 1985) mitigates this threat by applying conditional item exposure control to ensure no item exceeds a predefined maximum exposure rate. Psychometricians have also developed alternative methods, such as stratified item selection and test overlap control procedures, to maintain test security (e.g., Chang & Ying, 1999; Chen et al., 2014).

In addition to security concerns, a major challenge in transitioning national licensure examinations to computer-native formats is ensuring content validity. CAT algorithms must simultaneously manage various nonstatistical constraints (e.g., balancing word counts and avoiding mutually exclusive "enemy" items). To satisfy these requirements, researchers have proposed heuristic methods such as the maximum priority index (Cheng & Chang, 2009). Alternatively, the shadow-test approach (van der Linden, 2000) provides an optimal solution via mixed-integer linear programming. This method selects items to maximize an objective function (e.g., test information) while subjecting to a massive matrix of linear inequalities that represent blueprint specifications and test constraints.

Integrating the exposure control and constraint management algorithms is vital to the sustainable administration of CAT-based examinations. While adaptive item selection enhances estimation precision, heuristic and optimization methods provide practical safeguards to ensure assessments remain secure, valid, and feasible. Combining these methodologies allows national licensure authorities to achieve the psychometric efficiency of adaptive testing while upholding rigorous quality standards for professional certifications.

IV. Implication II: Validity Arguments within ECD

CAT enhances the precision of test administration; however, test efficiency is meaningless without validity. To address the limitations of rote knowledge evaluation and assess multifaceted professional competence (e.g., an architect's spatial reasoning or a physician's sequential diagnostic logic), assessments must adopt interactive, computer-native simulations. Examination authorities can strengthen validity arguments by leveraging ECD to systematically develop these complex tasks, as demonstrated by the case simulations in USMLE Step 3 and the tool-integrated platforms in the ABR Core Exam. By adopting ECD, examination authorities move beyond traditional test development to reimagine assessment as a rigorous exercise in evidentiary reasoning rather than a rote process of answering predetermined questions (Mislevy et al., 2003).

In traditional linear testing, subject matter experts often depend on their intuition to generate items, assuming a correct answer indicates competence. ECD reverses this process by requiring test developers to define the claims they wish to make about a candidate and to determine the behaviors that constitute sufficient evidence of those claims before designing tasks capable of eliciting those behaviors (Mislevy et al., 2003). This reasoning is operationalized in the conceptual assessment framework of ECD, which comprises three interdependent models: the student, evidence, and task models (Mislevy et al., 2003).

The student model is the foundation of the conceptual assessment framework; it defines the combination of knowledge, skills, and abilities an examination intends to measure, specifies the constructs prioritized by the examination board, and establishes the scope of construct validity (Pellegrino et al., 2001). The task model outlines the

interactive environmental structure of an assessment and determines the conditions, stimuli, and digital tools presented to the candidate to elicit the knowledge, skills, and abilities defined in the student model. Finally, the evidence model links the interactive environment (task model) with an individual candidate's proficiencies (student model) by dictating how interactions in a simulation are captured, scored, and evaluated. Specifically, the evidence model consists of two components: an evaluation rule and a measurement model (Mislevy et al., 2003). The evaluation rule (or scoring rubric) defines how observable variables are extracted from the complex test-related responses or the chronological log file; the measurement model (e.g., item response model) links these extracted variables to the latent traits identified in the student model (Mislevy et al., 2002).

This ECD framework facilitates the integration of process data analytics to further bolster validity. As exemplified by the ABR Core Exam, interactive, tool-integrated platforms generate granular log data capturing a candidate's cognitive trajectory (e.g., keystrokes, tool navigation, and search patterns). While conventional examinations remain blind to these cognitive trajectories, examination authorities can utilize advanced analytical methods (e.g., response-time modeling and sequence analysis) to extract actionable psychometric information from such voluminous data. For example, by applying hierarchical modeling frameworks (van der Linden, 2007), examination authorities can operationalize processing efficiency as a distinct evidentiary component. Furthermore, sequence analysis techniques such as Levenshtein distance (Levenshtein, 1966) enable test systems to compare a candidate's individual responses against an idealized, expert-defined course of action (He et al., 2021). By integrating these advanced analytics, examination authorities can verify that credentialed candidates have not only arrived at correct solutions but have done so by adopting the safe, efficient, and procedurally correct methods demanded by the profession. Thus, ECD provides a robust defense against the two greatest threats to validity in high-stakes testing: construct underrepresentation and construct-irrelevant variance (Messick, 1995) .

