Development of The Mandarin Language Proficiency Assessment Instrument Hanyu Shuiping Ceshi (HNC) as a Standard for Mandarin Language Skills for Unesa Students

Muhammad Farhan Masrur[1], Cicik Arista[2], Sueb[3], Khusnul Khotimah[4]

[1,2,3,4] Universitas Negeri Surabaya

ABSTRACT: This research aims to develop the Hanyu Shuiping Ceshi (HNC) as a standardized assessment tool for evaluating Mandarin language skills among UNESA students. Using the Research and Development (R&D) approach, the study was conducted in three evaluation phases: one-to-one, small group, and field testing. Results demonstrated that the HNC instrument achieved high content and construct validity, with average CVR scores of 0.90 for Level 1 and 0.88 for Level 2. Reliability was confirmed through Cronbach's alpha values exceeding 0.80. The results further revealed that HNC Levels 1 and 2 effectively measure basic to early intermediate Mandarin proficiency. The inclusion of contextualized test items enhanced critical thinking and communication skills applicable to real-world scenarios. Recommendations for further development include introducing interactive questions and expanding the instrument to Level 3 to facilitate a progressive and comprehensive learning experience. Overall, the HNC is a reliable tool for assessing academic achievement while fostering practical language competencies relevant to professional demands.

KEYWORDS: Research and Development; Assessment Instrument; Proficiency Test; Mandarin Language; Hanyu Shuiping Ceshi

INTRODUCTION

Mandarin language education in Indonesia has undergone significant transformation alongside the implementation of China's One Belt One Road initiative. The increasingly close ties between Indonesia and China, particularly in relation to their joint commitment in 2020 to strengthen bilateral cooperation, have positively impacted the growing demand for individuals proficient in Mandarin (Djelantik, 2015). However, Mandarin education in Indonesia faces challenges, especially regarding the quality standards of Mandarin teaching, including limited opportunities for learners to assess their language skills (Maria, 2019) Standardized assessment instruments, as one of the main tools in the education system, play a crucial role in achieving national educational goals (Lubis, 2013). Therefore, it is important to develop objective evaluation tools for assessing the quality of Mandarin education to ensure smooth and high-quality teaching and learning processes in the context of the growing bilateral relationship between Indonesia and China.

The current Mandarin proficiency test is the HSK (*Hanyu Shuiping Kaoshi*), an internationally recognized standardized test developed by the Chinese Language and Culture Center (*Hanban*). HSK evaluates the Mandarin proficiency of individuals, particularly non-native speakers, by testing four main skills: listening, reading, writing, and speaking. Comprising six levels, from HSK 1 (basic level) to HSK 6, HSK is used as a learning standard for Mandarin in many countries, including Indonesia (Peng et al., 2021). However, the HSK has drawbacks, including its relatively high cost, as well as limited availability in terms of time and location. These factors pose significant barriers for Mandarin learners, especially in higher education.

One higher education level requiring Mandarin language evaluation tools is the university level. In the context of implementing the Independent Learning-Independent Campus Curriculum (*Merdeka Belajar Kampus Merdeka* or MBKM), universities in Indonesia, as institutions for knowledge, research, and community service, face the demand to focus on achieving performance targets (Anggraeni et al., 2022). A key strategy for managing university performance is through the Main Performance Indicators for State Universities (*Indikator Kinerja Utama Perguruan Tinggi Negeri* or IKU-PTN), outlined in the Minister of Education and Culture's Decree, specifically Decree Number 3/M/2021. These indicators comprise three main aspects: graduate quality, faculty and teaching staff quality, and curriculum quality. Graduate quality is assessed through the employment of graduates in decent jobs and student experiences outside campus (Merdeka, 2020). Therefore, study programs must develop

Development of The Mandarin Language Proficiency Assessment Instrument Hanyu Shuiping Ceshi (HNC) as a Standard for Mandarin Language Skills for Unesa Students

relevant learning outcomes aligned with the needs of industries and job markets.

