大学入試研究ジャーナル第 36 号，52–59，2026

Reliability and Validity of Multiple-Choice Questions Measuring Higher-Order Thinking Skills in University English Entrance Examinations across Two Years

Hiroko Ueda (Kobe University)

This study investigates the validity and reliability of multiple-choice questions (MCQs) designed to assess analytical higher-order thinking skills (HOTS) in university English entrance exams. Using actual exam data from 2024 (N=121) and 2025 (N=141), the study evaluates whether a limited number of MCQs can effectively maintain sufficient validity (difficulty and discrimination indices) and provide meaningful evidence of reliability (KR-20 coefficient). Results revealed that MCQs from both years had appropriate difficulty and high discrimination, indicating good overall quality. However, the small item pool contributed to lower reliability, especially in 2025, due to frequent perfect scores. Increased lexical complexity and sentence length in 2025 did not correlate with higher difficulty or discrimination. These findings highlight the necessity of clear, well-structured items rather than superficial textual complexity, providing practical guidance for designing effective HOTS MCQ assessments under realistic constraints.

Keywords: Higher-Order Thinking Skills (HOTS), Multiple-Choice Questions (MCQs), Validity, Reliability, University Entrance Examinations

1 Introduction

Higher-order thinking skills (HOTS), including analyzing, evaluating, and creating in the revised Bloom's taxonomy (Anderson and Krathwohl, 2001; Ueda, 2021a, 2021b, 2025), are widely regarded as cross-disciplinary competencies (Yen and Halili, 2015). Although domain-general, HOTS must be assessed through concrete task formats. Contemporary reading theory, informed by models such as Kintsch's (1988) construction-integration framework, conceptualizes comprehension as an active process involving linguistic processing, knowledge activation, and higher-order reasoning, meaning that insufficient HOTS impede information integration (Laila and Fitriyah, 2022). Because English reading tasks require inference, integration, and evaluative judgment (Aziz and Rawian, 2022), they constitute a cognitively principled domain for eliciting HOTS.

English reading is well-suited to HOTS assessment because, although it requires basic vocabulary knowledge, its core cognitive operations rely primarily on domain-general reasoning processes. Furthermore, within Japan's entrance examination system, English is taken by most applicants and already includes HOTS-oriented items across national and public universities, offering a broad and representative platform for examining HOTS in high-stakes contexts (Ueda, 2025). Thus, this study treats English as a theoretically and practically justified domain for analyzing how HOTS are instantiated through operational test formats.

Policy frameworks reinforce this alignment: national curricula emphasize HOTS (Aryani and Wahyuni, 2020) and embed them into English examinations (Singh and Shaari, 2019). University entrance exams consequently act as powerful curricular signals that shape expectations for higher-order cognition. Investigating how these exams operationalize HOTS, therefore, contributes to broader discussions on how domain-general constructs are represented through domain-specific tasks.

Despite this emphasis, large-scale examinations face constraints in developing and scoring HOTS-oriented items, making the psychometric behavior of each item especially consequential. This is particularly salient in EFL reading contexts, where lower-order linguistic demands may interact with higher-order reasoning. While constructed-response formats are traditionally favored for assessing HOTS, they pose challenges such as evaluator

variability (Starch and Elliott, 1912; Zhao and Huang, 2020). Multiple-choice questions (MCQs), by contrast, offer objectivity and operational efficiency (Smith, 1982). Yet little empirical research has examined how HOTS-oriented MCQs function under authentic entrance examination conditions.

To address this gap, the present study analyzes the validity and reliability of HOTS-oriented MCQs using data from university English entrance examinations administered across two consecutive years. Specifically, it investigates how item characteristics and textual features influence item difficulty, discrimination (Kelley, 1939), and internal consistency (KR-20; Kuder and Richardson, 1937), as well as how lower-order comprehension demands interact with higher-order reasoning. By identifying conditions under which a small number of well-constructed MCQs can serve as psychometrically robust indicators of HOTS, the study contributes empirical evidence to ongoing discussions of higher-order cognition in language assessment.

Research question:

How do item characteristics influence the validity (difficulty and discrimination) and reliability (KR-20) of MCQs designed to assess HOTS in university English entrance examinations when only a limited number of such items can be included?

