TPM Vol. 32, No. 3, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

DEVELOPING AND VALIDATING A HIGHER-ORDER THINKING AND INNOVATION COMPETENCY ASSESSMENT: PSYCHOMETRIC EVIDENCE FOR A ROBUST ADAPTIVE TESTING ITEM BANK

SORAYA SANMUANG*
KHON KAEN UNIVERSITY, THAILAND

PRAKITTIYA TUKSINO *
KHON KAEN UNIVERSITY, THAILAND

KORNCHAWAL CHAIPAH
KHON KAEN UNIVERSITY, THAILAND

Abstract
This study aimed to develop and validate an assessment tool for evaluating higher-order thinking and innovation competency among Thai upper secondary students, with the ultimate goal of creating a high-quality item bank for future Computerized Adaptive Testing (CAT) applications. The instrument was designed as a situational judgment test with four response options aligned with Bloom's revised taxonomy. The development process included expert validation, pilot testing, and large-scale administration to a multi-stage random sample of 714 students from four provinces in Northeastern Thailand. Confirmatory Factor Analysis (CFA) was conducted to verify the construct validity, and Item Response Theory (IRT) with the Graded Response Model (GRM) was applied to assess item parameters and test information. Reliability was examined through Cronbach's alpha analyses. The CFA indices met recommended thresholds (χ^2 = 37.343, df = 25, p = .0535; RMSEA = 0.025; SRMR = 0.022; CFI = 0.992; TLI = 0.975.), indicating good model fit. GRM analysis revealed high discrimination parameters (a = 0.82–2.15) and well-ordered threshold parameters. Cronbach's alpha was 0.927. Competency cut scores were determined from Wright Map analysis of GRM threshold parameters, classifying students into four proficiency levels: (1) Insufficient competency ($\theta \leq -1.31$) ; (2) Emerging competency ($-1.31 < \theta \leq -0.35$) ; (3) Competent ($-0.35 < \theta \leq 0.56$); and (4) High competency ($\theta > 0.56$)
Keywords: Competency assessment, higher-order thinking and innovation, cut score, confirmatory factor analysis (CFA), Item Response Theory (IRT), Graded Response Model (GRM), item bank development

INTRODUCTION

Higher-order thinking and innovation competencies are increasingly recognized as core 21st-century skills by leading international and national education frameworks. The OECD Learning Compass 2030 highlights transformative competencies such as creating new value, reconciling tensions and dilemmas, and taking responsibility as essential for thriving in uncertain and complex futures (OECD, 2019). Likewise, UNESCO's transversal competencies framework emphasizes critical and creative thinking, problem-solving, and the ability to adapt knowledge to diverse and authentic contexts (UNESCO, 2016).

In Europe, the European Reference Framework on Key Competences for Lifelong Learning identifies creativity, critical thinking, initiative, and problem-solving as fundamental for personal fulfillment, employability, and active citizenship in the knowledge society (European Commission, 2019). Similarly, the World Economic Forum's Future of Jobs Reports (2023, 2025) consistently list analytical thinking, creativity, complex problem-solving, and innovation as among the top skills demanded by employers worldwide. Empirical labor market studies confirm that these competencies directly influence career readiness and are increasingly decisive for recruitment, career progression, and organizational innovation capacity.

In Thailand, recent education reforms have shifted toward a competency-based curriculum that prioritizes creativity, innovation, and problem-solving in real-world contexts (Ministry of Education, 2023). Higher-order thinking and innovation competencies enable students to analyze, evaluate, synthesize, and generate innovative solutions to complex challenges while fostering adaptability, resilience, and lifelong learning. In an era of rapid technological advancement, socio-economic transformation, and global challenges, such competencies are critical not only for individual success but also for driving national competitiveness and sustainable development.

Given these imperatives, there is an urgent need for valid, reliable, and context-driven assessment tools capable of capturing these multifaceted skills. This study responds to that need by developing and validating a scenario-based, context-specific item set designed to measure higher-order thinking and innovation competencies among Thai upper secondary students. These psychometrically sound items, analyzed through Confirmatory Factor Analysis (CFA) and the Graded Response Model (GRM), form a high-quality item bank that can be integrated into future Computerized Adaptive Testing (CAT) systems for more precise, efficient, and individualized competency assessment.

