Liz Hamp-Lyons
Yan Jin *Editors*

Assessing the English Language Writing of Chinese Learners of English

Springer

Assessing the English Language Writing of Chinese Learners of English

Liz Hamp-Lyons • Yan Jin
Editors

Assessing the English Language Writing of Chinese Learners of English

Springer

Editors

Liz Hamp-Lyons
University of Bedfordshire
Luton, Bedfordshire, UK

Yan Jin
Haoran Hi-Tech Building, Room 2203
Shanghai Jiao Tong University
Shanghai, China

ISBN 978-3-030-92761-5 ISBN 978-3-030-92762-2 (eBook)
https://doi.org/10.1007/978-3-030-92762-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

In loving memory of Professor Liz Hamp-Lyons

On March 9, 2022, Professor Liz Hamp-Lyons, co-editor of this volume, passed away at her home in Cambridgeshire, England. Words cannot express the sorrow and sadness *we feel at the loss of Professor Hamp-Lyons, an accomplished scholar, a great editor, a supportive mentor, a dear colleague, and a true friend!*

Professor Hamp-Lyons had a deep understanding of the Chinese educational contexts through her service as Chair Professor of English and director of the Asian Centre for Language Assessment Research at Hong Kong Polytechnic University from 1996 to 2003. She also played an instrumental role in the assessment-for-learning initiative in Hong Kong.

Professor Hamp-Lyons had maintained a lifetime interest in writing assessment and EAP materials development and assessment, as reflected in her numerous publications and editorial experience. She was both editor of Assessing Writing *(2002–2017) and the founding editor and co-editor of the* Journal of English for Academic Purposes *(JEAP) from 2002 to 2016. In recognition of her extraordinary service to the field of EAP, the Liz Hamp-Lyons Award was established in 2015 to recognize quality work published in JEAP.*

Professor Hamp-Lyons had also made significant contributions to the professional community of language assessment through her supervision of PhD students, support of early-career researchers, and consultancy work for a wide range of projects on language test development and implementation, including the College English Test in China. When planning this edited volume, for example, Professor Hamp-Lyons purposefully invited some early-career researchers and provided strong support to the authors *throughout the process, which is again a testament to her professional rigor and strong support to young researchers.*

In memory of Professor Hamp-Lyons, we would like to dedicate this edited volume to her and her loved ones.

March 12, 2022

Foreword

English language assessment permeates every aspect of the examination-driven Chinese societies (including mainland China, territories of Hong Kong and Macau, and Taiwan) and every stage of education of Chinese learners of English as a foreign language. It is widely acknowledged that China is the origin of large-scale written examinations of individual differences in ability (Martin, 1870). The impact of the imperial examination on the current education and assessment systems in China, although it was abolished in 1905, is still prevalent and permanent. The global dominance of the use of the English language in business, education, and science has further intensified the scale and scope of English language assessment in China. The number of Chinese speakers taking English language tests, and the number of English language tests they must take for different purposes throughout their education and career, is phenomenal.

Research and publications on assessment of Chinese speakers learning English as a foreign language are emerging, but are still under-represented in the existing literature outside China, especially when the sheer volume of assessments of Chinese learners of English is considered. Recently, three journal special issues and three edited volumes on assessment of English language abilities of Chinese learners have been published. *English-as-a-foreign-language assessment in Taiwan* (Vongpumivitch, 2012) and *High-stakes English language testing in China* (Qian & Cumming, 2017) by Language Assessment Quarterly, *English Language Assessment in China: Policies, Practices and Impacts* (Yu & Jin, 2014) by Assessment in Education. Three edited volume: *English Language Assessment and the Chinese Learner* (Cheng & Curtis, 2010) by Routledge, *English Language Education and Assessment: Recent Developments in Hong Kong and the Chinese Mainland* (Coniam, 2014) by Springer (Note: the focus of this book is perhaps slightly more on "education" than "assessment"), and *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums* (Yu & Jin, 2016) by Palgrave. Together, these publications make incremental contributions to understanding the constructs and consequences of assessing the English language abilities of

Chinese learners of English. However, none of these journal special issues or edited books has a specific focus on the assessment of a certain language ability. This edited volume on the assessment of writing ability is therefore a much-welcomed addition to the increasing knowledgebase of assessment of Chinese learners of English as a foreign language.

In Preface, the two editors, Professors Liz Hamp-Lyons and Yan Jin, outline concisely and convincingly three reasons why this edited volume – *Assessing the English Language Writing of Chinese Learners of English* – makes a timely, unique, and important contribution to our better understanding of the assessment of Chinese learners of English, an argument with which I agree wholeheartedly. The very special history of written examinations in China has shaped the examination-driven nature of education in China. The active young scholars and teacher-researchers in China have contributed to this edited volume with their voices grounded in the realities of the Chinese educational and cultural contexts, and their interpretation of the research findings is informed by the insights they have gained from being a member of the community of practice. However, it is through the expert editing and the profound first-hand experience and expertise and the forward-thinking vision of the two editors in research and practice of writing assessment that the quality and coherence of this edited volume have been brought about.

Part I of the edited volume presents studies on four large-scale English language tests (College English Test, Pearson Test of English Academic, Aptis, and General English Proficiency Test). It covers a broad range of research topics, from the use of different types of rating scales (holistic vs. analytic), linguistic features of writings produced in response to independent and integrated writing tasks (read-to-summarise and listen-to-summarise), writers' awareness and construction of targeted audiences in their formal and informal email-writing, use of strategies in writing by successful and less successful test-takers and its impact on their writing performance, use of source materials by high-proficiency and low-proficiency test-takers in different types of integrated writing tasks (read-to-write, listen-to-write, and read-listen-to-write), to the construction of performance standards (in this case, China's Standards of English Language Ability, CSE) and the use of such standards for formative assessment of English language writing. While the assessment in the studies reported in Part I was summative in nature, studies reported in Part II focused on writing assessment which was more formative and low-stakes in nature, for example, providing corrective feedback in writing assessment, building teacher assessment literacy for portfolio assessment, validating a rating scale for a locally developed test, and validating a writing proficiency scale of business English based on CEFR and CSE.