V. Strategic Recommendations for National Licensure Examinations

Based on the preceding analysis, this study proposes two primary frameworks for modernizing national licensure examinations: CAT and ECD. While these represent independent systems with distinct strategic objectives, they can be strategically integrated based on institutional needs and resource readiness. If the primary goal of an examination authority is to enhance measurement efficiency and administrative flexibility, the implementation of CAT offers an optimal solution. Conversely, if the objective is to improve ecological validity and assess complex professional competencies, the adoption of an ECD-based simulation model is paramount. For high-stakes licensure programs requiring both precision and authenticity, a comprehensive integration of CAT and ECD provides the most robust assessment architecture. Crucially, the adoption of these systems does not require a simultaneous launch; rather, authorities may choose to implement them sequentially or concurrently, depending on their specific technological maturity and operational constraints.

Specifically, national examination authorities could adopt a tiered assessment architecture that integrates CAT and ECD into a unified framework. For instance, CAT functions as the intake engine, efficiently identifying a candidate’s initial proficiency level. This proficiency estimate then serves as the 'strategic anchor' for the subsequent ECD-based simulation, determining the appropriate complexity and cognitive demands of the tasks presented. By aligning CAT’s ability estimation with ECD’s high-fidelity task design, the system ensures that complex professional simulations are adaptively calibrated to the candidate’s latent trait. This tiered integration empowers examination authorities to maintain operational efficiency while evaluating professional competencies. The following sections provide a detailed examination of the implementation of CAT and ECD and explore how each framework addresses challenges.

A. Applications of Adaptive Testing

The theoretical frameworks and empirical benchmarks examined in this study provide insights to assist examiners and policymakers in modernizing professional and

technical licensure examinations. For national examination authorities, the transition to CAT is particularly urgent and impactful in high-volume licensure programs, such as those for registered nurses, social workers, and nutritionists. In traditional linear testing models, test integrity relies on item disposability: Each unique test form is used only once before being released publicly (van der Linden & Pashley, 2010). Although this one-time linear testing approach ensures transparency and mitigates the risk of item recycling, it is inefficient and difficult to adapt for flexible scheduling and longitudinal equating.

A comprehensive transition to CAT offers the optimal solution. By applying a variable-length stopping rule, CAT can reduce average examination times by 50% without compromising measurement precision (Weiss, 2011). Furthermore, because CAT dynamically assembles tests from an item bank in real time to create a unique test for each examinee, examination authorities can move away from rigid annual or biannual administration and adopt year-round testing models. Additionally, because all items in a CAT item bank are mapped to a common metric, scores derived from different administrations are directly comparable, facilitating the longitudinal tracking of latent traits within the candidate population.

1. Practical Concerns Regarding CAT

Although CAT offers numerous advantages, it presents several practical challenges. The primary concerns in transitioning to CAT for national licensure examinations involve maintaining test security, ensuring measurement fairness, and addressing operational stakeholder concerns. Regarding test security, unlike traditional linear testing models that rely on item disposability, CAT depends on the continual reuse of items from high-quality item banks. Consequently, repetitive item exposure increases the risk of item leakage. In high-stakes testing, test-preparation organizations or coordinated candidate networks can systematically harvest items, rapidly compromising test banks and inflating scores. Therefore, the validity of CAT-based licensure systems is predicated on rigorous exposure controls. One such control is the Sympson–Hetter method, which restricts the frequency of highly informative items using probabilistic models (Sympson & Hetter, 1985). Alternatively, examiners can implement procedures such as two-stage exposure control (Chao & Chen, 2023) or test

overlap rate control (e.g., Chen et al., 2014) to enhance test security. In combination with sufficient item bank refreshment, these mechanisms can effectively mitigate the risk of item leakage, thereby ensuring the validity of the assessment.