2 Method

2.1 Research Design and Rationale

This study employs an empirical, quantitative design to analyze MCQs developed to assess analytical skills, a core component of HOTS, using actual university entrance examination data. Standard item analysis indices, difficulty, discrimination (Kelley, 1939), and KR-20 reliability (Kuder and Richardson, 1937), were applied to evaluate the validity and reliability of these items based on authentic testing contexts rather than theoretical assumptions.

MCQs were selected for their ability to minimize evaluator subjectivity and ensure actionable psychometric outputs, addressing long-standing concerns about inter-rater variability in constructed-response formats (Starch and Elliott, 1912; Zhao and Huang, 2020). Unlike constructed-response items, MCQs facilitate objective evaluation through statistical indicators, reducing the influence of expressive skills and enabling a more direct measurement of HOTS. Given the limited number of HOTS items typically included in high-stakes exams due to time constraints, this study investigates strategies to maintain validity and reliability under such conditions, thereby contributing practical insights to educational assessment in university entrance contexts.

2.2 Data for Analysis

The analyzed MCQs were drawn from the science-track English sections of Kobe University's Kokorozashi Special Selection Exams in 2024 (N = 121) and 2025 (N = 141). Three 2024 items and four 2025 items were selected because they required HOTS.

The 2024 items asked examinees to classify three experimental actions into six steps of the engineering design process: "The following (A) through (C) are parts of the instructions for an experiment. Which of the six engineering design process steps does each experiment instruction fall into? Choose the most appropriate step and write the number on your answer sheet (Kobe University, 2024, p. 5)." The functional structure is summarized below (Tables 1 and 2). Owing to copyright restrictions, items are presented in summarized form.

Table 1 *Summary of Items in 2024 (A-C)*

Item	Summary
A	Converts an idea into a design plan, including leak-prevention features.
B	Generates multiple ideas, compares them, and selects the suitable one.
C	Builds a model and observes and records the results.

Table 2 *Summary of Steps in 2024 (Options)*

Step (Option)	
1. Define the problem	4. Build and test a model
2. Brainstorm solutions	5. Assess the results
3. Develop a solution	6. Share a report

Based on these summaries, the 2024 item aligns with the Analyze level of Bloom's revised taxonomy (Anderson and Krathwohl, 2001), as examinees must differentiate among similar abstract steps and attribute

each action (A-C) to its underlying functional purpose.

For comparison, the 2025 items included a structurally similar analytical classification item. The item required examinees to classify four examples (A-D) into three theoretical categories, reciprocity theory, game theory, and the theory of planned behavior, based on a reading passage: "Reciprocity theory, game theory, and the theory of planned behavior are useful in various situations in daily life and social policies. Each of the following examples A to D shows specific applications of each theory. Based on the passage, classify each example by writing 'R' for reciprocity theory, 'G' for game theory, or 'P' for the theory of planned behavior on the answer sheet (Kobe University, 2025, p. 8)." This task similarly required examinees to distinguish among conceptually related theoretical constructs and assign each example to the appropriate principle.

Qualitative passage features, including Flesch Reading Ease scores (Flesch, 1948), passage length, and number of answer options, are summarized in Table 3. The 2025 passage demonstrated greater difficulty (Flesch score: 43.6 vs. 64.2), increased length (807 vs. 449 words), and fewer answer choices (three vs. six), while average option texts became substantially longer (291 vs. 55 words).

Table 3 *Comparison of Qualitative Factors of English Passages and Multiple-Choice Options between the 2024 and 2025 Passages*

Year	Flesch-Kincaid readability score	Passage length	Number of options	Average length of options (words)
2024	64.2 (Standard)	449	6	55
2025	43.6 (Difficult)	807	3	291

Lexical profiles were analyzed using the EnglishProfile (Cambridge University Press and Assessment, n.d.) to classify word tokens by CEFR level (A1-C2). As shown in Table 4, the 2025 passage contained higher proportions of B1-B2 vocabulary (35.7% vs. 28.0%) and C1-C2 vocabulary (3.6% vs. 1.8%), indicating increased lexical difficulty relative to 2024.