The studies reported in this volume demonstrate an increasing maturity and diversity of research into writing assessment by younger Chinese scholars and teacher-researchers working in different assessment contexts and using different research methodology. As the editors write, there is "plenty to be celebrated, still a lot to do if we are to reach an agreed understanding of what 'best practice' in writing assessment means". They also note that there are, for example, "no peer or self-assessment, or writing collaboration, etc." reported in this volume. It would be unfair to suggest that a single edited volume should cover every writing assessment method or every aspect of writing assessment. However, what is missing in this edited volume perhaps suggests some potentially promising areas for research into writing assessment in China and elsewhere. For example, our research efforts could focus on how to promote and engage teachers as the key agents and meaning-makers of formative and summative writing assessment; how to research writing across borders to better understand the global impacts of the assessment of Chinese learners of English beyond the geographical boundaries; how to utilise and integrate technology as part of the construct of writing assessment as well as for test delivery and validation research; how to promote the effective use of automated, personalised evaluation feedback to learners; and most importantly, how to embrace and challenge in equal measure writing assessment as a social practice, in relation to, for example, the challenges and contrasts in educational equality and quality exacerbated by our very own act of assessing students' writing at different educational levels and contexts.

It is a great honour to write the foreword to this volume edited by the two academic veterans and visionaries of our field of language assessment. I would like to take this opportunity to personally thank them for their inspiration, for being role models, and for their leadership. This edited volume is another example of their continuous and tireless hard work to support younger scholars and teacher-researchers; it is a must-read for anyone interested in writing assessment.

Professor of Language Assessment,
University of Bristol, Bristol, UK
12/12/2020

Guoxing Yu PhD

Preface

This book focuses specifically on work in the assessment of English language in writing in mainland China, the territories of Hong Kong and Macau, and Taiwan. It is unique in its focus on work done by Chinese researchers and practitioners in and from the Chinese-speaking world, aiming to ensure that the voices we hear in these texts are grounded in the realities of Chinese educational and cultural contexts. The editors combine strong familiarity with expertise in language assessment in China and strong background in writing assessment:

Professor Yan Jin, of Shanghai Jiao Tong University in Shanghai, is a leading scholar in English language assessment in China, and chair of the National College English Testing Committee of China, in charge of developing a test which assesses [among other components] the writing skills of approximately 20 million undergraduates in Chinese universities every year. She is also the founding co-editor of the Springer Open Access journal *Language Testing in Asia*.

Professor Liz Hamp-Lyons' long career include 15 years in Hong Kong universities, 2 years at the University Sains Malaysia, 1 year at the University of Melbourne, and 12 years in US universities. She has been senior consultant to the CET since 2006, and since 2008, she has been associated with the Centre for Research on English Language Learning and Assessment at the University of Bedfordshire, where she is now a visiting professor. She has been researching and publishing about writing assessment since 1990.

Readers coming to this book may ask: ***why a book specifically about the writing of Chinese learners of English***?

The first reason is the sheer number of first-language Chinese speakers using, and learning, English and learning to write in and through English. Many of these young people are learning English not for social reasons, but because of its importance in international business, industry and politics, and especially for its dominance in research publications. As Montgomery (2013) has pointed out, 'native speakers of English are a small, shrinking minority, outnumbered by non-native speakers 4 to 1'

(p. 46). In mainland China's 2688 'regular' universities, there were more than 30 million undergraduate students in 2019, and approaching 3 million postgraduate students. There are also approximately 80,000 undergraduate students in Hong Kong's universities who study principally through the medium of English. In Macao, the University of Macao has more than 10,000 students, and English is the medium of instruction in most disciplines. In Taiwan, however, the picture is more blurred because the vast majority of high school graduates enter universities: an average of just 8% of courses at each university uses English as the language of instruction. Additionally, Chinese L1 speakers are studying a whole range of subjects and courses in many parts of the world, notably in native English-speaking countries, that is, Britain (120,000 in 2018), the USA (370,000 in 2018), Canada (140,000 in 2018), Australia (160,000 in 2018), New Zealand (111,000 in 2018), and most often their coursework is delivered through the medium of English. Taken together, these numbers of Chinese learners of English as a foreign language and learners of English as an academic language in international contexts indicate a great need for teaching and learning materials, but also for principles, methods and practice in language assessment.

The second reason why this book is needed is the very special history of testing in China. China can claim by far the longest history of examination-driven education in the world. Experts differ as to the exact moment at which education in China developed an assessment structure formal and stable enough to be referred to as a national system. Teng (1943) cited the 11th and 14th editions of the Encyclopaedia Britannica to claim that some form of assessment through writing can be traced as far back as the Chou period (1111–771 BCE) in early China; he also describes his exhaustive process of reading every extant text (English and Chinese) where one might expect references to examinations. He found none in Greek, Egyptian or Roman history, or in mediaeval or modern history until the nineteenth century. However, Elman (1991), deepening the studies of the early examination period with the benefit of modern travel and technology's access to texts and establishing more specific definitions, describes the 'simple recruitment process for the Han Imperial Academy during the Han period (206 B.C.E. – 220 C.E.)'. At that time, disciples of masters who were learned in the Five Classics studied to become expert in one classical text, and were examined orally on it, and granted government positions if they were successful (Elman p. 9). But Rui Wang (2012) dates the formal beginning of the Imperial examinations to 605 CE, late in the short-lived Sui period, and the highly competitive civil service examinations with their rigid and ritualized structures, and their processes for ensuring that there would be no cheating appear to have been taking on full shape by 622 CE, and earlier in some provinces and cities.

This long history does much to explain the test-driven nature of education in China and other Confucian-heritage countries generally, and especially its long-lasting and powerful imperial examinations system (Cheng, 1998; DuBois, 1964; Miyazaki, 1976; Suen, n.d.; Yu & Suen, 2005). It is unsurprising that when, much more recently, the concept of relatively large-scale formal examinations spread beyond China to Europe, in Britain, their earliest proponents, such as Lord Macaulay and John Stuart, proposed a process by which men presenting themselves as candidates for civil service posts would receive written questions and write answers to be judged by appointed judges (VanWaarden, 2015). The earliest formal examination of English language proficiency, the Cambridge Certificate of English (CPE), was first introduced in 1913 at the University of Cambridge for use by 'Foreign Students who desire a satisfactory proof of their knowledge of the language with a view to teaching it in foreign schools' (UCLES, 1913: 5). CPE in its beginnings comprised (1) translation into French or German according to personal choice; (2) translation from French or German according to personal choice, plus some 'grammar' items which also embrace style; phonetics transcription, explaining some pronunciation terms, describing the articulation of vowels, and teaching (i.e., explaining how to teach the pronunciation of words with closely similar sounds); and (3) an English essay (2 h). In 1913, the essay subject choices were:

(a) The effect of political movements upon nineteenth-century literature in England
(b) English pre-Rafaellitism
(c) Elizabethan travel and discovery
(d) The Indian Mutiny
(e) The development of local self-government
(f) Matthew Arnold

There was no advice for the candidates (within the exam paper or test room) and no evidence that there was any kind of check on essay marking in terms of fairness or reliability (see Weir et al., 2013).