Traditional linear testing defines equity as “test form identity”—administration of the same questions to every candidate. This conventional view poses a significant challenge for CAT implementation, as candidates and stakeholders often perceive it as unfair that examinees are administered different sets of items. However, relying on this definition involves a measurement fallacy that leads to systemic inequality: linear testing provides precise results for average candidates but yields large measurement errors for those at the ability margins (Wainer, 2000). Accordingly, examination authorities adopting CAT must promote fairness as equal precision for measurement. By applying variable-length, fixed-precision stopping rules, CAT flattens the curve of conditional SEM, ensuring that pass/fail decisions are made with consistent precision for all candidates, regardless of test length. This standard of equal precision ensures that tests are truly fair and that no candidate’s score is compromised by structural inefficiencies in the test design (van der Linden & Glas, 2010).

Finally, examination authorities must address operational challenges regarding test transparency, candidate adaptation, and scoring standardization. The dynamic, algorithmic nature of CAT can be perceived as a "black box" by stakeholders, potentially undermining trust in the scoring process if not properly managed (Zenisky & Sireci, 2002). Furthermore, candidates accustomed to linear testing—where they can skip or revisit items—may experience anxiety when transitioning to adaptive environments, where item responses are final (Wise & Kingsbury, 2000). To address these concerns, authorities must move beyond technical implementation to develop communication strategies that educate stakeholders on the scoring logic. By ensuring that CAT is presented as a precise and standardized measurement tool rather than an opaque system, authorities can facilitate a smoother transition and maintain public confidence in licensure outcomes.

2. Strategic Roadmap for CAT Implementation

Based on the systematic analysis of the three international examination frameworks discussed previously, this study proposes a comprehensive roadmap to

guide the transition to CAT within national licensure systems. To facilitate this transition for high-volume national examinations, examination authorities should adopt a multiphase roadmap comprising the following four interdependent phases. Importantly, the progression through these stages does not follow a predefined or rigid timeline; rather, it is dynamically determined by the institutional readiness and operational maturity of the specific examination authority. Detailed analyses of the implementation challenges, institutional constraints, and cost-benefit considerations associated with each of these phases are further discussed in the subsequent sections of this study. The strategic focuses and core objectives of these four sequential phases are detailed as follows:

a. Item banking phase: Examination authorities must generate a sufficient item pool and establish a sustainable, long-term item generation procedure using techniques such as automated item-cloning templates. New items are invisibly seeded into existing test forms, allowing for seamless calibration through pilot tests with real candidates before they are added to the active CAT pool.

b. Adversarial simulations phase: Before CAT is adopted, exhaustive adversarial simulations must be conducted to tune the hyperparameters of testing algorithms and ensure that CAT systems are valid. Such fine-tuning requires determining the optimal stopping rules (precision thresholds) and adjusting exposure control parameters under ecologically valid conditions (e.g., by simulating organized cheating or rapid guessing behaviors) to ensure the system's robustness before operational launch.

c. Operational maintenance phase: Once CAT has been adopted, examination authorities must continually monitor test bank quality, item refreshment cycles, and changes in candidate ability distribution. This phase focuses on routine quality assurance, requiring examiners to detect item parameter drift or potential leaks to not only maintain the integrity of the CAT pool but also to enable the continuous validation and fine-tuning of CAT parameters.

d. Digital immune system: To preempt security threats, licensure systems must transcend reactive monitoring by constructing a "digital immune system" that leverages process data. For instance, by algorithmically analyzing response-time distributions alongside item responses, CAT systems can autonomously detect aberrant behaviors such as rapid correct responses to complex items—behavior that indicates

prior knowledge of the test or item. This proactive evaluation enables CAT systems to automatically quarantine compromised items and flag anomalous candidate profiles before the test's construct validity is undermined.