Table 4 *Comparison of CEFR Vocabulary Distribution in 2024 and 2025 Passages*

Year	A1 (Beginner), A2 (Elementary)	B1 (Intermediate), B2 (Upper-Intermediate)	C1 (Advanced), C2 (Proficient)
2024 (%)	70.2	28	1.8
2025 (%)	60.7	35.7	3.6

2.3 Data Analysis Method

Item analyses were conducted using Python (version 3.13.3) with pandas and numpy libraries. Difficulty was calculated as the proportion of correct responses per item, discrimination was derived using upper-lower group comparisons, and KR-20 was computed from item difficulty and total score variance to evaluate internal consistency reliability.

3 Results

This section reports the average difficulty index, discrimination index, and KR-20 reliability coefficients obtained from the item analysis of MCQs designed to measure the analytical skill of HOTS for 2024 (three items) and 2025 (four items), as summarized in Table 5. Ideal criteria for each index are provided for interpreting item quality, and due to institutional confidentiality restrictions, only aggregated mean values are presented.

Table 5 *Average Item Analysis Results and Ideal Criteria for MCQs Measuring the Analytical Skill of HOTS in 2024 and 2025*

Year/ Criteria	Mean Difficulty Index	Mean Discrimination Index	KR-20 Reliability
2024	0.59	0.83	0.68
2025	0.71	0.67	0.52
Ideal	0.30-0.70	≧0.30	≧0.70

Note:
1) Difficulty index values within 0.30-0.70 represent ideal item difficulty (Rafi et al., 2023).
2) Discrimination index values equal to or above 0.30 are considered effective (Rafi et al., 2023).
3) KR-20 coefficients equal to or greater than 0.70 indicate good stability (Batista et al., 2023).

3.1 Results for 2024

In 2024, the four items analyzed had average correct response rates of 59%, indicating that none of these questions were excessively easy or difficult. Each item fell within an appropriate difficulty range,

reflecting a balanced distribution across the test. Furthermore, all items demonstrated very high discrimination indices (greater than approximately 0.6), substantially exceeding the generally accepted threshold of 0.3. This indicates that the items effectively distinguished between higher- and lower-performing examinees, suggesting high item quality. The overall KR-20 reliability coefficient for this year was 0.68, indicating moderate internal consistency. While reliability coefficients of 0.70 or higher are typically considered desirable, the small number of items (three) is likely responsible for this slightly reduced value. Nevertheless, considering the limited number of items, the observed reliability is acceptable.

3.2 Results for 2025

In 2025, the four analyzed items had an average correct response rate of 71%, showing a slight tendency toward easier items. Although items with difficulty indices above .90 are considered excessively easy (Quaigrain and Arhin, 2017), all four items fell below this threshold. However, because the test comprised only four relatively easy items, the score distribution became compressed, and about 36% of the examinees (n = 51) achieved perfect scores, producing a ceiling effect.

All items displayed strong discrimination (above 0.5), yet the overall KR-20 reliability for 2025 remained low at 0.52, largely due to the limited item count and the ceiling effect.

4 Discussion

4.1 Summary of Main Findings

This study examined the quality of analytical HOTS MCQ items in 2024 and 2025 using KR-20 reliability, difficulty indices, and discrimination indices. In both years, item difficulty remained within acceptable ranges and discrimination was high, indicating generally sound item quality. However, the small number of items lowered the KR-20 reliability.

Comparisons across years showed that although the 2025 items featured greater vocabulary size, sentence length, and lexical complexity than those of 2024, correct response rates increased and discrimination decreased. This indicates that superficial linguistic complexity does not necessarily correspond to higher item difficulty or stronger discrimination.

4.2 Interpretation of Results in Light of Theory and Prior Research

4.2.1 Difficulty of HOTS Items and the Interaction Between Lower- and Higher-Order Processes

Prior studies suggest that HOTS items tend to fall within the moderate to difficult difficulty range (Maryani et al., 2021). Yet the items in this study, particularly in 2025, were not highly difficult. Despite increased lexical and syntactic complexity, difficulty decreased, partly associated with the reduction of answer choices from six to three, and more critically, to clearer option wording that eased basic comprehension. This aligns with evidence that higher-order thinking builds upon successful lower-order cognitive processing (Mitana et al., 2018).

