Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, Vol. 26, No. 1, pp. 37–47
The Forum

Examining HSK 3.0 Before Implementation: A Validity-Oriented Review of Linguistic Features and Task Structure

Kedi Mo
Teachers College, Columbia University

INTRODUCTION

The Hanyu Shuiping Kaoshi (HSK; 汉语水平考试, "Chinese Proficiency Test") has long served as one of the most widely used standardized tests of Chinese language proficiency for non-native speakers (Peng et al., 2021; Teng, 2017; Wang, 2018). Since its introduction in the 1980s, the test has undergone several revisions in response to changing views of language ability, proficiency standards, and assessment design (Teng, 2017). The current version, HSK 2.0, was released in 2009 and implemented in 2010, consisting of six levels, which its developers claimed aligned with the Common European Framework of Reference for Languages (CEFR) from A1 to C2 (Hanban, 2010). The main test assesses listening and reading at all levels and adds writing from HSK 3 onward, while speaking is offered through a separate, optional test, the HSKK (Hanyu Shuiping Kouyu Kaoshi; 汉语水平口语考试, "Chinese Proficiency Speaking Test"). The listening and reading sections rely predominantly on selected-response item formats, while writing tasks range from sentence construction at lower levels to short topic-based responses at higher levels. Over time, a growing body of research has raised concerns about aspects of HSK 2.0, including the interpretability of its proficiency levels, its representation of communicative ability, and its usefulness for broader score-based decisions (Fu et al., 2014; Peng et al., 2021; Su & Shin, 2015; Teng, 2017; Wang, 2018). These concerns have kept questions of test validity central to discussions of Chinese proficiency assessment.

Against this background, a new framework for the HSK (i.e., HSK 3.0) emerged with the publication of the Chinese Proficiency Grading Standards for International Chinese Language Education in 2021 (Center for Language Education and Cooperation [CLEC], 2021). Commonly associated with the new test, this framework adopts a "three stages and nine levels" structure and provides more detailed specifications of language knowledge and language use across proficiency levels. Although HSK 3.0 will not be fully implemented until July 2026, its official test specifications and structural documents already provide a substantial basis for examining how the revised test conceptualizes Chinese proficiency. As such, the new framework is well suited to a pre-implementation inquiry into how the test has been redesigned and what those revisions may imply for future validity research.

Validity, in educational measurement, refers to the degree to which empirical evidence and theoretical rationales support the interpretations and uses proposed for test scores (Kane, 2006, 2013). A pre-implementation, validity-oriented examination accordingly focuses on whether the design of a new test offers a sound conceptual basis for those interpretations and uses, before operational evidence becomes available (Chapelle, 2021). A full discussion of the nature of test validity is beyond the scope of this paper, but in a nutshell, test validity is generally

examined via the inferences that connect test performance to score meaning and score use, with scholars trying to find out the degree to which empirical evidence and theoretical rationales support the interpretations and uses proposed for test scores.

This study, therefore, undertakes a pre-implementation, validity-oriented examination of HSK 3.0 by reviewing its revised linguistic framework and task structure in relation to key validity concerns previously identified in HSK 2.0. Since the new test has not yet been operationalized, it is impossible to validate HSK 3.0 through test performance data or score-based evidence. While this study is not "pre-operational testing," i.e., the thoughtful, deliberate, and systematic process of collecting and analyzing evidence to support the validity of an assessment prior to that assessment's operational use (Kenyon & MacGregor, 2012), materials have been made available to all stakeholders of HSK on its official website (CLEC, 2025) that allow for a pre-implementation discussion on whether the design of a new test offers a sound conceptual basis for score interpretation and use.

To examine whether the revised framework appears to provide a stronger conceptual basis for score interpretation and use at the level of test design, the remainder of the paper proceeds in three steps: the next section reviews the validity concerns identified in HSK 2.0; the following section examines how HSK 3.0 revises the test's linguistic specifications and task structure, with particular attention to whether and how those concerns appear to be addressed; and the final analytic section considers what these revisions imply for the validity of HSK 3.0 score interpretation and use, while flagging questions that will require empirical investigation once the test is implemented.