3. Institutional Alignment and Implementation Challenges

The proposed transition to CAT within national licensure examinations must be situated within the unique institutional and cultural context of the local examination system. Each phase of the roadmap necessitates a strategic alignment of psychometric innovation with existing administrative realities. Specifically, within the prevailing context of the local licensure system, the push for digital transformation must navigate long-standing institutional norms—ranging from rigid statutory interpretations of "test equity" to the logistical challenges of high-volume candidate cohorts. Taking the current domestic landscape as a primary example, the implementation of each phase must, at a minimum, address the following critical considerations (though the actual requirements are not limited to these points).

For item bank development, authorities must reconcile CAT's requirement for a massive item pool with the rapid item attrition inherent to traditional linear testing frameworks (e.g., PBT) because of mandatory public disclosure policies. Alongside this policy reconciliation, sustainable item generation strategies are imperative to maintain bank viability under high-stakes conditions. Domestically, the absence of pre-testing and the mandatory public disclosure of test items hinder CAT implementation. To obtain item parameters, authorities can seed unscored pretest items within current computer-based tests. To prevent rapid item bank depletion, policies must shift from full disclosure to partial or non-disclosure. If strict statutory regulations prohibit these adjustments, the only viable alternative is to utilize AIG to create items adaptively in real time and abandon them immediately post-administration (see Future Directions 1: Automated Item Generation for details on AIG).

During the system simulation phase, authorities must incorporate empirical validation to address the deeply rooted cultural perception of fairness as form identity. The CAT algorithms must be optimized to ensure that any variation in test content or length can withstand scrutiny under the lens of procedural equality. Furthermore, algorithms should be fine-tuned not only to maximize test precision but also to provide

a smooth and intuitive testing experience, such as by adjusting initial item difficulty to prevent early candidate frustration and minimizing test-induced anxiety.

Regarding operational maintenance, the primary institutional challenge lies in shifting from a periodic to a continuous quality assurance model. This transition requires administrative realignments to support routine monitoring of item bank quality, refreshment cycles, and parameter drift. Authorities must ensure that examiners are empowered and technically equipped to detect potential leaks and validate CAT parameters in real-time, moving beyond traditional manual auditing toward a more automated, data-driven oversight framework.

Finally, the deployment of an automated monitoring system must navigate the legal complexities of administrative due process. To ensure the legitimacy of licensure outcomes, algorithmic detections must function as evidentiary support for human-led committees rather than as autonomous decision-makers, thereby maintaining the balance between technological efficiency and legal accountability.

Ultimately, the feasibility of implementing CAT within a high-stakes national context depends on a rigorous evaluation of resource allocation and institutional readiness. Specifically, within the prevailing domestic landscape, transitioning to advanced models such as year-round testing requires substantial capital investment and human resource restructuring. For instance, a multiphase digital transformation initiative involves significant budgetary commitments for self-built testing facilities and the maintenance of high-specification IT infrastructure, including uninterruptible power systems and cybersecurity auditing.

From a cost-benefit perspective, while the initial expenditure is high, the long-term gains include a large amount of reduction in average testing time and in paper-based administrative workloads, and enhanced assessment precision that mitigates candidate fatigue. However, the successful scaling of CAT—particularly for high-volume licensure categories—remains contingent on addressing the acute shortage of specialized psychometric personnel and optimizing the geographical distribution of certified testing seats. Therefore, a sustainable transition must integrate fee structure adjustments with a clear demarcation of administrative responsibilities to ensure that technological modernization is both fiscally viable and socially equitable.

In brief, to overcome budget and personnel constraints, authorities should adopt a

phased implementation strategy. Initial deployment should target high-volume or high-risk examinations. Moreover, a centralized CAT platform provides economies of scale. Once the core system is established, it can serve all examination categories without redundant development, making long-term resource requirements lower than anticipated. Finally, inter-ministerial cost-sharing, such as co-funding the nursing licensure examination with the Ministry of Health and Welfare, can further alleviate financial burdens. Although domestic cohort sizes are smaller than those in the United States, CAT can achieve long-term economic viability by reducing test administration time, venue leasing, and invigilation costs.