The Mandarin Language Education Study Program (*Program Studi Pendidikan Bahasa Mandarin* or Prodi PBM) at Indonesian state universities emerged in 2009, following proposals from universities or mandates from the Directorate General of Higher Education. Over time, the Prodi PBM was successively introduced at five state universities: Tanjungpura University, Pontianak (2009), State University of Surabaya (2010), State University of Malang (2011), State University of Semarang (2011), and State University of Jakarta (2014) (Maria, 2019). The journey of Prodi PBM at these five universities has encountered several challenges, including a shortage of educators, limited teaching materials, and the absence of standardized Mandarin proficiency tests. These factors have hindered efforts to enhance students' Mandarin language skills, and these challenges remain unresolved (Maria, 2019). Additionally, Prodi PBM faces an expanded range of responsibilities, such as creating standardized Mandarin proficiency tests based on the Indonesian National Qualifications Framework (KKNI). This effort is essential for defining graduate profiles, achieving learning outcomes, obtaining program accreditation, and strengthening partnerships with various stakeholders, including initiatives from the Ministry of Research, Technology, and Higher Education. A key focus is to equip graduates with measurable Mandarin proficiency qualifications that align with national standards.

To meet this need, the Hanyu Shuiping Ceshi (HNC) has been proposed as a new Mandarin proficiency test. This instrument aims to stimulate the development of tools for assessing Mandarin proficiency while also supporting preparation for the HSK test. The HNC is designed to cater to diverse groups, including university students, school students, and the general public, offering them an opportunity to evaluate their Mandarin skills and obtain certificates recognized in the job market. This initiative underscores a commitment to enhancing Mandarin language proficiency and promoting sustainable development in Mandarin education across various levels.

The development of the HNC will adopt a Research and Development (R&D) approach, encompassing several critical stages. It will begin with a comprehensive literature review and a comparison with international standardized tests like the HSK, followed by an analysis of local needs tailored to the Indonesian context, including the MBKM curriculum and KKNI framework. The test design will involve developing appropriate levels and question formats, followed by pilot testing to assess validity, reliability, and question difficulty. Insights from these pilot tests will guide refinements and improvements to the instrument. Once finalized, the HNC will be broadly implemented in educational institutions and the wider public domain. To ensure ongoing relevance, reliability, and recognition in academic and professional contexts, periodic evaluations of the HNC will be conducted.

METHOD

The research applied the Research and Development (R&D) method, a systematic approach to identifying problems and creating tools to address them (Creswell, 2018). This methodology was chosen to ensure a comprehensive development process that aligns with real-world needs. Respondents were selected through random sampling, ensuring that their homogeneous characteristics eliminated the need for significant grouping. The study involved 50 intermediate-level students from UNESA's Mandarin Study Program, providing a diverse yet focused sample to test the instrument's effectiveness.

The research instrument is closely related to the research method, as it plays a role in providing assessments and evaluations in the learning context (Nurfillaili & Anggereni, 2016). The instrument and data collection technique employed in this study is a questionnaire. The use of questionnaires is deemed appropriate and suitable for collecting data relevant to the research questions.

The validity and reliability of the instrument were tested. Validity refers to the accuracy of a test or scale in measuring a research phenomenon (Azwar, 2016). The validity in this study was measured through content validity using the Content Validity Ratio (CVR) method by Lawshe and construct validity through factor analysis. The CVR for content validity is calculated using the formula described by Azwar :

$$CVR = (\frac{2ne}{ne}) - 1$$

Explanation:

ne = The number of Subject Matter Experts (SME) who rate an item as essential.

n = The total number of SMEs conducting the evaluation.

The CVR value, as explained by (Azwar, 2011), ranges from -1.00 to +1.00. If CVR>0.00\text{CVR} > 0.00CVR>0.00, it indicates that more than 50% of SMEs in the panel consider the item essential. The larger the CVR value above zero, the more essential the item and the higher its content validity.

Construct validity was evaluated through Exploratory Factor Analysis (EFA). The validity model employed the Kaiser-Meyer-Olkin (KMO) measure, which is considered valid if the KMO value is greater than 0.5 and the significance is less than 5%.

