Chapter 1
Large-Scale English Writing Assessment for Chinese Learners of English: An Introduction to Part I

Yan Jin

Abstract This chapter provides an introduction to Part I, which focuses on large-scale English writing assessment for Chinese learners of English. The chapter begins with an overview of the development of English writing assessment for Chinese learners of English over the past half century. This is followed by a discussion on the benefits of performance assessment and the inadequacy of task authenticity in large-scale writing assessments. A chapter-by-chapter summary is then provided for Chaps. 2, 3, 4, 5, 6 and 7 and brief comments are made on the strengths and weaknesses of each chapter. Finally, the chapter highlights the need for improving the construct validity of large-scale, standardized writing assessments so as to promote the teaching and learning of writing in English for real communicative purposes.

Keywords English language writing assessment · Large-scale writing assessment · Chinese learners of English

English language education in China has been changing rapidly in response to the changing social conditions and needs since the founding of the People's Republic of China in 1949 (Dai & Hu, 2009). In the 1950s, Russian was taught as the first foreign language in high schools and universities. The overwhelming predominance of Russian gave way to English in the 1960s when the relations between the two countries became increasingly strained. The first national teaching syllabus of English as a foreign language for institutions of higher education was published in 1962. During the mid-1960s to mid-1970s, English language education in mainland China was interrupted by the Cultural Revolution. After a 10-year hiatus, the National Unified Enrolment Examination (NUEE) for admission to higher education institutions was resumed in 1977 and English became a compulsory component of the NUEE in 1983.

Y. Jin (✉)
Shanghai Jiao Tong University, Shanghai, China
e-mail: yjin@sjtu.edu.cn

L. Hamp-Lyons, Y. Jin (eds.), *Assessing the English Language Writing of Chinese Learners of English*, https://doi.org/10.1007/978-3-030-92762-2_1

When the Matriculation English Test (MET), the English examination of the NUEE, was designed, the test developer aimed to achieve validity of its writing assessment by adopting direct writing tasks and avoiding the "contextless" and "constructless" (Hamp-Lyons & Condon, 2000: 10) approach of using multiple-choice questions (Li, 1990). Take the 1987 MET writing task as an example. The task involves a situation: An American student visits China and meets the candidate at a party, where the two, being seated next to each other, take each other's notebook by mistake. The candidate was supposed to send back the American's notebook with a letter explaining the circumstances and asking the American to send back his or her own notebook. Some of the key elements of a communicative writing task can be clearly identified in the task, for example, the purpose and the audience of writing. At the tertiary level, the national English teaching syllabus was revised in the mid-1980s. This was followed by the inception of the College English Test Band 4 (CET-4) in 1987 and Band 6 (CET-6) in 1989. Similar to the MET, the CET Writing adopted the format of composition writing. The writing task of the first CET-4 test in 1987 was a guided composition on the topic "Women in the Modern World". Since then, composition writing has remained a compulsory component of the CET, accounting for 15% of the total score (Jin, 2019).

The inclusion of direct writing tasks in the high-stakes English language tests has had major impact on teaching and learning in China. Before the 1990s, the use of English was largely missed out in English language teaching (ELT) and "ELT in schools . . . was a matter of teaching the form of English as knowledge" (Li, 1990: 396). When the MET was designed, the test developer was faced with a conflict: "on the one hand it cannot cut itself off from the state of the art of ELT in schools, on the other hand it must break away to achieve validity" (p. 369). The outcome was "a mixed test with two somewhat incompatible major components: the 'knowledge' component that represents a concession to the existing state and tests formal knowledge of grammar, vocabulary and phonetics in psychometric-structuralist tradition, and the 'use' component which is intended as an embodiment of new psycholinguistic-sociolinguistic concepts and tests the use of English as directly as possible through reading, writing, listening and speaking" (ibid.). A survey of the MET washback conducted in six provinces identified changes in teaching materials, teaching content, and extracurricular activities as a result of the introduction of the direct writing task, indicating clearly "a shift from formal linguistic knowledge to practice and use of the language" (Li, 1990: 402).

Similarly, a chief purpose of the CET was to promote the implementation of the national teaching syllabuses (Yang, 2003). A collaborative validation study was conducted during 1991–1995 by the National College English Testing Committee and the British Council, which demonstrated a steady, albeit small, increase in the mean scores of the CET writing (Yang & Weir, 1998). To further promote the teaching and learning of English language writing, a minimum score of the writing component was required for a CET certificate in the late 1990s. Since the introduction of this policy, teachers and learners in tertiary institutions have attached greater importance to the teaching and learning of writing. Further improvements in test takers' performances on the CET-4 Writing were observed: during the three five-

year periods from 1987 to 2001 (i.e., 1987–1991, 1992–1996, and 1997–2001), the mean scores (out of a total of 15 points) of the CET-4 Writing were 4.5, 6.0, and 7.5 for the entire test population, and 5.5, 7.5 and 8.5 for key universities (Jin & Yang, 2006).