In conclusion, while significant practical challenges and institutional hurdles remain to be overcome, it is anticipated that by transitioning to CAT through the aforementioned four stages, national examination authorities can modernize large-scale licensure programs through state-of-the-art measurement systems, ensuring operational efficiency, robust security, and psychometric fairness for all candidates.

B. Enhancing Ecological Validity Through ECD

Although CAT is efficient and provides an optimal solution for high-volume examinations, a different strategy is required to optimize licensure examinations for highly specialized and procedural programs, such as medicine, architecture, law, and structural engineering. For these fields, the primary threat to measurement validity is construct underrepresentation, which occurs when assessments fail to capture the interactive and procedural nuances of professional practice. Consequently, national examination authorities must comprehensively integrate authentic professional environments into their assessments; digitalization alone is insufficient. To fulfill its social contract and protect public safety, the licensure ecosystem must transition from validating rote knowledge (the “knows” tier) to assessing a candidate’s ability to respond to dynamic scenarios (the “shows how” tier). In the domestic context, this shift is particularly crucial for maintaining the "professional competence" standard required by the Professional and Technical Personnel Examination Act.

National examination authorities should prioritize the transitions to dynamic assessment models that evaluate procedural judgment over time rather than static evaluations. Examination authorities should also implement test scenarios that require

candidates to independently devise solutions without assistance from predefined answers. Another key component of such systems is synthetic time management, in which a virtual clock forces candidates to manage temporal pressures and ensures that the timing of an intervention is as psychometrically meaningful as the intervention itself. For instance, in medical licensure, candidates could be required to manage the condition of a virtual patient whose state responds continually to a candidate's sequence of orders, capturing the authentic, nonlinear reasoning required in clinical practice.

National licensure systems should also adopt a tool-first integration policy to ensure that credentials reflect the technological and informational demands of the contemporary workplace. Such a policy should reduce construct-irrelevant variance by providing candidates with the corresponding digital tools during examinations that are used in professional practice. This approach requires embedding specialized software or databases directly into testing platforms to assess procedural engagement. For instance, in legal licensure, this might involve closed-platform retrieval systems in which candidates must query searchable databases to establish precedents rather than relying on memory alone. Similarly, in architecture and engineering licensure, computer-aided design or building information modeling modules should be employed to enable candidates to manipulate three-dimensional models. By adopting such localized instruments and verifying designs on the spot, examination authorities can ensure that professionals have demonstrated the interactive competencies required in modern practice.

To address practical concerns regarding variability in digital tools in real-world settings, examination authorities must implement a "functional standardization" protocol. This involves providing functionally equivalent digital tools within the testing environment—ensuring that while specific interfaces may vary, they remain psychometrically identical in their measurement purpose. Consequently, candidates are evaluated on their mastery of primary professional functions rather than software-specific nuances. By calibrating these tools to represent industry standards, authorities can maintain psychometric fairness and ensure that no candidate is disadvantaged by technological disparities.

Drawing upon the structural components of ECD and the analysis of the three

international systems previously examined, this study organizes a four-phase roadmap for implementing advanced digital assessments in national licensure examinations. This integrated framework aims to guide the transition toward high-fidelity professional simulations in a manner that is both technologically feasible and psychometrically robust. The strategic focuses and core objectives of these four sequential phases are detailed as follows:

1. Environment architecture phase (task model design): Based on established medical benchmarks, examination authorities must define an ECD-based task model using digital tools, interactive simulation interfaces, and algorithms (e.g., virtual clocks or dynamic case trajectories) to assess the desired professional competencies.

2. Evidence model adjustment phase: To facilitate the ECD framework, psychometricians must establish rigorous evaluation rules. This phase involves extracting observable variables from complex log files generated in simulated environments, ensuring that every captured interaction from tool usage to decision timing is accurately mapped to a candidate's latent proficiencies.

3. Adversarial testing and fidelity validation phase: Before a test system is adopted, it must undergo exhaustive adversarial testing to ensure that interactive elements do not introduce construct-irrelevant variance and that automated scoring produces results consistent with expert judgments.