Development of The Mandarin Language Proficiency Assessment Instrument Hanyu Shuiping Ceshi (HNC) as a Standard for Mandarin Language Skills for Unesa Students

On the anti-image correlation diagonal, values greater than 0.5 are expected for all items. If any item has a value below 0.5, it is eliminated (Priyatno & Korelasi, 2013). Once all items meet the criteria, the next steps can be taken.

According to Gerber & Finn (2013), there are four basic steps in conducting factor analysis:
Calculating all correlation matrices for each variable, performing factor extraction, conducting rotation, naming each factor. All components must have values ≤ 0.5 for the factor analysis process to be considered valid. A valid HNC instrument will have higher effectiveness because it can accurately measure Mandarin language skills and provide useful feedback for improving students' skills. Effectiveness refers to the extent to which a plan can achieve its goals. The goal of assessing question effectiveness is to evaluate questions in terms of validity, reliability, difficulty level, discrimination power, and distractor functionality. These steps are taken to ensure that the instrument used is valid and reliable for measuring the Mandarin language skills of students in the Mandarin Study Program at UNESA.

RESULT AND DISCUSSION
Result
1.   Evaluation Results of Scores
Table 1: Average Student Scores at Each Evaluation Stage

Evaluation Step	Level	Student	Average	Category
One-to-One Evaluation	HNC Level 1	3	82,3	Very Good
One-to-One Evaluation	HNC Level 2	3	79,0	Good
Small Group Evaluation	HNC Level 1	6	80,2	Very Good
Small Group Evaluation	HNC Level 2	6	77,5	Good
Field Test	HNC Level 1	134	81,5	Very Good
Field Test	HNC Level 2	53	78,7	Good

The test trials were conducted in three stages: one-to-one evaluation, small group evaluation, and field test. Below are the results of each stage:
   a.   In this stage, HNC Level 1 and Level 2 tests were conducted on three students representing high, medium, and low ability levels. The evaluation results show: (1) The average score of students at HNC Level 1: 82.3 (very good); (2) The average score of students at HNC Level 2: 79.0 (good); (3) Students provided feedback on the clarity of instructions and the relevance of the questions to everyday life contexts.
   b.   In small groups (6 students), the tests were conducted simultaneously. The results show: (1) The average score for HNC Level 1: 80.2 (very good category); (2) The average score for HNC Level 2: 77.5 (good category); (3) Some questions in Level 2 were found too challenging for students with lower abilities, prompting revisions to less appropriate answer choices.
   c.   Field Test, the test involved students from the Mandarin Language Study Program at FBS UNESA, with the following distribution:
      1)   HNC Level 1: Class 2023A (38 students), 2023B (37 students), 2023C (36 students), and 2023I (23 students).;
      2)   HNC Level 2: Class 2022A (29 students) and 2022B (24 students).
      3)   he results of this test indicate that the HNC instrument effectively measures Mandarin proficiency at the basic and initial levels. Minor revisions were made to improve the relevance and appropriateness of question difficulty levels, especially for Level 2, to ensure optimal results.

The average scores at the trial stages demonstrate that HNC Level 1 and Level 2 meet the categories of "good" to "very good." This indicates the alignment of the questions with the students' ability levels. Tests that match learners' abilities enhance their confidence and foster motivation to learn (Anderson & Krathwohl, 2001; Li, 2024) Therefore, the design of HNC questions relevant to students' abilities not only assesses academic achievement but also has a positive psychological impact on learners.

Development of The Mandarin Language Proficiency Assessment Instrument Hanyu Shuiping Ceshi (HNC) as a Standard for Mandarin Language Skills for Unesa Students

2. Difficulty Level and Discriminatory Power

Table 2: Statistical Analysis of HNC Level 1 and Level 2 Questions

Parameter	Test Level	Range of Values	Difficulty Level	Discriminatory Power (≥ 0,3)	Reliability (Cronbach α)
Difficulty Level	Level 1	0,35–0,70	Moderate	90% Questions Meet Criteria	0,85
	Level 2	0,30–0,65	Moderate	85% Questions Meet Criteria	0,82
Discriminatory Power	Level 1	≥ 0,30	Good		
	Level 2	≥ 0,30	Good		

a. The difficulty level of the questions for both levels is categorized as moderate (0.30–0.70), indicating a balanced distribution between easy and difficult questions. This aligns with the guidelines for question construction, as moderate difficulty is considered ideal for effectively measuring competencies (Arikunto, 2006).