According to the construction-integration model (Kintsch, 1988), comprehension difficulty arises not primarily from surface complexity but from the demands involved in integrating multiple textual propositions. When MCQ options became less ambiguous in 2025, lower-order burden decreased, allowing examinees to engage more directly in integration, thereby lowering effective cognitive load even for HOTS-targeted items.

This supports Laila's (2022) view that insufficient activation of higher-order processes can hinder integration; from a cognitive-processing perspective, when integration is relatively straightforward, the cognitive demand for HOTS tends to diminish. The findings suggest that cognitive challenge in HOTS assessment must stem from reasoning pathways required for option discrimination, rather than from text length or vocabulary difficulty.

4.2.2 Discrimination Decline and the Cognitive Demands of Distractors

HOTS items typically show high discrimination (Hamamoto Filho et al., 2020; Hermi and Achour, 2015), a pattern largely confirmed in this study. However, discrimination weakened in 2025, likely

due to clearer option wording that resulted in more perfect scores and a ceiling effect.

Although reducing the number of answer choices increases the chance level of correct responses, prior research indicates that examinees are unlikely to engage in blind guessing and instead tend to rely on educated guessing (Rodriguez, 2005). Accordingly, the decline in discrimination observed in 2025 is more plausibly attributable to ceiling effects associated with clearer option wording.

The contrasting patterns between 2024 and 2025 illustrate how HOTS function within reading comprehension tasks, requiring inferential, analytical, and evaluative judgments (Aziz and Rawian, 2022). When distractors no longer demand integration of textual propositions or evaluation of competing interpretations, both high- and low-ability examinees can answer correctly, reducing discrimination.

From the perspective of HOTS as cross-disciplinary competencies, this underscores a key principle: HOTS emerge only when items require learners to navigate ambiguity and synthesize information across cues. If options test only lexical or surface comprehension, they no longer operationalize the domain-general reasoning processes underlying HOTS. Thus, discrimination reflects the extent to which items require examinees to:

1) eliminate distractors through inferential reasoning,
2) integrate multiple propositions,
3) evaluate logical coherence.

When option clarity removes these reasoning demands, discrimination naturally declines.

4.2.3 Reliability Constraints and the Nature of HOTS as Domain-General Constructs

The small number of HOTS items in both years produced low KR-20 values, but this should not be interpreted as evidence of weak measurement. HOTS are cross-disciplinary (Yen and Halili, 2015) and operate across multiple cognitive layers. Because they are not designed to function as an isolated subscale, internal-consistency indices, which assume unidimensionality and sufficient item volume, are not appropriate indicators of measurement quality in this context.

Instead, reliability should be conceptualized from a construct-representation perspective, where the focus is on whether items authentically elicit domain-general reasoning processes within a domain-specific reading format. Given that the full reading test was reliable and the HOTS items displayed strong discrimination and conceptually aligned reasoning demands, the low subset KR-20 values merely reflect the psychometric constraints of small item counts and do not undermine construct validity.

This interpretation aligns with contemporary measurement theory, which holds that HOTS are best assessed through rich task formats embedded within disciplinary contexts, rather than isolated psychometric scales.

4.3 Contributions and Implications

4.3.1 Integrating Reading Theory into HOTS Item Design

By situating the results within Kintsch's construction-integration framework, this study challenges long-standing assumptions about text complexity and HOTS difficulty. Increases in surface-level complexity do not guarantee increased cognitive demand. Instead, difficulty emerges when tasks require examinees to activate, integrate, and evaluate textual information, central operations in higher-order thinking.

Thus, designers of HOTS-oriented MCQs must focus on:

1) integration difficulty rather than vocabulary rarity,
2) conceptual ambiguity rather than sentence length,
3) evaluative reasoning rather than superficial textual load. This supports viewing English reading tasks as a cognitively principled domain (as established in the introduction) for HOTS assessment.

4.3.2 Designing MCQs That Truly Elicit Domain-General HOTS

Because HOTS require inferential, integrative, and evaluative reasoning, distractor design must be grounded in cognitive plausibility, not stylistic complexity. The study demonstrates that HOTS difficulty can be meaningfully calibrated by manipulating:

1) the ambiguity and reasoning depth embedded in distractors,
2) the number of response options,
3) the cognitive distance between correct and incorrect alternatives. These findings extend prior research by showing not merely that MCQs can assess HOTS, but how and under what design conditions they effectively do so within high-stakes entrance exam contexts.