The driving force for performance assessment, as noted in Yu (2014: 616), is the "close similarity or proximity between the performance and the construct of interest". By using performance-based tasks, writing assessments are more likely to achieve construct validity. Performance assessment in high-stakes contexts, however, presents practical challenges. In pursuit of fairness, the provision of context has to be compromised to standardize testing conditions, or at the very least, reduce contextual variability and elicit comparable performances for consistent scoring by trained raters. Topic bias should also be avoided by carefully monitoring possible differential item functioning due to test takers' gender, disciplinary background, socio-economic status, and so on.

To establish the communicative context in the writing tasks of a high-stakes test, the test designers have to make a serious attempt to specify in detail such contextual facets as task format, prompt, intended audience, genre, length of the output, and responding time. In particular, the input material, the length of the output, and the response time need to be tightly controlled. As a result, however, task authenticity, the very strength of performance assessment, has been compromised. That is, the writing tasks may lack "interactional authenticity" (Bachman, 1991: 691) and test takers may not be engaged in activities of a truly communicative nature.

The inadequate authenticity is also reflected in scoring criteria, which are the *de-facto* constructs of writing tasks. Performances on essay writing tasks are generally scored for content relevance, discourse coherence and cohesion, and language quality. Cumming (2002: 73) noted that "formal tests" of writing should also fulfill ethical criteria of "confidentiality, prior orientation, fairness, and equality of opportunity" by assuming "a pragmatic, functional definition of second-language (L2) writing in which an examinee's text production is judged normatively in respect to conventions for a discourse type or domain". In large-scale writing assessments, such a functional, pragmatic ideology is often adopted. What is missing, however, is "a developmental orientation to foster creative, personal expression" or "a political orientation to challenge or critique societal norms" (ibid.: 75–76).

1.1 Outline of Chapters 2 to 7

The seven chapters in Part I provide a good coverage of large-scale tests currently in use in China and beyond, including two international tests, Pearson Test of English-Academic (PTE Academic) and Aptis-General, and two tests developed mainly for local uses, College English Test (CET) and General English Proficiency Test (GEPT). The tests concerned are on a large scale and are used for making high-stakes decisions. An extended introduction to Chaps. 2–7 is provided below.

Chapter 2 by Shaoyan Zou reports a study of the impact of two types of rating scales, a holistic scale and an analytic scale, on the CET-4 essay writing task. The results favored the analytic scale for its better control of rater variation and scoring consistency. The categories of the analytic scale also functioned satisfactorily in discriminating test takers' writing performances. Follow-up interviews showed that teachers/raters preferred the analytic scale for its explicit performance descriptors and potential for diagnostic feedback. The only reservation about the analytic scale, in the view of the teachers/raters, was the practicality of using an analytic scale for a test with over 20 million test takers a year. A limitation of the study, as admitted by the author, is the lack of voices from test takers, whose views on task requirements, scoring criteria and score report may yield interesting findings about the strengths and weaknesses of each type of rating scale.

In Chapter 3, Yu-Hua Chen and Ying Zheng investigated the linguistic features of Chinese learners' English language writing in the independent (essay writing) and integrated (read-to-summarize and listen-to-summarize) writing tasks of PTE Academic. A comparison of the scores on the three tasks showed that Chinese learners achieved higher scores than non-Chinese test takers on the read-to-summarize and essay writing tasks. Further comparisons between the two groups on their use of recurrent word combinations, or lexical bundles, revealed that Chinese learners produced lengthier responses and used significantly more lexical bundles in all the three tasks. The read-to-summarize task elicited the most frequent use of prompt-based lexical bundles by both groups. The study affirms the need to re-define the construct of writing in English for academic purposes by incorporating the aspect of engaging with source materials of different modes. It is however worth noting that prompt-based summary writing differs from integrated essay writing, rendering it less comparable with independent essay writing.

Chapter 4 by Ying Chen and Xiaoxian Guan presents a study of how Chinese test takers conceptualize and construct audiences when working on the Aptis-General Writing Task 4. Think-aloud data were collected to look into test takers' processes of writing, and a follow-up questionnaire survey and face-to-face interviews were conducted to further tap into test takers' awareness and construction of targeted audience in the process of writing. Results showed that in both informal and formal email writing tasks, Chinese test takers took audience into consideration by analyzing the features of their audience and making efforts to meet the audience's expectations. Differences in audience-related strategies were identified between the informal and formal email writing. It is interesting to note that, although few test takers regarded the rater as their audience, they did take into consideration the rater by playing safe in their choices of words and structures, indicating that no matter how hard the test developer may have tried, test tasks could at best simulate real-life activities.