Happily, English language examinations have moved far since 1913: the examining/assessing of writing in English has, however, proceeded more slowly.

This brings us to a third reason why we feel this book is needed, which is: the authors themselves. The researchers of these studies are all themselves Chinese speakers as well as active young scholars and teacher-researchers, who know their students/test-takers very well, and therefore are likely to bring fresh insights/perspectives into Chinese learners' English language writing. The authors have been able to focus on a homogeneous group of learners (Chinese L1 learners of English) so that the studies in the book are free from the interference of the L1 variable, which is often an issue with studies done by researchers outside immediate Chinese-L1 contexts, which typically use whole classes or cohorts of learners where L1s are not a key variable. The majority of writing assessment studies published in international journals have been conducted in the USA (Zheng & Yu, 2019), a context where studies in this field most often use whole classes/cohorts of 'freshman English' college students, and where individual classes often comprise a mix of native speakers of English judged to have inadequate writing levels as well as significant numbers of non-native English speakers from a random assortment of countries, and the data is frequently not disaggregated by home language. The focus in this book on a single language, Chinese, provides a clearer picture of the state of research and teaching as well as language learning in this single language group.

We hope this book will provide a valuable source of information for researchers, teachers and curriculum designers in high schools and colleges where Chinese is the

L1 and English is a language used in instruction and where it is assessed. Apart from writing and writing assessment scholars, we hope it may also be useful to researchers, practitioners and students of language assessment more broadly who have an interest in assessment and education issues in China and other Asian contexts, as well to anyone who wants to consider the influences on writing assessments outside their own context and experience.

The chapters in the book are organized into two parts. The first part looks at how writing in English has been assessed in the past 25 years or so, and how it continues, in the main, to be assessed at present. This part is introduced by Professor Yan Jin, and reports a number of studies of large-scale writing assessments, including the impact of rating scales on the CET-4 Writing by Shaoyan Zou, the linguistic features of Chinese learners' English language writing in the PTE Academic by Yuhua Chen and Ying Zheng, Chinese learners' conceptualization of audiences in Aptis-General Writing by Ying Chen and Xiaoxian Guan, and the effectiveness of strategy use on performances in the GEPT Writing by Naihsin Li. The part also addresses issues involved in the use of source materials in integrated writing tasks by Yan Zhou and Ke Bin and the use of performance standards for the teaching, learning and assessment of English language writing by Mingwei Pan. The second part of the book explores directions for the future of English writing assessment in Chinese contexts, and is introduced by Professor Liz Hamp-Lyons. It includes a proposed framework for the assessment of academic writing by Cecilia Zhao, a report of a study to develop a rating scale for a specific English-learning purpose by Li Wang, and an attempt to build a validity argument in line with current validity theory by Li Liu and Guodong Jia. The final three chapters discuss two views of feedback on writing, the first by Jing Yang and the second by Icy Lee, Na Luo and Pauline Mak, and finally, a report on a project to build assessment literacy for portfolio assessment of writing by Ricky Lam.

Luton, Bedfordshire, UK Liz Hamp-Lyons
Shanghai, China Yan Jin

References

Cheng, L.-y. (1998). Impact of a Public English Examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24*(3), 279–301.

Cheng, L., & Curtis, A. (2010). *English language assessment and the Chinese learner*. New York: Routledge.

Coniam, D. (Ed.). (2014). *English language education and assessment: Recent developments in Hong Kong and the Chinese mainland*. Singapore: Springer.

DuBois, P. H. (1964). *A test-dominated society: China, 1115B.C.–1905A.D*. Invitational conference on testing problems, Educational Testing Services, Princeton.

Elman, B. (1991). Political, social and cultural reproduction via civil service examinations in late Imperial China. *Journal of Asian Studies, 50*(1), 7–28.

Martin, W. A. P. (1870). Competitive Examinations in China. *The North American Review*, 111 (228), 62–77. https://doi.org/10.2307/25109555

Miyazaki, I. (1976). (Translated by Conrad Schirokauer) *China's examination hell: The civil service examinations of imperial China.* Yale University Press.

Montgomery, S. (2013). *Does science need a global language? English and the future of research*. University of Chicago Press.

Qian, David D., & Cumming, A. (2017) Researching English Language Assessment in China: Focusing on High-Stakes Testing, *Language Assessment Quarterly*, 14 (2), 97–100, https://doi.org/10.1080/15434303.2017.1295969

Suen, H. K. (n.d.). The hidden cost of education fever: Consequences of the Keju-driven education fever in ancient China. In J. G. Lee (Ed.), *Education fever*. Kangwon National University Press. Available at http://suen.ed.psu.edu

Teng, S.-y. (1943). Chinese influence on the Western examination system: I. Introduction. *Harvard Journal of Asiatic Studies, 7*(4, September), 267–312.

UCLES. (1913). Annual Report, University of Cambridge Local Examinations Syndicate.

VanWaarden, B. (2015). John Stuart Mill on civil service recruitment and the relation between bureaucracy and democracy. *Canadian Journal of Political Science*, *48*(3), 625–645. Downloaded from ResearchGate 28-08-20.

Vongpumivitch, V. (2012). English-as-a-Foreign-Language Assessment in Taiwan. *Language Assessment Quarterly*, 9(1), 1–10. https://doi.org/10.1080/15434303.2012.649592

Wang, R. (2012). *The Chinese Imperial Examination System: An annotated bibliography*. Scarecrow Press

Weir, C., Vidakoviḉ, I., & Galaczi, E. (2013). *Measured constructs.* Cambridge University Press.

Yu, L., & Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy, 2*(1), 17–33.

Yu, G., & Jin, Y. (2014). English language assessment in China: policies, practices and impacts. *Assessment in Education: Principles, Policy & Practice*, 21(3), 245–250. https://doi.org/10.1080/0969594x.2014.937936

Yu, G., & Jin, Y. (Eds.). (2016). *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums*. Basingstoke: Palgrave.