4.3.3 Reinterpreting Reliability in Small HOTS Subsets: A Construct-Oriented Perspective

A key methodological consideration in this study is that the HOTS items were not designed to function as an independent subscale but were intentionally embedded within the broader reading comprehension test. Because internal-consistency coefficients such as KR-20 assume a unidimensional scale with a sufficient number of items, they are not appropriate indicators of measurement quality for such a small, non-subscale item set. As extensively noted in psychometric research, KR-20 and Cronbach's α systematically underestimate reliability when applied to very small item sets (Cortina, 1993; Tavakol and Dennick, 2011). Therefore, the low KR-20 values observed for the HOTS subset should be interpreted as a methodological artifact rather than evidence of weak measurement.

Within the context of HOTS assessment, what is more central is whether the items successfully elicit the domain-general reasoning processes that define higher-order thinking, even when embedded within a domain-specific reading task. In both years, the HOTS items demonstrated strong discrimination and clear cognitive authenticity, indicating that they engaged the intended reasoning processes despite their limited number. These item-level indicators provide more meaningful evidence of construct representation than internal-consistency estimates derived from an insufficient subset.

This construct-oriented interpretation aligns with contemporary perspectives in language testing, which emphasize that HOTS are best assessed through context-rich tasks rather than isolated psychometric subscales. It is also consistent with contemporary validity theory, which defines validity as an integrated evaluative judgment about the adequacy of score interpretations and uses, grounded in both empirical evidence and theoretical rationale (Messick, 1989; Kane, 2013). From this perspective, validity depends on whether the items faithfully represent the construct of higher-order thinking and elicit the intended reasoning processes, not on numerical indices derived from a small subset.

From this standpoint, a small number of strategically designed MCQs can meaningfully contribute to the overall assessment by enhancing the reasoning profile of the test and indirectly aligning the assessment with national curriculum expectations for higher-order competencies. Consequently, the validity of HOTS measurement in this study rests not on subscale-level reliability coefficients but on the conceptual quality, cognitive alignment, and psychometric functioning of the items. This shift, from structural reliability to construct-representation validity, reflects current debates on the assessment of higher-order thinking and underscores the value of embedding HOTS within authentic reading contexts, even when only a limited number of items can be included.

5 Conclusion

This study examined the performance of analytical HOTS-oriented MCQs in two consecutive entrance examinations. Despite appropriate difficulty and strong discrimination, the small number of HOTS items naturally constrained KR-20 estimates, particularly in the presence of ceiling effects. Importantly, increases in textual complexity did not raise item difficulty; clearer option wording and fewer distractors reduced lower-order processing demands and facilitated performance. These findings show that HOTS difficulty depends less on surface linguistic features than on how options engage inferential and integrative reasoning. They also highlight that item-level functioning, rather than subscale-level internal consistency, provides the most meaningful evidence of construct representation when HOTS items constitute a small subset of the test. These findings suggest that even a small number of well-designed HOTS items can enhance the cognitive profile of high-stakes reading

assessments. Future research should extend the validity argument by integrating response process evidence on how examinees engage with HOTS-oriented MCQs.

References

Anderson, L. W. and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Longman.

Aryani, E. J. and Wahyuni, S. (2020). An Analysis of Higher Order Thinking Skills Realization in Reading Comprehension Questions. *Language Circle: Journal of Language and Literature*, **15**(1), 83–89.

Aziz, M. and Rawian, R. (2022). Modeling higher order thinking skills and metacognitive awareness in English reading comprehension among university learners. *Frontiers in Education* (Lausanne), **7**. https://doi.org/10.3389/feduc.2022.991015

Batista, S. A., Ginani, V. C., Stedefeldt, E., Nakano, E. Y., and Botelho, R. B. A. (2023). Reproducibility and Validity of a Self-Administered Food Safety Assessment Tool on Children and Adolescent's Risk Perception, Knowledge, and Practices. *Nutrients*, **15**(1), 213. https://doi.org/10.3390/nu15010213

Cambridge University Press and Assessment. (n.d.). *EnglishProfile*. https://englishprofile.org/?menu=evp-online (2025, January 8).