VALIDITY CONCERNS IN HSK 2.0

Researchers have identified several interrelated concerns regarding the validity and implementation of the current HSK (2.0) test, which cluster into three areas: (1) a lowered difficulty threshold that appears misaligned with the external frameworks the test claims to reference, most notably the CEFR, (2) construct underrepresentation, particularly of speaking ability, and (3) an insufficient account of the professional target language use (TLU) domain. The first of these is most apparent at the lower levels, where vocabulary requirements, such as 150 words for Level 1 and 300 words for Level 2, appear insufficient to support the proficiency claims described in the test specifications. Several studies have pointed out that this design reflects the test developers' intention to make the test more accessible and thus attract a larger number of learners to promote Chinese language teaching and learning nationally and internationally (Luo et al., 2011; J. Zhang et al., 2012, as cited in Peng et al., 2021; Teng, 2017). However, this decision has contributed to concerns about the test's alignment with external proficiency frameworks, most notably the CEFR, which the test developers claimed to reference. Although such alignment problems are most pronounced at the lower levels, they are not confined to them. Peng et al. (2021, p. 332) provide a table mapping HSK levels to CEFR levels followed by summaries of findings that demonstrate such misalignments. For example, although HSK Level 4 has been aligned with CEFR B2, its writing component remains limited to sentence-level production and does not match the extended discourse expected at the B2 level, indicating the misalignment reaches the intermediate levels as well. These inconsistencies have raised concerns about score interpretation and the long-term credibility of the test.

A second major concern is construct underrepresentation, particularly in relation to speaking ability. Since the HSK does not include speaking as a mandatory component but instead offers a separate test, the HSKK, current HSK scores do not necessarily reflect learners' communicative speaking ability (Su & Shin, 2015; Wang, 2018). This structural separation may also shape test preparation practices, as learners are likely to focus on the skills tested in the main exam while giving less attention to oral proficiency. For example, they may concentrate their preparation on the listening, reading, and writing sections of the main exam while postponing or skipping the optional HSKK. Moreover, even the HSKK does not fully resolve this concern, as its task formats tend toward presentational oral production (e.g., repetition and retelling) instead of the kind of interactive communication that is central to communicative language ability (see, e.g., Bachman, 1989; Sato & McNamara, 2019). Related to this issue is the concern over the limited authenticity of test tasks. Selected-response formats dominate the listening and reading sections and only partially reflect real-world language use (Teng, 2017). At the lower levels, some listening materials have been described as overly simplified and lacking naturalness. In the writing section, tasks such as sentence construction are often decontextualized and therefore only loosely connected to authentic language use. In addition, some indirect writing tasks have been found to correlate more strongly with reading ability than with writing proficiency, raising further concerns about construct validity (Fu et al., 2014).

Finally, the HSK is positioned not only as an academic gateway but as a multi-purpose test whose stated score uses extend to employment and professional selection (Peng et al., 2021). Against this professional TLU domain, however, the test has shown limited utility in professional contexts. In the workplace, it is not commonly used as a primary criterion for recruitment or promotion, and employers have reported limited confidence in HSK scores as indicators of functional Chinese proficiency in professional settings. Wang (2018, p. 138) surveyed participants from organizations that incorporated Chinese language test certificates into their recruitment and promotion processes but found that more than 75% of them did not consider such certificates a mandatory requirement for hiring. Additionally, employers may interpret high HSK scores as reflecting test-taking ability rather than workplace communicative competence. Consequently, many organizations prefer to rely on their own internal assessments, while business users also report a lack of clear guidance on how HSK scores should be interpreted and used for professional selection purposes.

DESIGN FEATURES OF THE HSK 3.0 FRAMEWORK

The concerns outlined above provide the backdrop for HSK 3.0, which is intended not merely as an expansion of HSK 2.0, but as a substantive revision of how Chinese proficiency is conceptualized and assessed. Developed on the basis of the Chinese Proficiency Grading Standards for International Chinese Language Education (CLEC, 2021), the revised framework is examined in two dimensions reflected in currently available test specifications and test-structure documents: the specification of linguistic resources and the definition of communicative tasks.

Linguistic Revisions

With the new version of the HSK confirming its implementation in mid-2026, updated test specifications have been made publicly available (CLEC, 2025). Table 1 presents the differences between HSK 2.0 and HSK 3.0 in the specification of linguistic resources, including both character and vocabulary requirements. In the revised framework, HSK 3.0 organizes its nine levels into three explicitly named stages: Beginner (Levels 1–3), Intermediate (Levels 4–6), and Advanced (Levels 7–9). Whereas HSK 2.0 claimed a one-to-one correspondence with the six CEFR levels (A1–C2), HSK 3.0 is anchored in a more independently developed national standard that defines its nine levels internally, with any correspondence with the CEFR depending on empirical linking rather than an explicit design claim (CLEC, 2021). The highest band, Levels 7–9, is administered as a single integrated test encompassing listening, reading, writing, translation, and speaking, with scores from each skill contributing to an overall level classification based on ability estimates. Beyond publicly available descriptions indicating that test takers are further classified into four levels, namely "Below HSK (Level 7)," "HSK (Level 7)," "HSK (Level 8)," and "HSK (Level 9)," further details regarding the operationalization of these levels remain limited at the current stage.