Research Objective

The objectives of this study were to:

1) Develop a set of context-based, scenario-driven test items to measure higher-order thinking and innovation competencies among Thai upper secondary students, aligned with national and international competency frameworks.
2) Validate the structural model of the competencies through Confirmatory Factor Analysis (CFA) to ensure construct validity.
3) Analyze the psychometric properties of the test items using the Graded Response Model (GRM) under the framework of Item Response Theory (IRT).
4) Examine the reliability of the test using both Cronbach's alpha to ensure measurement precision across the ability spectrum.
5) Establish a high-quality item bank that can serve as the foundation for future Computerized Adaptive Testing (CAT) applications in assessing higher-order thinking and innovation competencies.

REVIEW OF RELATED LITERATURE

1) Higher-Order Thinking and Innovation Competency Model

This study developed a measurement model for higher-order thinking and innovation competencies among upper secondary students. The model comprises four sub-competencies, synthesized from both national and international competency frameworks (CBE Thailand, 2022; European Commission, 2018; OECD, 2018, 2019; UNESCO, 2015; World Economic Forum, 2020) and supported by recent empirical studies.

1) Critical Thinking

Critical thinking refers to the ability to analyze, evaluate, and synthesize information to make reasoned decisions and judgments. It involves questioning assumptions, identifying bias, and applying logic to problem contexts (European Commission, 2018; OECD, 2019). Duong (2022) emphasizes its importance in navigating complex information landscapes, noting that learners with strong critical thinking skills are better equipped to adapt to rapidly changing environments.

2) Creative Thinking

Creative thinking involves generating novel and valuable ideas, approaching problems from unconventional perspectives, and fostering innovation (OECD, 2019; WEF, 2020). Rusmini (2025) proposed the Connected Creative Problem Solving model, which integrates scientific reasoning with collaborative discourse to enhance creativity in educational contexts. Similarly, Plailek et al. (2022) demonstrated that hybrid learning approaches with peer coaching can significantly improve students' creativity and innovation in English language learning.

3) Systemic Thinking

Systemic thinking is the ability to understand interrelationships within complex systems, recognize patterns, and anticipate the implications of decisions (UNESCO, 2015). Upapong and Poonputta (2025) developed a validated tool to assess systemic thinking among upper primary students in Thailand, highlighting its role in preparing learners for interconnected global challenges.

4) Problem Solving

Problem solving refers to identifying challenges, generating alternative solutions, evaluating options, and implementing effective strategies (OECD, 2019). Research indicates that structured pedagogical models—such as hybrid learning combined with peer coaching—can foster both problem-solving and creative capacities (Plailek et al., 2022).

2) Synthesis of Competency Frameworks

The four sub-competencies align closely with global frameworks that highlight 21st-century skills as essential for lifelong learning and workforce readiness (European Commission, 2018; OECD, 2018, 2019; UNESCO, 2015; WEF, 2020). Thailand's national competency-based education framework (CBE Thailand, 2022) similarly prioritizes higher-order cognitive and innovative abilities.

From a measurement perspective, the integration of psychometric models such as Confirmatory Factor Analysis (CFA) for construct validation and the Graded Response Model (GRM) for item calibration provides robust evidence of validity and reliability. The validated item bank resulting from this process serves as a high-quality resource for future adaptive testing applications, enabling precise and efficient assessment tailored to individual learner profiles.

3) Item Response Theory and the Graded Response Model

The foundation of this study is Item Response Theory (IRT), which offers a probabilistic framework for modeling the interaction between test-taker ability and item characteristics (Embretson & Reise, 2013). Among the polytomous IRT models, the Graded Response Model (GRM), introduced by Samejima (1969), is particularly suited for competency assessments with ordered response categories. GRM provides two types of parameters: discrimination (α), indicating the ability of an item to differentiate between students of different ability levels, and threshold (β), representing the ability level required to transition from one response category to the next (Baker & Kim, 2017). Numerous studies have applied GRM in educational measurement to ensure precise ability estimation and to support defensible score interpretations.

4) Confirmatory Factor Analysis and Construct Validation

Confirmatory Factor Analysis (CFA) is widely used to validate the factor structure of competency assessments (Brown, 2015). The present study applied CFA to confirm that the items aligned with the theoretical constructs of Higher-Order Thinking and Innovation. Fit indices were interpreted according to Hu and Bentler's (1999) recommended thresholds (e.g., RMSEA < 0.06, CFI ≥ 0.95, TLI ≥ 0.95), ensuring that the measurement model reflected the intended competencies. Establishing construct validity through CFA strengthens the argument that the developed item bank measures the competencies it purports to assess.