4. Process-oriented data integration phase: Once implemented, the system should transition from outcome-based to process-oriented measurements. By using transaction logs, authorities can demonstrate that candidates have used safe, efficient, and procedurally correct methods to arrive at their solutions to test questions.

VI. Future Outlook and Conclusions

Adaptive testing and high-fidelity simulations can substantially enhance the precision and ecological validity of professional licensure examinations. However, the successful implementation of these computer-native architectures should not be viewed as an endpoint but rather as a foundation for ongoing technological and psychometric development. To ensure the long-term sustainability and diagnostic accuracy of national examination systems, licensure authorities must engage with cutting-edge approaches to intelligent assessment. This final chapter examines two key directions

that represent the state of the art in licensure ecosystems: automated item generation and cognitive diagnostic assessment.

A. Future Direction 1: Automated Item Generation

Implementing CAT and high-fidelity simulations can address fundamental challenges of precision and validity in national licensure examinations. However, these computer-native testing systems are subject to a major vulnerability: item starvation (Parshall et al., 2002). In year-round national licensure programs, item pools are rapidly depleted due to overexposure and the retirement of compromised items. Sustaining these pools requires the continuous generation of numerous psychometrically sound items, a process that is often cost-prohibitive and unsustainable when authorities depend exclusively on manual item drafting by subject matter experts (Gierl & Haladyna, 2013).

To address this challenge, examination authorities must transition from manual item drafting to algorithmic AIG, a multidisciplinary process that integrates cognitive psychology, advanced psychometrics, and natural language processing (Gierl & Haladyna, 2013). Subject matter experts and psychometricians working with AIG must jointly develop cognitive models to define the complex network of variables, constraints, and logical relationships inherent in professional tasks (Irvine & Kyllonen, 2002). Once a cognitive model has been defined, it can be transformed into a formal item template. An AIG algorithm can subsequently be used to systematically manipulate the defined variables within the template and rapidly generate thousands of structurally unique but psychometrically equivalent isomorphic items. For instance, in a pharmacology examination, a single AIG template governing drug–interaction logic could generate thousands of unique patient vignettes by systematically altering a patient's age, weight, and secondary comorbidities as well as contraindicating medications. Currently, domestic examination authorities utilize 'derived items' (衍生題) to expand item pools. This practice serves as a foundational step toward AIG. Upgrading this approach to a fully algorithmic AIG will further maximize item production efficiency.

AIG functions optimally when items are generated during the test itself. In other words, an item presented to a candidate should ideally be algorithmically generated by

the testing system based on the candidate's current performance or ability estimate before it is rendered on the screen. Because such items are individualized, they effectively curtail item harvesting and leakage, increasing test security (Bejar, 2002). However, the implementation of AIG must align with strict statutory frameworks, such as the Examination Act in Taiwan. Given the imperative of public trust, AIG cannot replace human examiners; it must function as an augmented tool. While AIG algorithms rapidly generate candidate items based on predefined templates, subject matter experts retain full authority over the final substantive review and approval. This "human-in-the-loop" approach complies with legal regulations, preserving assessment integrity while accelerating item production.

B. Future Direction 2: Cognitive Diagnostic Assessment

Although AIG can secure assessment items, CDA is required to enhance the utility of scoring for candidates. Traditional assessments typically yield a single overall score that is effective for ranking candidates and making binary pass/fail decisions, but they cannot provide detailed diagnostic information to reveal where and why a candidate was unsuccessful. CDA addresses this limitation by providing multidimensional categorical latent variables to determine whether a candidate has mastered a discrete set of fine-grained attributes essential to the profession (Rupp et al., 2010).

By applying CDA to the response vectors and granular log data generated by computer-native assessments, examination authorities can construct a comprehensive mastery profile for every candidate. Such profiles assist candidates in identifying and addressing their weaknesses, particularly for unsuccessful candidates (retakers). Instead of receiving only a unidimensional numerical score (e.g., "score: 58; cutoff score: 60"), a candidate evaluated using CDA receives a detailed diagnostic profile indicating non-mastery of specific areas (e.g., pediatric dosage calculations or infection control). This formative feedback transforms summative licensure examinations into educational interventions that guide targeted remediation, thereby expediting retakes and strengthening the professional competencies (Templin & Bradshaw, 2013).

