Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

INSTd: A Computer-Adaptive Assessment of Definition-Based Comprehension Skills

Noriko ARAI[a*], Naoya TODO[b], Teiko ARAI[c], Shingo SUGAWARA[d], Takuya MATSUZAKI[e] & Koken OZAKI[f]
[a]*National Institute of Informatics, Japan*
[b]*Tokyo Metropolitan University, Japan*
[c]*Takasaki City University of Economics, Japan*
[d]*Institute of Science for Education, Japan*
[e]*Tokyo University of Science, Japan*
[f]*University of Tsukuba, Japan*
*arai@nii.ac.jp

Abstract: This paper introduces INSTd, a computer-adaptive test designed to assess learners' ability to interpret dictionary-style definitions and apply them in appropriate contexts. Unlike conventional vocabulary tests that measure already-acquired word knowledge, this test presents definitions and evaluates learners' ability to acquire new words based on those definitions. We analyzed the correlation of the estimated abilities of INSTd with their vocabulary sizes among 1,221 sixth-grade students. Results show a moderate correlation with vocabulary size ($r = 0.343$, $p < .001$), indicating that the test, although related to vocabulary, captures a distinct aspect of reading ability.

Keywords: definition comprehension, vocabulary assessment, digital learning environment, item response theory

1. Introduction

In digital learning environments, such as online textbooks and platforms like Wikipedia, it is common practice to hyperlink unfamiliar terms to their dictionary definitions. It reflects a widely accepted assumption: that learners can interpret these definitions and apply them as needed to acquire new vocabulary. However, is this assumption justified?

Most prior research on vocabulary has focused on size or depth—what learners already know (Anderson, 1981; Begler, 2010; Sternberg, Powell & Kaye, 1987). Fewer studies have investigated whether learners can successfully interpret unfamiliar definitions when presented in context (Arai et al, 2025). This paper introduces INSTd, a new test designed to measure this often-overlooked ability. In each item, examinees are presented with a short dictionary-style definition and asked to select the appropriate examples or applications from multiple-choice options.

Empirical validation suggests that INSTd is psychometrically robust. Results from a study involving 1,221 sixth-grade students showed a moderate positive correlation between estimated ability scores on INSTd and vocabulary size ($r = 0.343$, $p < .001$). This suggests that while INSTd is related to vocabulary knowledge, it measures a different dimension of reading ability.

2. Test Design

INSTd is a computer-adaptive test designed to measure examinees' ability to read short, dictionary-style definitions and apply them accurately across different contexts. Each item presents a concise definition, a simple instruction, and four multiple-choice options, with one or more correct answers. The test interface is deliberately minimal, featuring one item per screen and a clear "Submit" button, which helps examinees concentrate solely on interpreting the definition. Figure 1 illustrates an example of INSTd questions.

Definitions are primarily sourced from widely used Japanese educational materials, including the popular elementary-level dictionary Reikai Shougaku Kokugo Jiten (Sanseido, 2024), as well as Wikipedia and textbooks. Vocabulary terms are selected based on definitional clarity.

Each item is authored by an experienced developer and reviewed by at least two reviewers to verify that the definition is unambiguous, the options are appropriate, and the correct answers logically follow from the definition.

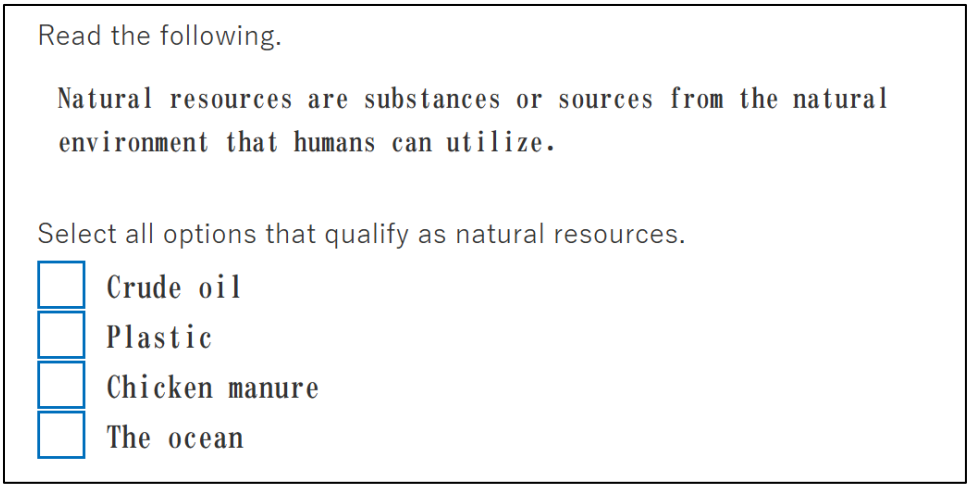

Figure 1. Screenshot of an INSTd question. Correct answer: 'crude oil'.