The most structurally significant change in linguistic specification is the explicit differentiation between character recognition and character writing in HSK 3.0. In HSK 2.0, character knowledge was represented as a single cumulative figure per level, reflecting the total number of unique characters expected at each stage without any formal distinction between receptive and productive character ability. The official HSK 2.0 syllabus does not provide explicit per-level character counts; rather, these figures can be derived from the unique characters contained in each level's vocabulary list (Hanban, 2010), as character knowledge in HSK 2.0 was effectively subordinate to vocabulary acquisition. The syllabus specified the vocabulary learners were expected to master at each level, and the associated character demands were an implicit by-product of that vocabulary requirement with no official benchmarks distinguishing how many characters test takers were expected to recognize versus produce in writing. HSK 3.0 replaces this unified treatment with two separate cumulative benchmarks: one for characters test takers are expected to recognize, and one for characters they are expected to produce in writing. As shown in the table, these two trajectories develop at different rates. Recognition requirements reach 1,940 characters by Level 6 and expand substantially to 3,088 at the 7–9 band, while writing requirements accumulate more gradually, reaching 850 by Level 6 and 1,200 across the full framework. Compared with HSK 2.0's cumulative character total of 2,663, the overall recognition scope in HSK 3.0 is modestly expanded, though the distribution across levels is notably different: recognition demands in 3.0 are somewhat lower than in 2.0 at Levels 5 and 6 (1,527 vs. 1,685 and 1,940 vs. 2,663, respectively), with the overall expansion concentrated at the advanced 7–9 band.

Vocabulary requirements are also expanded, and again both systems are cumulative. The scale of expansion, however, is more substantial than in the character domain. At the lower levels, vocabulary requirements approximately double from 150 to 300 words at Level 1 and from 300 to 500 at Level 2 with similar proportional increases through Level 4. The expansion is comparatively modest at Level 6, where the requirement rises from 5,000 to 5,400 words. The most substantive addition is the new 7–9 band, which more than doubles the overall framework ceiling from 5,000 to 11,000 words. Taken together, the revised character and vocabulary specifications reflect a more differentiated account of linguistic knowledge across proficiency levels, though the nature and extent of these changes vary considerably across the framework, with the largest increases concentrated at the upper end of the proficiency continuum.

TABLE 1
Comparison of Character and Vocabulary Requirements in HSK 2.0 and HSK 3.0

Level	Characters			Vocabulary	
	HSK 2.0	HSK 3.0		HSK 2.0	HSK 3.0
		Recognition	Writing		
1	174	246	50	150	300
2	348	371	100	300	500
3	618	655	250	600	1,000
4	1,064	1,096	400	1,200	2,000
5	1,685	1,527	550	2,500	3,600
6	2,663	1,940	850	5,000	5,400
7–9	N/A	3,088	1,200	N/A	11,000
Total	2,663	3,088	1,200	5,000	11,000

Note. All numbers are cumulative. HSK 3.0 has separate cumulative totals for recognition characters and writing characters. The official syllabus for HSK 2.0 provides no per-level character counts, but the numbers can be derived from the unique characters found in each level's vocabulary list (Hanban, 2010).

Task Revisions

HSK 3.0 further introduces task-specific requirements that are not present in HSK 2.0, which reflects its shift toward a more communicative and task-oriented framework. A direct comparison across the two versions is only partly possible because HSK 2.0 did not explicitly define communicative "tasks" or functional objectives but rather focused on the mastery of characters, vocabulary, and grammar and treated speaking as a uniformly separate and optional component through the HSKK, which test takers could choose to take independently of the main written examination. By contrast, HSK 3.0 restructures the role of speaking across its three stages. At Levels 1–2, the HSKK remains optional and separately administered. At Levels 3–6, however, the HSKK becomes a mandatory component bundled with the written test upon registration, with each written level paired with a corresponding HSKK. At Levels 7–9, speaking is fully integrated into a single combined examination rather than administered as a separate test. HSK 3.0 also introduces an explicit classification of tasks in the test specifying the types of activities that test takers are expected to carry out in Chinese. These task descriptors are organized by level and linked to particular language skills including listening, speaking, reading, writing, and translation. Table 2 presents the task descriptors for each level, including the number of tasks listed and a summary of their key characteristics. A closer review of all tasks within the same level also makes it possible to identify the language skills required to complete

each task. For example, Task Item 4 at Level 1 focuses on 问答天气 ("asking and answering about the weather"). Under this broader task label, the content further specifies objectives through verbal phrases such as 听懂 ("understand what is heard"), 询问 ("ask/inquire"), 介绍 ("introduce/describe"), and 看懂 ("understand what is read/seen").