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from Assessing Writing (2000–2018). Assessing Writing, 42, 100421. https://doi.org/10.1016/j.asw.2019.100421

Acknowledgements

We editors must thank the wonderful professional colleagues who have helped us, and our authors, by reviewing the draft chapters:

Nadia Behizadeh	Georgia State University
Emma Bruce	CRELLA, University of Bedfordshire
Tineke Brunfaut	University of Lancaster
Yeonsuk Cho	Educational Testing Service
Bill Condon	Washington State University (emeritus)
Deborah Crusan	Wright State University
Alister Cumming	University of Toronto (emeritus)
Tony Green	CRELLA, University of Bedfordshire
Jane Lockwood	FuturePerfect
Judy Parr	University of Auckland
Barry Reynolds	University of Macau
Sara Weigle	Georgia State University

We are also very grateful to the anonymous reviewers who provided very valuable feedback on the book as a whole and on its parts. We would also like to thank Ms Linli Chen, lecturer in the School of Foreign Languages, Shanghai Jiao Tong University, for helping us with manuscript preparation.

Finally, we would like to thank our authors for responding to our invitation to take part in this project. It was a pleasure working with them. We hope our readers will agree that the book as a whole provides a unique insight into a specific moment in time in research into English as a second language, as perceived, interpreted and taken forward by Chinese-background assessment scholars.

Contents

Part I Large-Scale English Writing Assessment for Chinese Learners of English

Part II Research and Practice of English Language Writing Assessment in China

About the Editors

Liz Hamp-Lyons began her career as a teacher of English for Academic Purposes, before researching the writing test component of ELTS (the precursor to the IELTS) for her PhD at the University of Edinburgh. Much of her PhD was published, with other chapters, in *Assessing Second Language Writing in Academic Contexts* (1991). She worked at the Universities of Michigan and Colorado before becoming Chair Professor and Department Head at the Hong Kong Polytechnic University. For many years she edited the journal *Assessing* Writing (Elsevier), and founded the *Journal of English for Academic Purposes* in 2002. She has consulted on large scale as well as more local writing assessments.

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 and President of the Asian Association for Language Assessment. 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.

Abbreviations

BE	Business English
CEFR	Common European Framework of Reference for Languages
CET	College English Test
CET-4	College English Test Band 4
CLA	Communicative Language Ability
CSE	China’s Standards of English Language Ability
CSE-W	China’s Standards of English Language Ability – Writing scales
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EGP	English for General Purposes
EMI	English as Medium of Instruction
ERPs	Event-Related Potentials
ESP	English for Specific Purposes
GEPT	General English Proficiency Test
GSEEE	Graduate School Entrance English Examination
HE	Higher Education
L1	First Language
L2	Second Language
MFRM	Many-Faceted Rasch Model
MnSq	Mean Square
NMET	National Matriculation English Test
PTE Academic	Pearson Test of English Academic
SEM	Structural Equation Modelling
SWPE	Standards of Writing Proficiency in English
TAPs	Think-Aloud Protocols
TEEP	Test in English for Educational Purposes
TEM	Test for English Majors
TEM-4	Test for English Majors Band 4
TWPE	Test for Writing Proficiency in English
WCF	Written Corrective Feedback

List of Figures

List of Tables

Part I
Large-Scale English Writing Assessment for Chinese Learners of English

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.

Chapter 2
The Impact of Rating Scales on the CET-4 Writing: A Mixed Methods Study

Shaoyan Zou

Abstract Holistic rating scale and analytic rating scale are two types of scales which are frequently utilized by raters and EFL teachers in scoring EFL writing performance. Given their significant differences in terms of the underlying assumptions about writing constructs and development, as well as the implications for essay rating processes, whether one type is superior to the other remains a contentious issue. However, to date only a few studies have been conducted to empirically compare the effectiveness of the two types of scales in rating EFL writing performance. Therefore, this study was undertaken to examine the impact of two rating scales, one holistic and the other analytic, on the rating of writing scripts produced in College English Test Band 4 (CET-4) by adopting the Mixed Methods approach. Results of the study indicated that the analytic scale was as effective as the holistic one in terms of scale discrimination, rating reliability and scale level functionality; However, the analytic scale lent itself more easily to the control of rating variations. The raters in general held more positive views toward the analytic scale although suggestions were also proposed to conduct minor revisions to the scale descriptions. In conclusion, findings of the study revealed the potential of analytic rating scales in large-scale EFL writing assessments and also generated implications for rating scale development and validation in other research contexts.

Keywords Rating scales · CET writing · Mixed methods research

2.1 Introduction

In recent decades, spurred by test users' increasing demand for the interpretability of test scores (Chapelle et al., 2008), holistic rating scales, especially those commonly used in large-scale EFL writing assessments have aroused growing concerns (e.g. Knoch, 2009; Lee et al., 2010). The major concerns are that these holistic

S. Zou (✉)
Qingdao Agricultural University, Qingdao, China
e-mail: amandazsy@qau.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_2

scales are often designed intuitively, hence may not closely represent the features of the corresponding writing performance for one thing, and for another, the scoring criteria adopted in such holistic scales usually rely on impressionistic terminology which may be open to subjective interpretations (Knoch, 2009).

As such, analytic rating scales have gained renewed momentum particularly in the domain of EFL writing assessment. Compared with holistic rating scales, analytic rating scales are more advantageous in capturing examinees' specific weaknesses and strengths in writing, which is especially useful for second language learners whose writing skills are still under development (Lee et al., 2010). Therefore, the recent years has witnessed increasing research attempts to develop and validate analytic rating scales for EFL writing tests (e.g., Berger, 2015; Knoch, 2009; Lee et al., 2010).

Inspired by such a research trend, an analytic rating scale was developed for the writing assessment of College English Test Band Four (the CET-4 writing), one of the largest-scale EFL tests in China. Specifically, the scale is expected to adequately address the concerns over the holistic rating scale currently used in the CET-4 writing (e.g., Fei & Zhao, 2008; Jian & Lu, 2005). Featuring a multi-method approach, the scale developing process took advantage of document analysis, intuitive judgment and Rasch model analysis in pinpointing useful rating categories, providing adequate descriptors, and calibrating difficulty level of the descriptors. The analytic rating scale developed in this study is comprised of five levels and four rating categories (see Table 2.1 for a demo of the scale).