In Chapter 5, Naihsin Li examined learning strategies employed by Taiwanese learners of English and the effectiveness of strategy use on performances in the GEPT High-Intermediate Writing test. Data were collected through a questionnaire survey among GEPT test takers, focusing on five categories of learning strategies: cognitive strategies, affective strategies, seeking practice opportunities, planning and evaluation, and self-regulation. An SEM analysis showed that metacognitive

strategies governed or controlled the use of other types of strategies. A comparison of successful and unsuccessful writers revealed that the two groups had different patterns of learning strategy use and that the unsuccessful group committed significantly more errors and more varieties of errors. The effect of learning strategy use on test performance, however, needs to be interpreted with caution because no causal relationship was proved and test performance could have been affected by test takers' use of test-wise strategies. Nonetheless, the findings of this study have valuable implications for the instruction of learning strategies specific to the skill of English language writing.

Chapter 6 by Yan Zhou and Ke Bin addresses the issue of construct validity of integrated writing tasks with a special focus on the use of source materials by Chinese learners of English. Using questionnaire surveys, the study explored the amount and the pattern of source material use in three types of integrated writing tasks: read-to-write, listen-to-write, and read-listen-to-write. The low proficiency group reported to have used less material in the listen-to-write task than the other two tasks. Different patterns of source material use between the two proficiency groups and across the three types of tasks were also observed. The study suggests that the ability to use an appropriate amount of source materials for achieving various purposes is integral to the construct of integrated writing tasks and that source material at an appropriate level of difficulty is essential for integrated writing tasks. A limitation of the study is its sole reliance on self-reporting data, producing little direct evidence for the source material use in integrated writing tasks.

Chapter 7 by Mingwei Pan addresses the issue of how performance standards can be developed and used for the teaching, learning and assessment of English language writing. The chapter begins with an introduction to the context of the China's Standards of English Writing (CSE-W) project, followed by a review of the literature of EFL/L2 writing ability and existing proficiency scales of English language writing, thus laying the foundation for the definition of the construct of the CSE-W. The process of collecting and calibrating the descriptors of CSE-W subscales was then reported in detail. In the second part of the chapter, the application of the CSE-W was explored, focusing on the use of the CSE-W subscales for formative assessment of English language writing. Finally, challenges facing the application of the CSE-W for the teaching, learning and assessment of English language writing were discussed and suggestions were made as to the appropriate use of the CSE-W for assessment purposes.

1.2 Key Issues in Large-scale Writing Assessment

The issues addressed in the six chapters may be familiar to language testing researchers and practitioners, and they may not even be unique to the writing assessments of Chinese learners of English. Bachman (2010: x), however, noted that "the enormity of the enterprise in this (assessing Chinese learners' English)

context magnifies kinds of problems that are faced by language testers everywhere, and makes it more difficult to find justifiable solutions".

In the Chinese context, a test could easily derive its "extrinsic power" from its size and official authority. "For a test to be truly, positively powerful", however, Li (1990: 394) argued, "its extrinsic strength needs to be combined with intrinsic strength". Construct validity gives a test its intrinsic strength. Weir (2005) highlighted the role of contextual facets in operationalizing test constructs (see Jin, 2020 for two case studies). In conventional writing tests, constraints imposed on the context of writing by the need for standardization, however, may pose a threat to the construct representation. An analysis of the contextual facets of the writing tasks covered in Part I suggests possible construct under-representation: test takers are typically required to write a short essay within the time limit. In terms of genre, the international tests include essay writing, summary, or email writing, whereas the local tests assess argumentative writing only. Argumentative essay writing however may not be the most relevant target-language-use activity for the test population. In a needs analysis of English for professional purposes conducted among university graduates in mainland China, argumentative essay was found to be the least common type of writing in workplaces (Jin & Hamp-Lyons, 2015). An analysis of the scoring criteria of the writing tasks also indicates that writing is viewed more as a language problem than a writing problem, probably because linguistic features can be more objectively scored than ideas and styles of writing. More than two decades ago, Hamp-Lyons and Kroll (1997: 18) cautioned against what they call a "snapshot approach" to writing assessment and argued for expanding the definition of the construct of second-language writing beyond what conventional tests have attempted to assess. Cumming (2002: 78) also asked a rhetorical question: If writing can fulfill emancipatory functions in educational practices, why can't it do so in assessment contexts as well?