This level of specification helps streamline the process of identifying which skills are being activated across tasks, and the task types and their key characteristics reveal a clear pattern of progression across levels. At the lower levels, tasks appear to be primarily situated in everyday contexts and generally involve recognition and constrained-response formats, such as identifying personal information, understanding basic descriptions, and answering simple questions. These tasks require relatively limited language production and place greater emphasis on basic comprehension and simple exchanges. At the intermediate levels, task types then expand to include greater control over sentence- and paragraph-level production and involve a wider range of communicative functions. For example, the task demands begin to include narrating experiences, describing processes, and participating in more extended interactions on familiar topics. More complex and contextually embedded activities are further identified through increasingly demanding task requirements at the higher proficiency levels. At the highest level (Levels 7–9), task types encompass academic research, professional communication, and translation across domains such as legal and medical contexts. They are situated in business, academic, and institutional settings (e.g., reporting information, discussing abstract topics, and engaging with formal texts) in professional and other advanced communicative scenarios.

TABLE 2
Summary of Task Types, Skills Assessed, and Key Characteristics Across HSK 3.0 Levels

Level	Task Number	Task Types	Key Characteristics	Skills Assessed
1	15	Basic identification of personal info, weather, travel, shopping, and study.	Focuses on simple Q&A and recognizing basic factual information in daily life.	L, S, R
2	17	Descriptions of situations, comparisons of items, and simple social interactions.	Includes simple descriptions and comparisons; adds basic writing like filling out forms or writing dates.	L, S, R, W
3	22	Narrating experiences, discussing transportation rules, and simple office/nature topics.	Focuses on general tasks; requires narrating events (beginning, process, result) and writing simple sentences.	L, S, R, W
4	30	Discussing social phenomena, new	Tasks involve a certain level of complexity; requires	L, S, R, W

		technology, economy, and historical stories.	relatively fluent expression and paragraph-level writing.	
5	28	Business handling, study reports, sharing research, and discussing ancient history/tech.	Focuses on more complex situations; requires more fluent exchange, reporting, and writing short essays.	L, S, R, W
6	24	Formal policy/regulation documents, food safety, career positioning, and traditional thought.	Involves complex situations; requires high-level fluency, formal elaboration, and full-length article writing.	L, S, R, W
7–9	30	Academic research, market analysis, legal/medical translation, and professional debate.	Professional/high-level scenarios; emphasizes deep analysis, academic rigor, and professional translation.	L, S, R, W, T

Note. L = listening; S = speaking; R = reading; W = writing; T = translation. For Levels 1–2, the speaking test is optional. For Levels 3–6, candidates must take the separate HSKK test. For Levels 7–9, the speaking component is included in the main HSK test itself.

VALIDITY IMPLICATIONS OF THE HSK 3.0 REVISIONS

The revisions introduced in the upcoming HSK invite reflection on whether score interpretation and use are more likely to align with the test's intended purposes. Again, the discussion is necessarily prospective as it considers what the revised framework promises rather than what has yet been demonstrated through operational testing. Building on the previously established framing, four specific and interrelated questions guide the discussion: (1) how clearly the test specifies what is being measured; (2) how broadly proficiency is represented within the task framework; (3) how plausibly test performance maps onto language use beyond the testing context; and (4) how effectively scores may support intended decisions, particularly at the upper proficiency levels.

Compared with HSK 2.0, HSK 3.0 specifies linguistic knowledge in a more differentiated and cumulative manner, with a clearer three-stage, nine-level progression, an expanded lexical scope, and an explicit distinction between character recognition and character writing. These changes make the intended performance domain more explicit by defining with greater precision the linguistic knowledge and abilities expected at each level. By providing a more detailed account of how language demands develop across stages, this specification may improve the interpretability of score differences by linking them to a more clearly articulated proficiency continuum. A more detailed internal proficiency framework may therefore strengthen score interpretation within the HSK system. However, stronger internal specification alone does not

establish external alignment, so claims of equivalence with frameworks such as the CEFR would still require separate empirical linking and standard-setting research.