However, despite the endeavors involved in the scale developing stage, whether this analytic scale will be robust in scoring the CET-4 writing scripts or not still needs further scrutiny. Therefore, this study undertakes to examine the effectiveness of the analytic scale and the existing holistic scale. The aim is two-fold: one is to gather validity evidence for the newly-developed analytic scale, and to identify potential problems that jeopardize the scale validity, and the other is to enrich the meager literature concerning empirical comparison of the two types of rating scales, thereby exploring the prospect of analytic rating scale in large-scale EFL writing assessments.

2.2 Literature Review

2.2.1 A Review of Theoretical Arguments

According to Weigle (2002), holistic scoring involves assigning a single score to a script based on the rater's overall impression of the whole script. Whereas, analytic scoring takes into account a list of elements, each of which is judged separately and then assigned a single score (Hamp-Lyons, 2016).

Over the past decades, holistic scales have been widely used in writing assessment due to a number of advantages (see Shaw, 2004; Shaw & Weir, 2007; Weigle, 2002; White, 1984, 1985): First, they are highly efficient and practical in terms of

Table 2.1 A demo of the analytic rating scale

	Linguistic range and control	Content and idea	Discourse organization	Linguistic appropriacy
Level A	Can use a sufficient range of vocabulary to convey information and ideas effectively. Can use some fairly advanced words or phrases as well as common idioms, proverbs, sayings or expressions to convey meaning more precisely or to enhance the expressing effect. Both the choice of words and the use of collocations are accurate and idiomatic, thus making the expression clear and distinct. Can effectively and accurately use a variety of sentence patterns, syntactic or grammatical structures.	The content is closely related to the topic and the opinions are expressed in a clear and distinct manner. Can explain or illustrate issues and ideas with adequate examples or in proper ways. Can effectively organize and use facts and details to support one's arguments. Can produce clear, well-organized and well-developed text, demonstrating a sufficient understanding of the topic.	Can appropriately and flexibly use a wide variety of cohesive devices to indicate the logical relations between sentences and to make the text clear and well-organized. Can maintain the cohesion and consistency of the content through echoing the points mentioned before. Good paragraph structures, with clearly expressed main points and adequate supporting details. Has an adequate control of the connections between paragraphs as well as the logical relations between parts and the whole.	Can express appropriately and idiomatically in written language based on specific contexts. Has certain awareness of the audience, and the language style and register (formal or informal) are rather appropriate. Can express feelings or attitudes in a manner that is appropriate to the context. Can idiomatically use cultural references and figures of speech to convey information effectively.

time and cost. Second, they are useful for discriminating across a narrow range of assessment bands and suitable for arriving at a rapid overall rating, especially for large-scale assessments, thus enhancing rating reliability. Third, they focus the rater's attention on the strengths of the writing rather than the weaknesses so that writers are rewarded for what is well done instead of what is underachieved. However, holistic scales have been criticized for being devoid of any real theoretical underpinning, thus leading some researchers to challenge their validity (Shaw & Weir, 2007). With multiple categories collapsed into a single score, the same score from different raters may reflect vastly different constructs because raters, more often than not, bear their own rating criteria in mind during the rating process. For example, some raters pay more attention to grammatical accuracy, while others attach great importance to syntactic complexity. Another significant drawback of holistic scales lies in their inability to capture examinees' specific weaknesses and strengths in writing. According to Lee et al. (2010), this drawback can be even more conspicuous for EFL learners whose writing skills are still under development and who are more likely to show uneven profiles across different aspects of writing.

In contrast to holistic scales, the primary advantage of an analytic rating scale is that it can provide useful diagnostic information about a test taker's performance, especially his/her literacy progress (Hamp-Lyons, 1986, 1991; Shaw & Weir, 2007; Weigle, 2002). Moreover, analytic scales are considered more helpful in rater training as inexperienced raters can understand and apply rating criteria more easily (Weigle, 2002; Weir, 1990). Notwithstanding all the advantages, there are also some concerns over analytic rating scales, such as the high cost and the 'halo effect' problem associated with the use of such scales (Weigle, 2002), thus posing questions for the effectiveness of analytic scales in essay rating.

2.2.2 A Review of Empirical Research

Despite the seemingly abundant theoretical discussion, only a few empirical studies have been conducted to examine the effectiveness of the two types of scales. Moreover, findings of the existing research are a little mixed. On one hand, some studies found that the use of holistic scales could result in a higher generalizability coefficient and dependability coefficient, thus contributing to satisfactory inter-rater reliability (Barkaoui, 2007; O'Loughlin, 1994). For instance, in Barkaoui's (2007) study, four experienced English teachers scored EFL learners' writing scripts using holistic scale and analytic scale alternatively. The results showed that holistic rating scale contributed to higher rating reliability, while analytic rating scale caused greater variation between raters, which could be addressed only through increasing the number of writing scripts.

On the other hand, some research discovered that analytic rating scales could lead to higher rating reliability due to their advantages in making finer and more accurate distinctions between test takers' performances (Barkaoui, 2008; Li, 2014; Li, 2015; Song & Caruso, 1996; Sun & Han, 2013; Wiseman, 2008). For instance, Song and Caruso (1996), after scrutinizing holistic and analytic scores assigned by English and ESL teachers, found significant differences in the holistic scores, but not the analytic ones. They concluded that raters' teaching and rating experience might have an impact on their holistic scores, however, these factors didn't seem to affect analytic scoring as analytic scales focus raters' attention on the same aspects of the essays, thus mitigating the effects of rater-related variables. Li (2014) also compared two sets of data elicited through holistic scoring and analytic scoring of the writing scripts produced in Test for English Majors Band-4 (TEM-4) separately. He found that analytic rating scale was more advantageous than holistic scale in terms of scale discrimination, inter- and intra-rater reliability, and interaction bias between raters and examinees.

In addition to the above two lines of conclusions, some researchers argue that there's no significant differences between holistic rating scale and analytic rating scale in terms of their rating reliability (e.g., Bacha, 2001). They can both lead to higher agreement within and between raters, although analytic scales can generate more diagnostic information for EFL learners' writing proficiency.