5) Reliability in Educational Measurement

Test reliability, as measured by Cronbach's alpha (Cronbach, 1951), indicates the internal consistency of the assessment. These values exceed the recommended 0.80 threshold for high-stakes educational assessments (Nunnally & Bernstein, 1994), ensuring score stability and interpretability.

6) Cut Score Determination Using Wright Maps

Cut score setting is a critical process for classifying student proficiency levels in a valid and interpretable manner. In this study, the Wright Map technique (Bond & Fox, 2015) was applied to align item thresholds and person abilities on a common logit scale, facilitating evidence-based decision-making about performance boundaries. Using the Graded Response Model (GRM), threshold averages were calculated to represent the ability levels required for progression through each scoring category.

Building on established competency-level frameworks (Bates, 2019; Colby, 2021; Hess et al., 2020), four performance levels were defined: Insufficient competency,
Emerging competency, Competent and High competency

This combined approach—integrating psychometric evidence from Wright Map visualization with theoretically grounded performance descriptors—ensures that the cut scores are both statistically defensible and pedagogically meaningful. Such clearly defined levels also support transparent interpretation of results and targeted instructional planning.

METHODOLOGY

Research Design

This study employed a quantitative research design aimed at developing and validating a context-based, scenario-driven assessment tool for measuring higher-order thinking and innovation competency among Thai upper secondary students. The research followed a sequential process encompassing item development, expert validation, pilot testing, large-scale administration, and psychometric analysis using modern measurement techniques.

Participants and Sampling Procedure

A total of 714 upper secondary students from the Northeastern region of Thailand participated in the study. The sampling process utilized a multi-stage random sampling method to ensure representation across varied educational contexts and student ability levels.

Stage 1: Four provinces were randomly selected from the Northeastern region using simple random sampling to ensure geographic diversity.

Stage 2: Stratified random sampling of schools within each selected province was conducted to ensure inclusion of both urban and rural contexts, thereby capturing students with diverse learning experiences, resource availability, and socio-cultural backgrounds. The sampling strategy was purposefully designed to obtain participants whose abilities span across all four targeted competency levels of higher-order thinking and innovation. This diversity in competency distribution was essential to enable robust psychometric analyses—first, to validate the construct structure through Confirmatory Factor Analysis (CFA); second, to estimate Graded Response Model (GRM) parameters, ensuring that each item can effectively discriminate among students at different proficiency bands for precision across the entire ability continuum. Furthermore, the availability of ability data across the full spectrum supports accurate Wright Map analysis, allowing both item difficulty thresholds and student ability estimates to be plotted on a common scale. This mapping

facilitates the determination of empirically grounded cut scores for classifying students into distinct competency levels with high validity and reliability.
Stage 3: Classrooms were randomly selected within each school, and all students in the selected classes were invited to participate in the assessment to maintain the natural variation of abilities within the sample.

Instrument Development

The assessment instrument was developed based on a synthesized competency framework integrating multiple authoritative sources, including the OECD Learning Compass 2030 (OECD, 2019), UNESCO transversal competencies (UNESCO, 2016), the European Key Competences for Lifelong Learning (European Commission, 2019), the World Economic Forum's Future of Jobs Reports (WEF, 2023), and Thailand's latest competency-based education policy (Ministry of Education, 2023). This integration ensured alignment with internationally recognized 21st-century competencies and national education reform priorities.

All items were designed as scenario-based, context-rich tasks requiring students to apply higher-order cognitive processes—analysis, evaluation, and creation—consistent with the revised Bloom's taxonomy (Anderson & Krathwohl, 2001). Items targeted behavioral indicators of higher-order thinking and innovation competencies, encouraging the integration of critical thinking, problem-solving, and creativity in authentic contexts.

Each item employed a four-option multiple-choice format, anchored to a 1–4 scale representing ascending competency mastery. Items were systematically developed to represent all competency indicators, with six items per behavioral indicator, ensuring comprehensive coverage of the construct. The item set was intentionally designed to span a wide range of difficulty levels, enabling precise measurement across the full spectrum of student ability. This approach aligns with best practices in item bank development, which require a sufficiently large, diverse, and difficulty-balanced pool of calibrated items to maintain measurement precision, ensure content representativeness, and support efficient item selection in future Computerized Adaptive Testing (CAT) applications (Hambleton & Swaminathan, 1985; OECD, 2019; van der Linden, 2018).