b. Questions with extreme difficulty levels (< 0.30 or > 0.70) were revised to enhance the effectiveness of the instrument. This is supported by Item Response Theory (IRT), which emphasizes the importance of aligning question difficulty with the test population (Reeve, 2024).

c. Most questions have a discriminatory power of ≥ 0.30, indicating the test's ability to significantly differentiate between high- and low-performing students. High discriminatory power is an indicator of good question quality, as it demonstrates the ability to identify differences in participants' abilities (Peers, 2006).

3. Validity and Reliability Analysis

Table 3: Question Validity Based on CVR and EFA Analysis

Validity	Parameter	HNC Level 1	HNC Level 2
Content Validity	Average CVR Score	0,90	0,88
Construct Validity	Loading Factor ≥ 0.5	95% Questions	90% Questions

a. Content Validity: All test questions were reviewed by three native Mandarin-speaking lecturers, who rated them as "highly relevant" using the Content Validity Ratio (CVR). This resulted in an impressive average score of 0.90.

b. Construct Validity: Exploratory Factor Analysis (EFA) confirmed that all test items achieved factor loadings above 0.5, indicating that they effectively measure the intended constructs.

c. Reliability: The test demonstrated high reliability, with Cronbach's alpha scores of 0.87 for Level 1 and 0.83 for Level 2, showcasing excellent internal consistency.

The CVR-based content validity scores (0.90 for Level 1 and 0.88 for Level 2) affirm that the test items are highly relevant to their intended measurement goals. The CVR method is widely recognized for its systematic and quantitative approach to evaluating content relevance (Almanasreh et al., 2019). Furthermore, reliability scores exceeding 0.80 meet the benchmarks for dependable instruments in educational and psychological contexts (Kline, 2015).

The robust content validity and reliability of the HNC instrument highlight its effectiveness in achieving learning objectives. Using modern evaluation techniques, this tool offers a reliable and systematic way to support the measurable development of Mandarin language skills among students.

DISCUSSION

1. Effectiveness of the Instrument

The HNC Level 1 effectively serves as a reliable tool for assessing basic Mandarin language skills, especially in listening and reading. This is reflected in students' average scores, which are consistently categorized as "very good," and in the difficulty level, which is well-suited for beginners. These findings align with the principles of the "Comprehensible Input" theory, which suggests that learners acquire a new language most effectively when exposed to material that is just slightly beyond their current proficiency level. This approach proves particularly effective for Mandarin as a second language, ensuring students can gradually and confidently build their skills (Erbaugh, 2022; Krashen, 1982).

HNC Level 2 demonstrates effectiveness in measuring more complex initial abilities, such as the use of simple sentence structures and understanding the context of everyday conversations. Bloom's taxonomy in learning supports the importance of

Development of The Mandarin Language Proficiency Assessment Instrument Hanyu Shuiping Ceshi (HNC) as a Standard for Mandarin Language Skills for Unesa Students

evaluation based on analysis and synthesis at higher levels of ability (Bloom, 1956; Seaman, 2011). However, some items require revision to ensure alignment with students' skill levels.

HNC Levels 1 and 2 work hand-in-hand as effective tools for assessing Mandarin language proficiency, covering the range from basic to early intermediate levels. Grounded in well-established learning theories, these levels are carefully designed to provide evaluations that align with students' abilities. However, refining certain aspects of HNC Level 2 could further improve its accuracy and effectiveness as a language assessment tool. These enhancements are vital to ensuring that the learning process not only achieves its goals but also fosters the steady and long-term growth of students' Mandarin language skills.

2.   Relevance to HSK Standards and Learning Materials

Competency-based language learning focuses on creating a curriculum and evaluation methods that meet learners' needs and deliver meaningful outcomes (Burns & Richards, 2018). When evaluation tools consider the context of learners, they not only make learning more effective but also motivate students to improve their communication skills (Ellis & Shintani, 2013). This approach aligns with global language instruction standards, such as the HSK, which is designed to assess Mandarin proficiency step by step, addressing both educational and professional requirements. By using modern theories and methods, tools like these play a vital role in fostering sustainable and practical Mandarin language learning.