However, providing such detailed feedback in high-stakes licensure contexts necessitates a robust security framework. To safeguard test integrity, formative feedback should be reported at the level of latent competency dimensions or cognitive

attributes rather than revealing specific item-level content or correct answers. By focusing on these high-level diagnostic profiles, examination authorities can facilitate candidate remediation while minimizing the risk of item harvesting or strategic "test-taking" behaviors by coaching organizations. Domestically, the examination authorities' current practice of tagging past items to evaluate foundational competencies aligns conceptually with CDA. Transitioning to formal CDA models will systematically provide multidimensional diagnostic feedback to professional educational systems.

C. Necessity of Test Modernization

On the basis of a unified theory of validity (Messick, 1995), this study argues that national licensure examinations are essential to fairly and validly assess candidates. National examination authorities must uphold their social contract with the public: They alone possess the authority to restrict entry into a profession to protect the common weal. When a board licenses a professional, such as an attorney or a physician, it provides an implicit guarantee that the licensed individual has the ability to defend civil liberties or protect human life. Licensing professionals based on examinations that are disconnected from the realities of practice is both inefficient and risky. If examination authorities depend on such outmoded, low-fidelity assessments that cannot measure core professional competencies, they fail in their ethical duty. Therefore, as Kane (2013) argued, modernizing the licensure ecosystem is essential to preserve public well-being.

AIG can secure testing, and the implementation of CDA can transform licensure examinations from static gatekeepers into active instructors that guide professional improvement. By adopting these advanced assessment technologies, national examination authorities can fulfill their mandate and transition from merely administering tests to cultivating a highly competent, thoroughly tested, and deeply trusted professional workforce.

References

American Board of Radiology. (n.d.). Diagnostic radiology certification: Eligibility requirements. Retrieved March 1, 2026, from https://www.theabr.org/get-

certified/diagnostic-radiology/

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Becker, G. J., & Dunnick, N. R. (2008). Intended consequences of computer-based core and certifying examinations in diagnostic radiology. *American Journal of Roentgenology*, *191*(5), 1302-1305. https://doi.org/10.2214/AJR.08.1232

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Erlbaum Associates.

Bennett, R. E. (2005). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 201-217). Wiley. https://doi.org/10.1002/9780470712993.ch11

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. https://doi.org/10.1177/014662168200600405

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1-20. https://doi.org/10.1007/s11336-014-9401-5

Chang, H. H., & Ying, Z. (1999). *A*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211-222. https://doi.org/10.1177/01466219922031338

Chao, H. Y., & Chen, J. H. (2023). Controlling the minimum item exposure rate in computerized adaptive testing: A two-stage Sympson–Hetter procedure. *Applied Psychological Measurement*, *47*(7-8), 460-477. https://doi.org/10.1177/01466216231209756

Chen, J. H., & Chao, H. Y. (2025). Maximin criterion for item selection in computerized adaptive testing. *Behavior Research Methods*, *57*(7), 180. https://doi.org/10.3758/s13428-025-02673-8

Chen, S. Y., Lei, P. W., Chen, J. H., & Liu, T. C. (2014). General test overlap control: Improved algorithm for CAT and CCT. *Applied Psychological Measurement*, *38*(3), 229-244. https://doi.org/10.1177/0146621613513494

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369-383. https://doi.org/10.1348/000711008X304376

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational easurement*, *34*(2), 141-161. https://doi.org/10.1111/j.1745-3984.1997.tb00511.x

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2002). An examination of the contribution of computer-based case simulations to the USMLE Step 3 examination. *Academic Medicine*, *77*(Suppl. 1), S80-S82. https://doi.org/10.1097/00001888-200210001-00026

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

Dickison, P., Haerling, K. A., & Lasater, K. (2019). Integrating the National Council of State Boards of Nursing clinical judgment model into nursing educational frameworks. *Journal of Nursing Education*, *58*(2), 72-78. https://doi.org/10.3928/01484834-20190122-03