3. INSTd as an IRT-Based Computer Adaptive Test

We implemented INSTd as a computer-adaptive test based on the 2PL model (van der Linde & Glas, 2010). The probability that an examinee with ability θ_i will answer item j correctly is given by the 2PL model:

$$P_j(\theta_i) = \frac{1}{1 + \exp\left[-k \cdot a_j(\theta_i - b_j)\right]}.$$

Here, k is a constant approximating the normal ogive model ($k = 1.7$ in INSTd), a_j is the item discrimination parameter, and b_j is the item difficulty parameter.

The average factor loading of the items was 0.521, and the IRT-based reliability coefficient was 0.768. These values indicate that the items measure a coherent latent trait and provide sufficiently precise ability estimates for diagnostic purposes.

These results confirm that INSTd meets essential psychometric criteria and can reliably assess definition-based comprehension across a wide range of examinee populations.

During the allotted five-minute testing period, examinees are instructed to answer as many questions as possible, as accurately as possible. The items presented to each examinee are selected based on their prior responses and accuracy, following a computer-adaptive testing algorithm explained in Section 3. As a result, each examinee completes a different number of items and encounters a unique sequence of questions.

4. Results

To investigate the relationship between INSTd and traditional measures of vocabulary knowledge, we analyzed its correlation with a vocabulary size survey conducted among 1,221 sixth-grade students in City A. In this survey, each examinee was presented with a list of 60 words and asked to indicate whether they knew each word. Responses were coded as binary values (1 = known, 0 = unknown or unsure), and the total number of known words was calculated as an estimate of vocabulary size.

The 60 words were carefully selected based on familiarity ratings from a large-scale database provided by NTT Communication Science Laboratories (Amano & Kondo, 1998). Words were chosen to span a wide range of familiarity levels, including highly familiar terms (e.g., “bank”), moderately familiar terms (e.g., “economy”), and low-familiarity terms (e.g., “recoinage”). This design ensured that the vocabulary size measure could differentiate examinees across a broad range of lexical knowledge.

The correlation between INSTd ability estimates and vocabulary size was found to be $r = 0.343$, statistically significant at $p < .001$. While moderate in strength, this correlation suggests that INSTd measures a construct that is related to—but distinct from—vocabulary breadth.

This result suggests that while hyperlink-based introduction of unfamiliar terms in digital learning platforms may support learners with extensive vocabularies, it may not adequately serve those with more limited lexical knowledge and underdeveloped definition-based comprehension skills as measured by INSTd.

5. Conclusion

This study introduced INSTd, a computer-adaptive test designed to assess learners’ ability to interpret and apply dictionary-style definitions—an ability that underpins many forms of learning in digital environments but has rarely been assessed directly.

Through rigorous psychometric development and large-scale validation, INSTd has been shown to capture a distinct dimension of lexical competence, one that complements vocabulary size. Its scalable design and diagnostic precision make it a promising tool for both educational assessment and personalizing digital reading experiences.

Acknowledgments

This research was supported by the Japan Society for the Promotion of Science (JSPS) under Grants-in-Aid for Scientific Research JP16H01819 and JP21H04416 and by Reading Skill Test, Inc. We also thank the Institute of Science for Education for providing access to INSTd data and screen captures used in this study.

References

Amano, S., & Kondo, T. (1998). Estimation of mental lexicon size with word familiarity database. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (Vol. 5, pp. 2119–2122).

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association.

Arai, T., Todo, N., Ozaki, K., & Arai, N. H. (2025). *Assessing the ability to read and interpret mathematical definitions: Insights from INSTm*. In Proceedings of the 25th IEEE International Conference on Advanced Learning Technologies (ICALT 2025). IEEE.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194.

Sanseido. (2024). *Reikai Gakushū Kokugo Jiten* (11th ed.) [例解学習国語辞典・第11版, in Japanese]. Sanseido.

Sternberg, R. J., Powell, J. S., & Kaye, D. B. (1987). Measuring the depth of vocabulary knowledge: Toward a bridge between vocabulary learning and reading comprehension. *Journal of Educational Psychology, 79*(4), 269–275.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-6