Considering the ongoing arguments on the two types of scales and the mixed findings yielded from existing research, the present study is undertaken to re-examine the impact of two rating scales, one holistic and the other analytic, on the CET-4 writing. Specifically, the study is intended to address the following questions:

1. To what extent do the two scales differ in terms of discriminating power, rater reliability, rating variation and scale level functionality?
2. To what extent can the rating categories on the analytic scale function as intended?
3. How do raters of the CET-4 writing perceive the effectiveness of the two scales?

2.3 Research Design

To solve the research questions, a convergent parallel design of the mixed-methods approach was adopted. According to Creswell and Clark (2017), the advantage of this type of design lies in that the researcher can collect and analyze both quantitative and qualitative data during the same research phase, and the two sets of data can then be merged to provide an overall interpretation of the research questions.

2.3.1 Research Instruments

2.3.1.1 Instruments of the Rating Experiment

The first instrument adopted in the rating experiment is the existing holistic rating scale for the CET-4 writing which consists of five band levels and three rating categories: content relevance, language quality and discourse coherence. As with the operational rating of the CET-4 writing, the holistic scale is presented in Chinese to facilitate the raters' interpretation of the scale. The second instrument is the draft version of the analytic scale. At this stage, the scale is comprised of four sub-scales: *Linguistic Range and Accuracy*, *Content and Idea*, *Discourse Organization*, and *Linguistic Appropriacy*, with each substantiated by detailed level descriptions. As with the holistic scale, the analytic scale also entails five band levels.

In addition, 30 writing scripts produced in the operational CET-4 in June 2011 and June 2016 respectively were authorized to be utilized by the CET committee. These scripts which had been used as benchmark essays during the operational rating process address two different topics, with half of them entitled *Online Shopping* and the other half involving writing a *Thank-you* letter.

2.3.1.2 Instruments of the Interview

The instrument adopted in the interview was a semi-structured interview guideline which was designed based on the scale usefulness framework proposed by Knoch (2009). The guideline mainly deals with the raters' perceptions of the two scales, including scale clarity (perceived validity), scale completeness (perceived validity), scale operability (practicality), feedback for teaching (impact), correspondence with the CET-4 writing performance (authenticity), as well as its efficiency in rater training (perceived reliability).

2.3.2 Participants

21 raters selected from various CET-4 marking centers in China took part in the rating experiment. Among them, 15 were females and 6 males. 19 of them had more than 5 years of teaching experience and had been involved in the operational marking of CET-4 writing for more than 3 times. 6 of the raters held a PhD degree while the others held a master degree.

For the semi-structured interviews, 10 participants (4 males and 6 females) were invited. All of them were experienced raters of the CET-4 writing and several of them were rating experts appointed by the CET committee.

2.3.3 Data Collection

The rating experiment began in early January, 2017 and lasted for approximately one and a half month. To control the potential order effects caused by the use of the two rating scales, a counterbalanced design was adopted. Specifically, the participants were divided into two groups with 10 in the first group and 11 in the other. The rating procedures were as follows: in the first session, Group1 started to rate the 30 CET-4 writing scripts using analytic scale while Group 2 accomplished the same assignments by using the existing rating scale; Whereas, in the second session, the two groups utilized the two scales in reverse order.

Following the rating experiment, the interviews were conducted one on one through WeChat, a popular communication software in China. The raters were encouraged to comment on the analytic scale and further, to make suggestions for the potential improvement of the scale. All the interviews were live recorded by a digital voice recorder.

2.3.4 Data Analysis

2.3.4.1 Quantitative Analysis

The rating data were submitted to Many-faceted Rasch Model (MFRM) analysis using the computer program of FACETS version 3.80.0 (Linacre, 2013). MFRM can be used to calibrate a number of facets involved in the rating onto a common logit scale, thus allowing for a direct comparison of the two scales. Specifically, indices like examinee separation, rater separation, scale level functionality and rating category functionality were closely looked at. According to Fisher (1992), the examinee separation ratio is conventionally used as an effective indicator of the discrimination of the rating scale because it measures the spread of examinees' proficiency relative to their precision. Meanwhile, rater separation demonstrates how the raters as a group differ in terms of their severity or leniency. The scale level response structure and category response structure are indices indicating whether all the scale levels and scale categories can be used as intended in the rating process. Of particular interest to this study were infit and outfit mean squares (MnSq) which revealed variations in rating and assessed global model fit.

2.3.4.2 Qualitative Analysis

Recordings of the interviews were transcribed and coded. A coding scheme was developed *a priori* by taking into account the aspects involved in the interviews. The interview data were then classified according to the themes. It is worth mentioning that to ensure the accuracy and appropriateness of the coding schemes, a second qualified researcher was invited for a double check.

2.4 Results and Discussion

2.4.1 Effectiveness of the Two Scales

To facilitate comparison of the two scales, the key statistics resulting from the two MFRM analyses are summarized in Table 2.2.

Examinee Discrimination

As can be seen from Table 2.2, the examinee separation ratio for the holistic rating session is 14.88, indicating that the measures of examinee proficiency are statistically different, and the examinee separation indices by the two scales are quite similar, although the one yielded by the analytic scale is a little higher. Meanwhile, the separation reliability estimates of the two scales are also very close to each other, with the one by the analytic scale being slightly higher.

Table 2.2 Key statistics for the two rating scales

Qualities	Indices	The existing rating scale	The analytic scale
Examinee discrimination	Examinee separation ratio	14.88	14.99
	Separation reliability	.98	1.00
Rater separation	Rater separation ratio	5.18	4.11
	Separation reliability	.87	.94
Variation in ratings: 0.7~1.3	% Unexpected responses	0.8%	0.6%
	% Rater misfit	9.5%	4.8%
	% Rater overfit	28.6%	9.5%
Scale properties	Level functionality	Well-spread	Well-spread

According to Knoch (2009), a higher examinee separation ratio indicates a more discriminating rating scale since more levels on the rating scale are attended to in the rating process. Following this logic, we can conclude that the two scales are both effective in discriminating the examinee proficiency levels, and the analytic scale is slightly more powerful than the holistic scale.

Rater Separation and Reliability

As shown in Table 2.2, the rater separation ratio resulted from the analytic rating process is 4.11 which is moderately lower than the one by the holistic rating session (5.18). According to Knoch (2009), "a well-functioning rating scale would result in small differences between raters in terms of their leniency and harshness as a group" (p. 205). Eckes (2011) also echoed this point by stating that the rater separation ratio close to 1 would be ideal because all raters would form a single, homogeneous class. Following these criteria, it is fair to say that the raters as a group were more similar in terms of their severity in using the analytic scale than in using the holistic scale.