Expert Validation and Pilot Testing

Five experts in educational measurement and instructional design reviewed the items for content validity, clarity, and alignment with the competency framework. The Content Validity Ratio (CVR) was calculated for each item to ensure adequacy before pilot testing. The pilot test was conducted with 90 students from diverse backgrounds to identify potential ambiguities and adjust item wording accordingly.

Preliminary Reliability Analysis

The pilot study with 90 upper secondary students was conducted to evaluate language clarity, cognitive demand, and preliminary reliability before the main data collection. Internal consistency was examined using both Cronbach's alpha and ordinal alpha (polychoric correlation-based), appropriate for polytomous items scored on a 1–4 scale. Results indicated satisfactory to excellent reliability, with an overall Cronbach's alpha of 0.89 all subscales exceeded the commonly accepted reliability threshold of $\alpha \geq .70$ (Nunnally & Bernstein, 1994) and ordinal $\alpha \geq .80$, indicating high internal consistency. Most items demonstrated corrected item–total correlations ≥ .30, suggesting good homogeneity within each construct. Items with correlations below .30 were revised for clarity, while category step ordering was confirmed to be monotonic, supporting the use of the 4-point scoring scale.

Participants and Ethical Considerations

The finalized assessment was administered to a main sample of 714 upper secondary students. In confirmatory factor analysis (CFA), the adequacy of the sample size is critical to obtaining stable and valid parameter estimates. According to Hair, Anderson, Black, Babin, and Tatham (2006), an appropriate sample size should be approximately 5–20 times the number of estimated parameters, with a preference for larger samples when the measurement model includes multiple latent variables and complex structures. Additionally, CFA assumes that the underlying population distribution is multivariate normal, making a large and well-distributed sample essential.

For Item Response Theory (IRT), specifically the Graded Response Model (GRM), De Ayala (2009, pp. 223–224) recommends a minimum of 500 respondents to ensure accurate estimation of discrimination and threshold parameters. The sample of 714 students in this study therefore exceeded both CFA and GRM requirements, ensuring sufficient statistical power and precision for the intended analyses.

This study received ethical approval from the Khon Kaen University Human Research Ethics Committee (Protocol ID: HE663399) in accordance with:

- The International Declaration of Helsinki principles (World Medical Association, 2013)
- Thai National Research Ethics Committee guidelines
- American Educational Research Association ethical standards (AE RA, 2011)

All participants provided informed consent through a secure digital platform with multilingual support (Thai and English). To protect participant confidentiality, rigorous data anonymization protocols were implemented, including:

- Assignment of unique identifiers replacing personal information
- AES-256 encryption for data transmission and storage
- Restricted access controls allowing only authorized research personnel to handle data
- Automatic data retention limits of five years before secure deletion

These procedures ensured compliance with both national and international ethical standards for research involving human participants, while also supporting the reliability and validity of the data for advanced psychometric analyses such as CFA, GRM parameter estimation, and Wright Map–based cut score determination.

RESULTS

1) The construct validity was examined through confirmatory factor analysis (CFA) of the measurement model for assessing the competencies of upper secondary school students.

The construct validity of the Higher-Order Thinking and Innovation model was tested with a four-factor structure comprising Critical Thinking (C), Creative Thinking (CT), Systemic Thinking (ST), and Problem Solving (PS). Model fit indices indicated an excellent fit: $\chi^2 = 37.343$, df = 25, p = .0535; RMSEA = 0.025; SRMR = 0.022; CFI = 0.992; TLI = 0.975. These results support the adequacy of the hypothesized four-factor structure, as shown in Table 1.