Dillon, G. F., Boulet, J. R., Hawkins, R. E., & Swanson, D. B. (2004). Simulations in the United States Medical Licensing Examination (USMLE). *Quality and Safety in Health Care*, *13*(Suppl 1), i41-i45. https://doi.org/10.1136/qshc.2004.010025

Eggen, T. J. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249-261. https://doi.org/10.1177/01466219922031365

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of the American Medical Association*, *287*(2), 226-235. https://doi.org/10.1001/jama.287.2.226

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gunderman, R., Williamson, K., Fraley, R., & Steele, J. (2001). Expertise:

Implications for radiological education. *Academic Radiology*, *8*(12), 1252-1256. https://doi.org/10.1016/S1076-6332(03)80708-0

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Han, K. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*(7). https://doi.org/10.3352/jeehp.2018.15.7

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, Article 104170. https://doi.org/10.1016/j.compedu.2021.104170

Hollingsworth, C. L., Wriston, C. C., Bisset, G. S., Strife, J. L., Bosma, J. L., Gerdeman, A. M., & Becker, G. J. (2010). American Board of Radiology certifying examination: Oral versus computer-based format. *American Journal of Roentgenology*, *195*(4), 820-824. https://doi.org/10.2214/AJR.09.3618

Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part II: Theory and practices of test equating in item response theory. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(3), e1543. https://doi.org/10.1002/wics.1543

Huang, H. Y. (2023). Modeling rating order effects under item response theory models for rater-mediated assessments. *Applied Psychological Measurement*, *47*(4), 312-327. https://doi.org/10.1177/01466216231174566

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Lawrence Erlbaum Associates.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707-710.

Lin, C. J., & Spray, J. A. (2000). Effects of item-selection criteria on classification testing with the sequential probability ratio test (Research Report 2000-8). ACT, Inc.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135. https://doi.org/10.1177/01466210122031957

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), S63-S67. https://doi.org/10.1097/00001888-199009000-00045

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3-62. https://doi.org/10.1207/S15366367ME0101_02

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, *15*(4), 363-389. https://doi.org/10.1207/S15324818AME1504_03

National Council of State Boards of Nursing. (2023a, April 3). *NCSBN launches Next Generation NCLEX exam*. https://www.ncsbn.org/news/ncsbn-launches-next-generation-nclex-exam

National Council of State Boards of Nursing. (2023b). *2023 NCLEX-RN test plan*. https://www.ncsbn.org/public-files/2023_RN_Test%20Plan_English_FINAL.pdf

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National

Academy Press.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schuwirth, L. W. T., van der Vleuten, C. P. M., & Donkers, H. H. L. M. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education*, *30*(1), 44-49. https://doi.org/10.1111/j.1365-2923.1996.tb00716.x

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal*, *326*(7390), 643-645. https://doi.org/10.1136/bmj.326.7390.643

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*(4), 405-414. https://doi.org/10.3102/10769986021004405

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, *24*(5), 5-11. https://doi.org/10.3102/0013189X024005005

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). Navy Personnel Research and Development Center.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251-275. https://doi.org/10.1007/s00357-013-9129-4

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Routledge.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*(1), 21-29. https://doi.org/10.1177/01466219922031149

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory*

and practice (pp. 27-52). Kluwer.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*(3), 283-302. https://doi.org/10.1111/j.1745-3984.2005.00015.x

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287-308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8_1

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*(4), 17-27. https://doi.org/10.1111/j.1745-3992.1998.tb00632.x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473-492. https://doi.org/10.1177/014662168200600408

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1-27.

Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part I: Usability and pilot testing. *Nurse Educator*, *34*(2), 56-59. https://doi.org/10.1097/NNE.0b013e3181990849

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*(1-2), 135-155.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). American Council on

Education/Praeger.

Zara, A. R. (1999). Using computerized adaptive testing to evaluate nurse competence for licensure: Some history and forward look. *Advances in Health Sciences Education*, *4*(1), 39-48. https://doi.org/10.1023/A:1009866321381

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, *15*(4), 337-362. https://doi.org/10.1207/S15324818AME1504_02