Moving beyond domain definition, another important feature of HSK 3.0 is its broader and clearer representation of proficiency through communicative tasks and language use across multiple skills. In this respect, the revised test specifications reflect a shift from a narrower, knowledge-based description of language ability toward a broader account of proficiency as communicative performance. The treatment of speaking across the three proficiency stages illustrates both the progress made and the limits of that progress. In HSK 2.0, the HSKK was uniformly optional and separate, meaning that speaking could be entirely absent from a candidate's proficiency record regardless of level. HSK 3.0 introduces a more differentiated arrangement: speaking remains optional at Levels 1–2, becomes a mandatory bundled component at Levels 3–6 where registration for the written test requires concurrent registration for the corresponding HSKK, and is fully integrated into a single combined examination at Levels 7–9. This staged approach represents a structural improvement over HSK 2.0, particularly at the intermediate levels where oral assessment is now required rather than elective. At the same time, the concern identified in relation to HSK 2.0 is not fully resolved. At Levels 1–6, the speaking component continues to be delivered through the current HSKK, whose task formats (e.g., sentence repetition and text retelling) tend to underrepresent the kind of interactive communication central to communicative language ability (Bachman, 1989; Sato & McNamara, 2019). Whether the speaking component integrated into the Levels 7–9 examination departs from these formats cannot yet be determined, as its detailed task design has not been made publicly available. Accordingly, the revision broadens and deepens the operational role of oral assessment, but whether the construct of speaking is more validly represented will depend substantially on whether these task formats are revised to better capture interactive oral ability, particularly those of the current HSKK at Levels 1–6.

One further question regards how plausibly HSK scores can be read as reflecting language use beyond the test itself. The revised framework specifies communicative task domains more explicitly, ranging from everyday interaction at lower levels to academic, institutional, and professional tasks at higher levels. By making task domains more visible, HSK 3.0 strengthens the conceptual link between score meaning and broader language use. In this respect, the framework appears to offer a more plausible basis for inferring real-world language ability than HSK 2.0. Nevertheless, this remains a design-level advantage rather than a demonstrated empirical one. Whether operational items and tasks will actually instantiate these intended domains with sufficient authenticity remains to be examined after implementation.

A final implication concerns score use, especially at the upper levels. The HSK 3.0 test specifications describe upper-level task types in considerably more concrete terms than HSK 2.0, including formal discussion, research-related communication, and specialized translation in domains such as legal and medical contexts. For institutions and employers, this provides a clearer basis for understanding what an upper-level score is intended to entail. In principle, the richer description should make upper-level scores more interpretable for decisions about academic participation and professional selection where HSK 2.0 was often seen as insufficient (Peng et al., 2021; Wang, 2018). However, the publicly available documents stop short of providing explicit guidance on how scores should be used to support particular institutional or employer decisions. The interpretive richness of the new task descriptors is, therefore, not yet matched by a comparably detailed account of score use.

DISCUSSION AND CONCLUSION

This study examined HSK 3.0 through a pre-implementation review of its revised linguistic framework and task structure in relation to key validity concerns previously identified in HSK 2.0. Overall, the review suggests that HSK 3.0 is not simply an expanded version of the existing test but a revision that redefines how Chinese proficiency is described and, potentially, how it may be interpreted. Based on the currently available test specifications and structural documents, the new framework appears to offer a clearer specification of linguistic progression and a structurally more integrated treatment of oral assessment, which together provide a more explicit connection between proficiency levels and communicative task domains. In this sense, HSK 3.0 appears to provide a stronger conceptual basis for score interpretation than HSK 2.0.

At the same time, the evidence considered here remains largely design-based rather than implementation-based. For this reason, the present analysis does not demonstrate that HSK 3.0 already provides a valid solution to the concerns raised about HSK 2.0. Rather, the revised framework should create more favorable conditions for a stronger validity argument. Among the concerns identified in HSK 2.0, domain definition appears to be the most clearly addressed at the level of framework design given the more differentiated and cumulative specification of linguistic resources and the explicit articulation of communicative task types across levels. Yet, the concern over construct underrepresentation of speaking is partially addressed, with the mandatory bundling of the HSKK at Levels 3–6 and the full integration of speaking at Levels 7–9 representing a meaningful structural shift from HSK 2.0's uniformly optional arrangement. Nevertheless, the underlying question of whether the HSKK's presentational task formats adequately represent communicative speaking ability remains open. The concern over the professional TLU domain shows the least resolution at the design level. While upper-level task descriptors are more concrete, no explicit guidance on score use for professional selection has been provided. In addition, the authenticity of actual test tasks cannot yet be evaluated on the basis of framework documents alone.