When the power of a test is exercised by its users, the test will have washback on teaching and learning. Bachman (2010: x) observed that "for many of these tests (English language tests in China), providing 'positive washback' on instruction is explicitly stated as a purpose" and that "(T)he intended consequence of promoting positive washback on instruction is perhaps the single characteristic that distinguishes many of these tests from high-stakes language tests in other parts of the world". The most worrying problem about the washback of standardized writing tests relates to a writing style specifically used for examination essays: to achieve higher scores, learners are encouraged or trained to memorize model essays and produce linguistically or structurally beautiful essays with vacuous ideas, referred to as "new eight-legged essays". In the imperial examinations during the Ming and Qing dynasties, test takers were required to read Confucian classics and produce eight-part responses to examination questions. That is, the responses must follow the sequence of (1) breaking the topic, (2) receiving the topic, (3) beginning discussion, (4) initial leg, (5) transition leg, (6) middle leg, (7) later leg, and (8) conclusion (Elman, 2009: 696). So the term "eight-legged essays" is often used to refer to essays which have a fixed structure but lack novel ideas. Washback of writing tests,

therefore, should be high on the research agenda, so as to promote the teaching and learning of writing in English for real communicative purposes.

The enormous size of the enterprise of language testing in the Chinese context may also take its toll on individual learners. In a study of the TWE, the writing component of the paper-based TOEFL, Hamp-Lyons and Kroll (1997: 21) commented that "(W)e understand a great deal less about our test takers from countries around the world than we need to" and that "(T)his is the great underresearched aspect of language testing". Although this book is not devoted specifically to studies of test taker characteristics, an exclusive focus on Chinese learners of English would no doubt facilitate a better understanding of this group of writers. It is however worthwhile noting that the term Chinese learners is "a trade-off between generalization and diversity" and that we should "avoid reduction and oversimplification through labelling as 'Chinese' or a false sense of sameness and homogeneity" (Cortazzi & Jin, 2011: 314). Given the huge variability among Chinese learners of English, research of English writing assessments needs to explore how Chinese learners as groups are affected by test variables as well as "the many individual factors related to background, experience and personality" (Hamp-Lyons & Kroll, 1997: 21).

References

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671–704.

Bachman, L. F. (2010). Foreword. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. x–xii). Routledge, Taylor & Francis Group.

Cortazzi, M., & Jin, L. (2011). Conclusions: What are we learning from research about Chinese learners? In L. Jin & M. Cortazzi (Eds.), *Researching Chinese learners: Skills, perceptions and intercultural adaptations* (pp. 314–319). Palgrave Macmillan.

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*, 73–83.

Dai, W., & Hu, W. (2009). *Research of the development of foreign language education in China (1949–2009)*. Shanghai Foreign Language Education Press.

Elman, B. A. (2009). *Berkshire Encyclopedia of China* (pp. 695–698). Berkshire Publishing Group LLC.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: principles for practice, theory and research*. Hampton Press.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 writing: Composition, community, and assessment* (TOEFL Monograph 5). Educational Testing Service.

Jin, Y. (2019). Testing tertiary-level English language learners: The College English Test in China. In L. I.-W. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 101–130). Routledge.

Jin, Y. (2020). Context validity in language assessment: Test operations and conditions for construct operationalization. In L. Taylor & N. Saville (Eds.), *Lessons and legacy: A tribute to Professor Cyril J Weir (1950–2018)* (pp. 83–104). Cambridge University Press.

Jin, Y., & Hamp-Lyons, L. (2015). A new test for China? Stages in the development of an assessment for professional purposes. *Assessment in Education: Principles, Policy & Practice, 22*(4), 397–426.

Jin, Y., & Yang, H. (2006). The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum, 19*(1), 21–36.
Li, X. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual & Multicultural Development, 11*(5), 393–404.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
Yang, H. (2003). The 15 years of the CET and its impact on teaching. *Journal of Foreign Languages, 3*, 21–29.
Yang, H., & Weir, C. J. (1998). *The validation study of the College English Test*. Shanghai Foreign Language Education Press.
Yu, G. (2014). Performance assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 615–630). Wiley.

Yan Jin is a professor of linguistics and applied linguistics at Shanghai Jiao Tong University, China. She started her career as a language tester in 1991 and has been involved in the development and research of large-scale, high-stakes language tests for three decades. She is currently Chair of the National College English Testing Committee of China. She is co-editor-in-chief of the Springer open-access journal *Language Testing in Asia* and is also on the editorial boards of *Language Testing, Language Assessment Quarterly* and a number of academic journals published in and outside China.