Table 1
Model Fit Criteria for Confirmatory Factor Analysis of the Higher-Order Thinking and Innovation Competency Model

Fit Index	**Value**	**Recommended Cut-off**	**Interpretation**
χ^2 (Chi-square)	37.343	-	Non-significant (p > .05) indicates good model fit
df	25	-	-
p-value	0.0535	> .05	Supports model fit
RMSEA	0.025	< 0.06	Excellent fit
SRMR	0.022	< 0.08	Excellent fit
CFI	0.992	> 0.95	Excellent fit
TLI	0.975	> 0.95	Excellent fit

2) Item Bank for Higher-Order Thinking and Innovation Competency

The item quality analysis of the Higher-Order Thinking and Innovation Competency bank was conducted using the **Item Response Theory (IRT)** framework, specifically the **Graded Response Model (GRM)** for polytomous items scored on a four-point scale. The analysis estimated the **discrimination parameter (a)**, also referred to as the slope parameter (α), and the **difficulty parameters (b)**, represented by **threshold parameters (β_1, β_2, β_3)**.

All 78 items in the competency bank (100%) met the selection criteria for inclusion in the final item bank. The discrimination parameters (α) ranged from 0.822 to 2.681, with a mean of 1.752, indicating that the items could effectively differentiate between examinees across medium to high ability levels. The mean threshold parameters were $\beta_1 = -1.310$, $\beta_2 = -0.353$, and $\beta_3 = 0.562$, demonstrating appropriate coverage of the ability continuum (−3 to +3 logits).

These findings confirm that the item bank is psychometrically sound and suitable for use in developing an online **Computerized Adaptive Testing (CAT)** system for student competency assessment. The full parameter estimates are presented in **Table 2**.

Table 2 The discrimination parameters (α) and threshold parameters (β_1, β_2, β_3)

Item	α	β_1	β_2	β_3	Interpretation
Item_1	1.311	-0.935	-0.099	0.927	Moderate-to-high discrimination; Covers a wide ability range.
Item_2	2.681	-2.128	-1.267	0.009	Very high discrimination; Covers a wide ability range.
Item_3	1.905	-1.829	-0.933	-0.187	High discrimination; Easy item.
…	…	…	…	…	…
Item_77	1.527	-2.066	-0.823	0.254	**Moderate-to-high discrimination**; lower thresholds indicate easier starting category.
Item_78	1.368	-2.402	-0.855	0.410	**Moderate discrimination** ($\alpha \approx 1.30$–1.50); thresholds show easier item with moderate-level final category.

3) Reliability Analysis

For each component of the competency item bank. The Higher-Order Thinking and Innovation competency comprises four components—Critical Thinking, Creative Thinking, System Thinking, and Problem Solving—covering 13 behavioral indicators and 78 items, with an overall reliability of 0.978, considered good according to psychometric standards. When examined by individual components, the reliability coefficients ranged from 0.899 to 0.943 are presented in Table 3.

Table 3

Reliability of the Higher-Order Thinking and Innovation Competency Item Bank

Competency	**Component**	**Reliability (α)**
Higher-Order Thinking and Innovation	1) Critical Thinking (C)	0.899
	2) Creative Thinking (CT)	0.943
	3) System Thinking (ST)	0.899
	4) Problem Solving (PS)	0.912
	Total	**0.978**

4) Determination of Competency Level Cut Scores from Wright Map Analysis Using Item Threshold Parameters (b1–b3) in the GRM for Higher-Order Thinking and Innovation

The cut score criteria for classifying the competency levels in higher-order thinking and innovation among upper secondary students were developed using the Item Response Theory (IRT) with the Graded Response Model (GRM). The analysis utilized the threshold parameters of the 78-item competency assessment bank—namely b_1, b_2, and b_3—which indicate the ability levels required to obtain scores in each category of an item. The mean threshold values were calculated and visualized through a Wright Map to illustrate the distribution of all three thresholds, thereby identifying the ability levels necessary to achieve higher score categories. The results revealed mean threshold values of b_1 = –1.39, b_2 = –0.71, and b_3 = 0.27, which were then applied as cut scores for classifying students' competency levels as follows: - $\theta \leq -1.31$: Level 1 – Insufficient competency, $-1.31 < \theta \leq -0.35$: Level 2 – Emerging competency, $-0.35 < \theta \leq 0.56$: Level 3 – Competent, $\theta > 0.56$: Level 4 – High competency. These results are presented in Figure 1 and Table 4.