Variation in Ratings

In assessing global model fit, the first index to be examined is the standard residual. As Table 2.2 shows, only 0.8% of standard residuals caused by the holistic rating session are higher than 3, indicating a satisfactory model fit. Likewise, in the analytic rating session, the standard residuals greater than 3 merely take up 0.6%, suggesting a good model fit as well. However, as Eckes (2011) noted, "such a result does not preclude that specific parts of the measurement system exhibit deviations from model expectations" (p. 58). Therefore, a closer look at the rater statistics is necessary so as to uncover potential variation in ratings.

Considering that the primary intention of the MFRM was to stringently compare the effectiveness of the two scales in the CET-4 writing and to inform further scale revision where necessary, a more severe control range of 0.7~1.3 (Bond & Fox, 2015; McNamara, 1996; Wright & Linacre, 1994) is adopted when examining the rater fit statistics. Following this criterion, raters with fit values exceeding 1.3 are considered misfitting because they assign ratings too inconsistently and show more variation than the model would predict; Whereas, raters with fit values less than 0.7 are overfitting as their ratings exhibit less variation than expected. Knoch (2009) holds that a well-functioning rating scale would result in fewer raters who rate either inconsistently or unduly consistently. According to Knoch, an overfit can be attributed to two possible reasons: one reason is that the raters were rating too consistently, the other involves the raters' overuse of the inner levels (i.e. Levels 2, 3, 4) of a rating scale in order to play safe. As is shown by Table 2.2, in the holistic rating session, there are 9.5% raters with fit values greater than 1.3, suggesting that their ratings display too much variation than expected. Meanwhile, 28.6% raters are overfitting with fit values less than 0.7, indicating that their ratings exhibit less variation than intended.

By contrast, when using the analytic scale, the raters' fit statistics are much more satisfactory. Only one rater shows more variation than predicted, and 9.5% raters are lack of supposed variation. Taking these findings together, it can be concluded that the analytic scale can lend itself more easily than the holistic scale to the control of rating variation.

Scale Level Functionality

The first commonly used indicator of the functionality of the scale levels is the average measure of each scale level, which represents the average of the examinee measure modeled to generate the observations in a given level (Eckes, 2011). The underlying assumption is that average measures should progress monotonically with the increase of scale levels. As Table 2.3 shows, the observations of the five levels on the holistic scale range from 83 to 169, all of which greatly exceed the minimum requirements of 10. Meanwhile, the average measures of each level increase monotonically from −6.51 logits to 5.41 logits, indicating that the five levels on the holistic scale are used reasonably by the raters.

Another indicator of scale level functionality is the outfit MnSq value for each level which compares the average examinee measure with the expected examinee measure that the Rasch model would predict for each scale level. According to Bond and Fox (2015), this outfit MnSq value should not exceed 2. As is shown in Table 2.3, the outfit MnSq values of the holistic scale are all around the expected value of 1. Finally, the level thresholds progress clearly from −5.27 logits to 4.66 logits. All of the threshold calibrations stay well within the expected range of 1.4~5.0 logits (Bond & Fox, 2015). Based on these findings, it is fair to say that the levels on the existing rating scale are properly ordered and generally functioned as intended.

Table 2.3 Level statistics for the existing rating scale

Level	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1	85 (14)	−6.51	1.0	NONE
2	140 (23)	−3.16	.9	−5.27
3	169 (28)	.56	.9	−1.39
4	132 (22)	3.01	.9	1.99
5	83 (14)	5.41	1.1	4.66

The same set of statistics for the analytic scale are listed in Table 2.4. As the table shows, the observations for each of the five levels on the analytic scale are all well above 10, suggesting that the five levels on this scale are also reasonably attended to by the raters. Besides, the average measures increase from −4.66 logits to 4.10 logits as the scale levels move up. The outfit MnSq values are either equal or very close to the expected value of 1. Finally, the level thresholds advance monotonically with the increase of the levels. Specifically, there is a clear progression from −3.93 logits to 4.27 logits. Given these findings, we can conclude that the five levels on the analytic scale can generally be used as intended.

To sum up, the results as presented above offer answers to the first research question: in general, the two scales are both effective in terms of scale discrimination, rating reliability and scale level functionality; However, the analytic scale lends itself more easily to the control of rating variation as the raters' behaviors are more consistent both at group level and at individual level.

2.4.2 Category Functionality of the Analytic Scale

According to Eckes (2011), the analysis of the category facet is able to generate insight into the relative difficulty of each rating category and to test the assumption that these categories can work together to define a single latent dimension. The statistics relating to the category utility of the analytic scale are presented in Table 2.5.

Table 2.4 Level statistics for the analytic scale

Level	Observed count (%)	Average measure	Outfit MnSq	Threshold calibration
1	361 (14%)	−4.66	1.0	NONE
2	517 (21%)	−2.46	1.1	−3.93
3	714 (28%)	−.02	1.0	−1.53
4	673 (27%)	2.34	1.0	1.20
5	254 (10%)	4.10	1.1	4.27

Table 2.5 Category measurement report

			Infit		Outfit		Correlation
Category	Measure	S.E.	MnSq	Std	MnSq	Std	
Linguistic appropriacy	.74	.07	.85	−2.6	.92	−1.1	.89
Linguistic range and accuracy	.20	.07	.84	−3.0	.86	−2.4	.90
Discourse organization	−.30	.07	1.12	2.0	1.13	2.1	.87
Content and idea	−.65	.07	1.09	1.5	1.13	2.2	.86
Mean	.00	.07	.97	−.5	1.02	.3	.88
S.D.	.61	.00	.09	1.7	.08	1.3	.01

As is shown by the table, the four categories differ significantly from each other in terms of their difficulty measures, with *Linguistic appropriacy* being the most difficult one (0.74 logits), while *Content and idea* being the easiest (−0.65 logits). This result implies that it is relatively more difficult for examinees to gain a higher score on the category of *Linguistic appropriacy* than on *Content and idea.* Meanwhile, the mean square fit indices of the four categories all stay within the narrow range of 0.7~1.3, which indicates satisfactory data-model fit. Besides, the correlation estimates of the four rating categories range from 0.86 to 0.90, all of which are well above 0.3, suggesting that the four rating categories can work together to measure one single latent construct.