Incorporating a contextual approach into test questions, based on theories of context-driven learning, has additional benefits. It encourages learners to develop a deeper understanding of the language and apply critical thinking to real-world situations (Brown et al., 2000; Zhu et al., 2024). The tool developed here aligns with the HSK Levels 1 and 2, making it a great fit for beginners. By adding real-world context to the questions, it helps students learn Mandarin while also improving their problem-solving skills and ability to use the language in everyday life.

Moreover, this tool's alignment with international standards and graduate goals highlights its ability to build strong language skills. Ongoing updates to include more challenging levels will only make it more effective, ensuring it continues to support learners in achieving their goals.

3.   Implications of Use

The HNC test serves as both a formative and summative assessment tool for students in the Mandarin Language Program at UNESA. Formative assessments that adopt a contextual approach provide deeper insights into students' abilities, fostering a more adaptive and personalized learning process (Wiggins & McTighe, 2011). By incorporating scenarios based on everyday life, the test also supports the development of communication skills that are directly applicable to workplace settings. This aligns with the standards for teaching Mandarin to non-native speakers, which emphasize the importance of real-world relevance in teaching materials (Orton & Scrimgeour, 2019). Through this approach, the HNC test goes beyond evaluation, equipping students with the skills needed to navigate communication challenges in professional and global environments.

What sets the HNC test apart is its flexibility and focus on context, making it an effective tool for both formative and summative assessments. By promoting adaptive learning and emphasizing real-life relevance, the test not only evaluates student progress but also cultivates practical communication skills essential for workplace success. As a result, the HNC test plays a key role in preparing UNESA Mandarin Language Program students to meet global challenges with confidence and competence.

4.   Recommendations for Development

Incorporating interactive elements, such as listening comprehension exercises with audiovisual components, can significantly enhance students' listening skills. This approach aligns with Mayer (2014) research, which highlights the power of multimedia learning in improving cognitive abilities. Additionally, expanding the test to include Level 3 is essential for supporting advanced, tiered learning. This expansion reflects the principles of the "Zone of Proximal Development" (ZPD), which emphasizes designing progressive learning experiences that challenge students within their potential capabilities (Vygotsky, 1978; Xi & Lantolf, 2021).

By integrating interactive components and extending the test to a higher level, the HNC instrument can better support effective and structured Mandarin language learning. These enhancements not only strengthen students' listening skills but also introduce challenges that foster growth and mastery. As a result, this instrument will facilitate comprehensive language competency development, equipping students with the skills needed to excel academically and professionally in the future.

CONCLUSION

Research findings confirm that the HNC instrument is both relevant and effective in assessing the Mandarin language skills of UNESA students at the basic to early intermediate levels. The instrument demonstrates strong content and construct validity, with impressive CVR scores of 0.90 for Level 1 and 0.88 for Level 2. Its reliability is equally noteworthy, as reflected by Cronbach's alpha values exceeding 0.80. Comprehensive trial evaluations—conducted across one-to-one sessions, small groups,

and field tests—revealed that HNC Levels 1 and 2 consistently fall within the "good" to "very good" categories. The tests exhibit a balanced level of difficulty and strong discriminatory power, ensuring they effectively measure student capabilities. Beyond academic assessment, this instrument motivates students to cultivate practical communication skills, aligning with the demands of the workplace and preparing them for real-world challenges.

REFERENCES

1) Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221.

2) Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

3) Anggraeni, A., Supriatnaningsih, R., Pamuji, D. Y., & Pramono, S. A. (2022). Capaian Pembelajaran Lulusan Prodi Pendidikan Bahasa Mandarin UNNES Berbasis Kebutuhan Dunia Kerja di Bidang Industri. *Jurnal Cakrawala Mandarin*, *6*(2), 173–184.

4) Arikunto, S. (2006). Prosedur penelitian tindakan kelas. *Bumi Aksara*, *136*(2), 2–3.