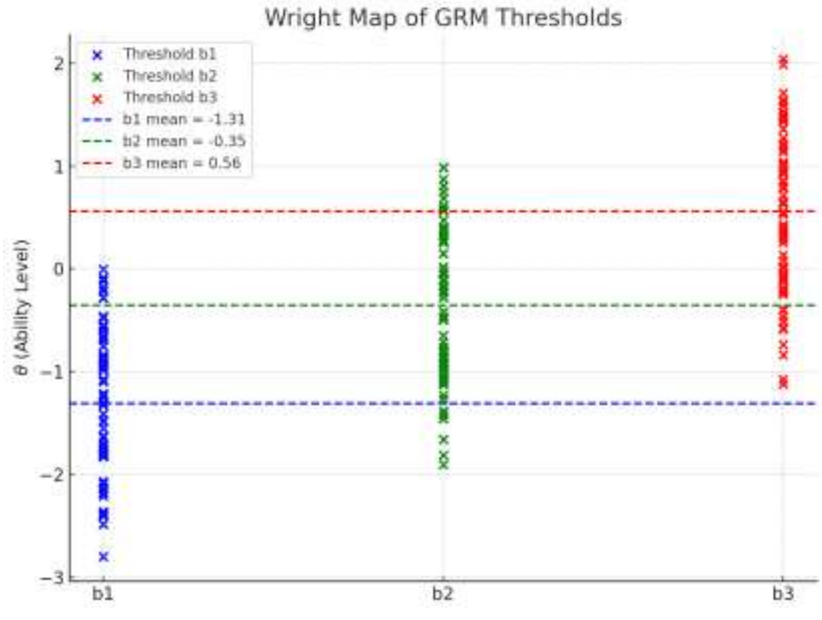

Figure 1. Wright Map of GRM thresholds (β_1, β_2, β_3) for higher-order thinking and innovation competency assessment. The horizontal dashed lines represent the mean threshold values used for cut score determination.

Table 4 Competency Level Cut Scores and Behavioral Characteristics for Higher-Order Thinking and Innovation Based on Wright Map Analysis Using the Graded Response Model (GRM)

Ability Level (θ)	**Level**	**Competency Level**	**Behavioral Characteristics of Students**
$\theta \leq -1.31$	1	Insufficient competency	Unable to distinguish problem components; fails to consider or apply reasoning in decision-making

–1.31 < θ ≤ –0.35	2	Emerging competency	Can analyze and identify problem-solving directions but may overlook causes and consequences
–0.35 < θ ≤ 0.56	3	Competent	Demonstrates critical thinking, considers new ideas, and makes decisions using evidence
θ > 0.56	4	High competency	Generates innovative ideas, considers multiple perspectives, and evaluates potential outcomes

DISCUSSION

The results of this study reflect a rigorous, multi-stage validation process that ensures the competency assessment tool for higher-order thinking and innovation is psychometrically sound and practically applicable.

First, the Confirmatory Factor Analysis (CFA) yielded a good model fit, with all indices (χ^2/df, CFI, TLI, RMSEA, SRMR) within the recommended thresholds. This provides strong evidence that the theoretical competency framework is consistent with the empirical data, confirming that the items accurately measure the intended constructs.

Second, the Item Response Theory (IRT) analysis using the Graded Response Model (GRM) offered precise discrimination (a) and threshold (b1, b2, b3) parameters for all 78 items. High discrimination values indicated that the items effectively differentiate between students of varying proficiency levels, while well-ordered thresholds reflected the increasing ability required for higher score categories.

Third, the reliability analysis showed high internal consistency, with Cronbach's alpha values ranging from 0.899 to 0.943 for subcomponents and 0.978 overall. This demonstrates the stability and dependability of the measurement across the entire item bank.

Finally, the cut score determination utilized Wright Map visualization of GRM thresholds to classify students into four distinct competency levels. The Wright Map visualization revealed clear thresholds (b1 = –1.31, b2 = –0.35, b3 = 0.56) that align with the progressive complexity of cognitive and innovative tasks. These thresholds correspond to behavioral characteristics ranging from insufficient competency—where students struggle to distinguish problem components—to high competency—where students generate innovative ideas and evaluate multiple perspectives effectively.

The behavioral descriptors for each level provide a concrete reference for teachers and administrators, ensuring that score interpretation leads to actionable decisions.

Importantly, the combination of CFA validity evidence, IRT precision, high reliability, and well-defined cut scores creates an item bank that is ideally suited for integration into a Computerized Adaptive Testing (CAT) system. Such a bank supports optimal item selection algorithms, maintains measurement precision across ability ranges, and aligns adaptive testing outputs with meaningful competency levels for both instructional and policy applications.