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, **78**(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233. https://doi.org/10.1037/h0057532

Hamamoto Filho, P. T., Silva, E., Ribeiro, Z. M. T., Hafner, M. de L. M. B., Cecilio-Fernandes, D., and Bicudo, A. M. (2020). Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. *São Paulo Medical Journal*, **138**(1), 33–39. https://doi.org/10.1590/1516-3180.2019.0459.r1.19112019

Hermi, A. and Achour, W. (2015). Difficulty, discrimination and cognitive level of Microbiology exam questions of the Faculty of Medicine of Tunis. *La Tunisie Médicale*, **93**(8–9), 487–490. PMID: 26815509.

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

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, **30**(1), 17–24. https://doi.org/10.1037/h0057123

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, **95**(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kobe University. (2024). *Rei6 Kokorozashiri Sougoumondai I[2024 Kokorozashi Science Comprehensive Items I]*. Center for the Next Generation. https://drive.google.com/file/d/1h2ERGNJja0_aJVR_RfTK-W6PoDmXf77e/view (2025, March 18).

Kobe University. (2025). *Rei7 Kokorozashiri Sougoumondai I[2025 Kokorozashi Science Comprehensive Items I]*. Center for the Next Generation. https://drive.google.com/file/d/1TBu_LPY2mSmPBr9ZWdHUH0Ve7JSaTDN2/view (2025, March 18).

Kuder, G. F. and Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, **2**(3), 151–160. https://doi.org/10.1007/BF02288391

Laila, I. and Fitriyah, I. (2022). An Analysis of Reading Comprehension Questions in English Textbook Based on Revised Bloom's Taxonomy. *Journal of English Teaching*, **8**(1), 71–83.

Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., and Fitrianawati, M. (2021). HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers. *Journal of Turkish Science Education*, **18**(4), 674–690.

Messick, S. (1989). Validity, in R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–104), American Council on Education and Macmillan.

Mitana, J. M. V., Muwagga, A. M., and Ssempala, C. (2018). Assessment of higher order thinking skills: A case of Uganda Primary Leaving Examinations. *African Educational Research Journal*, **6**(4), 240–249. https://doi.org/10.30918/AERJ.64.18.083

Quaigrain, K. and Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, **4**(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., and Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination. *Pedagogical Research*, **8**(1), em0145. https://doi.org/10.29333/pr/12657

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement, Issues and Practice*, **24**(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Singh, R. K. V. and Shaari, A. H. (2019). The analysis of Higher-Order Thinking skills in English reading comprehension tests in Malaysia. *Geografia: Malaysian Journal of Society and Space*, **15**(1), 12–26.

Smith, J. K. (1982). Coverging on Correct Answers: A Peculiarity of Multiple Choice Items. *Journal of Educational Measurement*, **19**(3), 211–220. https://doi.org/10.1111/j.1745-3984.1982.tb00129.x

Starch, D. and Elliott, E. C. (1912). Reliability of the Grading of High-School Work in English. *School Review*, **20**(7), 442–457. https://doi.org/10.1086/435971

Tavakol, M. and Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, **2**, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Ueda, H. (2021a). *Analysis of National English Test Items in Japan for Higher-Order Thinking Skills Using the Revised Bloom's Taxonomy*. [Unpublished master's thesis]. University College London Institute of Education.

Ueda, H. (2021b). *Analysis of National English Test Items in Japan for Higher-Order Thinking Skills Using the Revised Bloom's Taxonomy*. [Paper presentation]. JACET 60th Commemorative International Convention, Online, Japan.

Ueda, H. (2025). Elevating Educational Evaluation: A Case Study on the Implementation of Higher-Order Thinking Skill Assessment in English Entrance Examinations. *The Journal of University Admissions Research*, **35**, 63–70. https://www.sakura.dnc.ac.jp/archivesite/wp-content/uploads/2025/03/Journal2025_35-09.pdf

Yen, T. S. and Halili, S. H. (2015). Effective teaching of higher-order thinking (HOT) in education. *The Online Journal of Distance Education and e-Learning*, **3**(2), 41–47.

Zhao, C. and Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, **67**, Article 100911. https://doi.org/10.1016/j.stueduc.2020.100911