5) Azwar, S. (2011). *Sikap manusia: Teori dan pengukurannya* (2nd ed.). Pustaka Pelajar.

6) Azwar, S. (2016). Reliabilitas dan validitas aitem. *Buletin Psikologi*, *3*(1), 19–26.

7) Bloom, B. S. (1956). Taxonomy of Educational Objectives : The Classification of Educational Goals, Handbook I Cognitive Domain. In *Educational Objectives*. New York : Longmans, Green and Co.

8) Brown, J., Collins, A., & Duguid, P. (2000). Situated Cognition and the Culture of Learning. *Psychology of Education: Major Themes: Pupils and Learning*, *2*, 207–230.

9) Burns, A., & Richards, J. C. (2018). *The Cambridge guide to learning English as a second language*. Cambridge University Press.

10) Creswell, J. W. (2018). *Penelitian Kualitatif & Desain Riset. Memilih di antara Lima Pendekatan* (Ahmad Lintang Lazuardi, Trans.). Yogyakarta: Pustaka Pelajar.

11) Djelantik, S. (2015). *Asia-Pasifik: Konflik, Kerja Sama, dan Relasi dan Antarkawasan*. Yayasan Pustaka Obor Indonesia.

12) Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. Routledge.

13) Erbaugh, M. S. (2022). The acquisition of Mandarin. In *The crosslinguistic study of language acquisition* (pp. 373–455). Psychology Press.

14) Gerber, S. B., & Finn, K. V. (2013). *Using SPSS for Windows: Data analysis and graphics*. Springer.

15) Kline, P. (2015). *A handbook of test construction (psychology revivals): introduction to psychometric design*. Routledge.

16) Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Oxford Pergamon Press.

17) Li, D. (2024). An interactive teaching evaluation system for preschool education in universities based on machine learning algorithm. *Computers in Human Behavior*, *157*, 108211.

18) Lubis, A. (2013). Pelaksanaan Standar Nasional Dalam Dunia Pendidikan. In -. Yogyakarta: Rineka Cipta.

19) Maria, M. (2019). Problematika Program Studi Pendidikan Bahasa Mandarin di Universitas Negeri di Indonesia. *Jurnal Cakrawala Mandarin*, *2*(2), 1–16.

20) Mayer, R. E. (2014). *Introduction to multimedia learning.*

21) Merdeka, M. B. K. (2020). Buku Panduan Merdeka Belajar-Kampus Merdeka. *Direktorat Jenderal Pendidikan Tinggi Kementerian Pendidikan Dan Kebudayaan*.

22) Nurfillaili, U., & Anggereni, S. A. (2016). Pengembangan instrumen tes hasil belajar kognitif mata pelajaran fisika pada pokok bahasan usaha dan energi SMA Negeri Khusus Jeneponto Kelas XI Semester I. *JPF (Jurnal Pendidikan Fisika) Universitas Islam Negeri Alauddin Makassar*, *4*(2), 83–87.

23) Orton, J., & Scrimgeour, A. (2019). *Teaching Chinese as a second language: The way of the learner*. Routledge.

24) Peers, I. (2006). *Statistical analysis for education and psychology researchers: Tools for researchers in education and psychology*. Routledge.

25) Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing*, *38*(2), 326–337.

26) Priyatno, D., & Korelasi, A. (2013). Regresi dan Multivariate dengan SPSS. *Yogyakarta: Gava Media*.

27) Reeve, B. (2024). Item Response Theory [IRT]. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 3689–3697). Springer.

28) Seaman, M. (2011). BLOOM'S TAXONOMY. *Curriculum & Teaching Dialogue*, *13*.

29) Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.

30) Wiggins, G., & McTighe, J. (2011). *The understanding by design guide to creating high-quality units*. ASCD.

31) Xi, J., & Lantolf, J. P. (2021). Scaffolding and the zone of proximal development: A problematic relationship. *Journal for the Theory of Social Behaviour*, *51*(1), 25–48.

32) Zhu, W., Hua, Y., & Wang, L. (2024). " I felt like I was on campus" creating a situated learning environment through Instagram. *Research & Practice in Technology Enhanced Learning*, *19*.