Overall, the currently available HSK 3.0 documents point to a substantive shift in how Chinese proficiency is conceptualized for assessment. In a broader context, the new test is a more ambitious and defensible framework, but its validity still needs empirical demonstration. Future research should examine how the revised construct is operationalized in live test forms, whether mandatory oral assessment at intermediate levels improves score validity, and how task performance generalizes to academic and professional use. The available documents point to a conceptual shift and provide a clearer starting point for validation, yet the central question remains whether this stronger design will yield stronger validity in practice.

DECLARATION OF AI USE

During the preparation of this work, the author used Elicit (Basic plan) to assist in locating relevant scholarly literature and Claude (Opus 4.8) to assist with copyediting and formatting references in APA 7 style. All conceptual framing, analysis, interpretation, and conclusions are the author's own. After using these tools, the author reviewed and edited all content as needed, independently verified all citations, and takes full responsibility for the content of the publication.

REFERENCES

Bachman, L. F. (1989). Language testing–SLA research interfaces. *Annual Review of Applied Linguistics, 9*, 193–209. https://doi.org/10.1017/S0267190500000891

Center for Language Education and Cooperation. (2021). *Chinese proficiency grading standards for international Chinese language education.* Beijing Language and Culture University Press.

Center for Language Education and Cooperation. (2025). *汉语水平考试（HSK）考试大纲* [Syllabus for the Chinese Proficiency Test]. https://hsk.cn-bj.ufileos.com/3.0/%E6%96%B0%E7%89%88HSK%E8%80%83%E8%AF%95%E5%A4%A7%E7%BA%B21219.pdf

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE Publications. https://doi.org/10.4135/9781071878811

Fu, H., Zhang, J., Li, Z., Zhang, T., & Xie, N. (2014). 新汉语水平考试 HSK(六级)的性别公平性评估 [The evaluation of the gender equity of new HSK (Level 6)]. *考试研究* [Examination Research], (1), 35–38.

Hanban. (2010). *新汉语水平考试大纲* [*The new HSK test syllabus*]. The Commercial Press.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

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

Kenyon, D. M., & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (1st ed., pp. 295–306). Routledge. https://doi.org/10.4324/9780203181287

Luo, M., Zhang, J., Xie, O., Huang, H., Xie, N., & Li, Y. (2011). 新汉语水平考试(HSK)质量报告 [Report on the quality of new Chinese proficiency test (HSK)]. *中国考试* [China Examinations], (10), 3–7.

Peng, Y., Yan, W., & Cheng, L. (2021). Hanyu Shuiping Kaoshi (HSK): A multi-level, multi-purpose proficiency test. *Language Testing, 38*(2), 326–337. https://doi.org/10.1177/0265532220952972

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics, 40*(6), 894–916. https://doi.org/10.1093/applin/amy032

Su, Y., & Shin, S.-Y. (2015). Test review: The new HSK. *Iranian Journal of Language Testing, 5*(2), 91–103. https://www.ijlt.ir/article_114409_a2974eef97d75c8a589cdd2ebe89dba8.pdf

Teng, Y. (2017). Hanyu Shuiping Kaoshi (HSK): Past, present, and future. In D. Zhang & C.-H. Lin (Eds.), *Chinese as a second language assessment* (pp. 3–19). Springer. https://doi.org/10.1007/978-981-10-4089-4_1

Wang, S. (2018). *Investigating the consequential validity of the Hanyu Shuiping Kaoshi (Chinese proficiency test) by using an argument-based framework* [Doctoral dissertation, McGill University]. McGill University eScholarship. https://escholarship.mcgill.ca/concern/theses/4q77ft93g

Zhang, J., Li, P., Li, Y., Xie, N., & Huang, L. (2012). 对汉语水平考试(HSK)的新思考 [New thinking on the New HSK]. *中国考试* [China Examinations], (2), 50–53.

Kedi Mo is an Ed.D. student in Applied Linguistics at Teachers College, Columbia University. His research interests include technology-mediated language education, AI-driven tools for language teaching and teacher training, and automated language assessment and test validation. Correspondence should be sent to Kedi Mo, E-mail: km3690@tc.columbia.edu.