CONCLUSION AND RECOMMENDATIONS

Conclusion

This study successfully developed and validated a robust item bank for assessing higher-order thinking and innovation in upper secondary students. The sequential process—structural validation via CFA, psychometric calibration via IRT/GRM, confirmation of internal consistency, and establishment of cut scores—resulted in a measurement tool that is both statistically rigorous and educationally meaningful.

By meeting all psychometric criteria and mapping scores to clear competency descriptors, the item bank is fully prepared for deployment in a CAT environment. This ensures adaptive delivery, precise ability estimation, and efficient testing, thereby enhancing the practicality and scalability of competency-based assessment in schools.

Recommendations

1) Integration into CAT Systems – Implement the validated item bank within adaptive testing platforms to provide personalized and efficient assessments.
2) Teacher and Administrator Training – Equip educators with the skills to interpret CAT-based scores and link them to targeted interventions.
3) Ongoing Calibration – Periodically update item parameters and cut scores to maintain accuracy and relevance.

4) Cross-Context Validation – Test the item bank in diverse educational settings to confirm generalizability.
5) Policy-Level Adoption – Incorporate CAT-based competency assessment into national and regional educational policy frameworks.

REFERENCES

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
2. Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8
3. Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698
4. CBE Thailand. (2022). *Competency-based education frameworks in Thailand*. Ministry of Education, Thailand.
5. Duong, B. H. (2022). *Complexities in teaching competencies*. RISE Programme.
6. 12/Complexities_Teaching_Competencies_Vietnam.pdf
7. Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists* (2nd ed.). Psychology Press. https://doi.org/10.4324/9780203807630
8. European Commission. (2018). *Key competences for lifelong learning*. Publications Office of the European Union. https://op.europa.eu/en/publication-detail/- /publication/297a33c8-a1f3-11e8-99ee-01aa75ed71a1
9. Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Prentice Hall.
10. Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., & Sarsted, M. (2014). *A primer on partial least squares structural equation modeling (PLS-SEM).* SAGE Publications.
11. Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. https://doi.org/10.1007/978-94-017-1988-9
12. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.
13. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118
14. International Test Commission. (2014). ITC guidelines for translating and adapting tests (2nd ed.). *International Journal of Testing, 14*(2), 164–232. https://doi.org/10.1080/15305058.2013.799736
15. Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
16. Linacre, J. M. (2021). *A user's guide to WINSTEPS: Rasch-model computer programs*. https://www.winsteps.com/
17. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*.Routledge. https://doi.org/10.4324/9780203056615
18. Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
19. OECD. (2018). *The future of education and skills: Education 2030*. OECD Publishing. https://doi.org/10.1787/9789264300002-en
20. OECD. (2019). *OECD learning compass 2030: A series of concept notes*.OECDPublishing. https://www.oecd.org/education/2030-project/
21. Plailek, T., Essien, A. M., & Sawangdee, Y. (2022). Enhancement of undergraduate students' competency in creating English learning innovation through hybrid learningwith peer coaching. *Journal of Educational Issues, 8*(1), 250–262. https://doi.org/10.5296/jei.v8i1.19600
22. Rusmini, R. (2025). Connected Creative Problem Solving: An alternative model. *TEM Journal, 14*(1), 644–653. https://doi.org/10.18421/TEM141-78
23. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100–114. https://doi.org/10.1007/BF03372160
24. Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence, 34*(1), 120–151. https://doi.org/10.1177/0272431613511332
25. UNESCO. (2015). *UNESCO competency framework*. UNESCO. https://www.unesco.org/en/competency-framework
26. Upapong, T., & Poonputta, A. (2025). Developing a systems thinking skills assessment for upper primary students in Thailand. *International Journal of Evaluation and Research in*
27. *Education, 14*(2), xx–xx. https://files.eric.ed.gov/fulltext/EJ1466587.pdf

28. van der Linden, W. J. (2018). *Handbook of item response theory: Volume one: Models*. CRC Press. https://doi.org/10.1201/9781315374512
29. World Economic Forum. (2020). *The future of jobs report*. World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2020
30. World Economic Forum. (2023). *Future of jobs report 2023*. World Economic Forum. https://www.weforum.org/publications/future-of-jobs-report-2023/
31. World Economic Forum. (2025). *Future of jobs report 2025*. World Economic Forum. https://www.weforum.org/reports/future-of-jobs-report-2025/