In all, these results confirm the robustness of the four rating categories. That is to say, they could function ideally both as an individual and as a whole in discriminating the CET-4 writing performance. The results could offer adequate answers to the second research question.

The findings elicited thus far through quantitative analysis enabled a straightforward comparison of the effectiveness of the two scales in the CET-4 writing. However, to gain a more comprehensive and more in-depth understanding of the two scales, a closer look at the interview data should be essential.

2.4.3 Raters' Perceptions of the Two Scales

The raters' perceptions of the two scales were transcribed and then classified based on the themes entailed in the interview guide.

Overall Effectiveness

Almost all the raters expressed positive views on the overall effectiveness of the analytic scale. Rater 10 believed that compared with the existing rating scale, the analytic scale made the scoring of the CET-4 writing performance more reliable and more evidence-based. Also, the raters considered the analytic scale 'more detailed and more specific' (R1), thus 'it gives more useful guidance in rating' (R3) and 'it makes the rating well-grounded and more objective' (R2). To quote R2:

> In using the existing rating scale, I often hesitated in awarding scores for fear that some important information might be ignored. However, with the analytic scale, I feel more confident because everything is clearly set out in the scale.

There was only one rater (R7) who was slightly concerned over the effectiveness of the analytic scale, holding that:

> Although the analytic scale looks reasonably good, it may not be practical to rely solely on the scale.
>
> In the operational rating of CET-4, benchmark essays played a crucial role which sometimes even outweighed the role of the rating scale. However, it should be admitted that in the operational rating process, interpretations of the benchmark essays often varied from person to person, which to some extent would compromise the objectivity of holistic scoring.

Scale Clarity

As for the clarity of the two scales, the raters unanimously held more positive views toward the analytic scale. They thought that the level descriptions on the analytic scale were more distinct than those on the existing scale. Specifically, as R4 said:

> There are some key words at different levels of the analytic scale which can help to discriminate the CET-4 writing performance at different levels. For example, in the sub-scale of *Linguistic range and accuracy*, Level A is characterized by expressions like 'rich vocabulary', 'accurate expressions', 'diversified sentence structures' and so on, while Level B is filled with key words like 'basic vocabulary', 'circumlocutions' , 'lexical gaps', etc.

There are five raters who spoke highly of the sub-scale of *Linguistic range and accuracy* on the analytic scale, deeming it to be the most explicit one among the four sub-scales. As Rater 7 put it:

> When describing language errors, expressions like 'grave language errors' and 'occasional language errors' are frequently used by the existing scale of the CET-4 writing. The interpretations of such expressions, however, rely significantly on modifiers like 'grave' and 'occasional'. By contrast, descriptions on the sub-scale of *Linguistic range and accuracy* are more informative and more helpful, say, 'some grave grammatical errors occur in writing which distort the meaning conveyed', and also 'errors in collocation and usage occur occasionally, which after all do not hinder comprehension'.

Scale Completeness

According to the raters, the rating criteria and the information entailed in the analytic scale were more complete than those in the existing scale, and the analytic scale was sufficient for describing CET-4 writing performance. No criterion or information was reported unattended. For example, Rater 4 said that:

> The new scale has already accounted for everything in my mind, so I am not able to put forward any other rating categories.

Similarly, Rater 5 also mentioned that:

> The new scale is all-encompassing, including almost all the features characterizing the CET-4 writing scripts.

Scale Operability

Compared with the aspects discussed above, the raters' attitudes toward the operability of the analytic scale were a little mixed. Of them, five raters made quite positive comments on the operability of the scale. For example, Rater 4 made the following comments:

> The process of using the analytic scale was not as complicated as I had assumed it to be. In fact, rating analytically can be highly efficient if the raters were well trained. As is known to all, the spoken test of CET adopts an analytic scoring approach. At the beginning, the raters may rate a bit slowly, however, once they become accustomed to the rating scale, they can rate both quickly and efficiently.

Similarly, Rater 5 also reported that he felt a bit unaccustomed when using the analytic scale at first, but after rating a few scripts, he became more confident.

On the other hand, three raters expressed their concerns over the rating efficiency of the analytic scale. Rater 3 thought that scoring the writing scripts with the analytic scale would be extremely time-consuming. As Rater 7 put it:

> There are abundant descriptions on the analytic scale; it is therefore very exhausting to use it in rating the CET-4 writing scripts. It takes more time to rate a single script because each rating criterion has to be taken into account.

In addition to the above comments, there were two raters who seriously doubted the operability of the analytic scale. Their major concern was also related to the time consumed by analytic scoring, thus they considered it to be essentially impractical for such a large-scale test. As Rater 10 said:

> If every single piece of writing were to be assigned four analytic scores, the work load would become extremely heavier and more raters would be recruited, which would lead to a higher cost.

Feedback for teaching

All the raters expressed affirmative viewpoints on the feedback that the analytic scale would be able to provide, highlighting that:

> Such a scale would be very useful in that it could make the teaching of English writing more focused (Rater 6).

As Rater 2 said:

> Compared with the existing rating scale, the analytic scale is more advantageous in providing feedback on the test takers' writing performance. The feedback, I believe, will in turn promote the effectiveness of the teaching and learning of English writing.

Also, this point was further echoed in the following quote:

> It is no denying that the existing rating scale is very effective in terms of discriminating the CET-4 writing performance. However, even writing performance assigned at the same proficiency level may differ from each other in that some of them might excel at language use, whilst the others might do better in organization or content. It is in this respect that the analytic scale would excel the current holistic scale (Rater 3).

Usefulness for Rater Training

As for the usefulness of the scales for rater training, eight raters held a more positive attitude toward the analytic scale. For example, Rater 1 said:

> I think the analytic scale will be more effective in rater training since it is fairly detailed and the level division is also reasonable. At least, I feel, I myself would benefit from rater training featuring such an informative scale.

Even Rater 7 who had previously expressed some doubts over the operability of the analytic scale acknowledged that:

> In the operational scoring of the CET-4 writing, due to the vagueness of the existing rating scale, the raters' interpretations of the scoring points sometimes deviate from what has been informed during the rater training. Well, I think with the analytic scale, this problem might be solved.

Despite these positive remarks, however, two raters cast some doubts on the usefulness of the analytic scale in rater training. As Rater 8 said:

> With so many descriptors, how can the efficiency of rater training be guaranteed? I'm afraid that the rater training will be turned into a laborious process which costs more but gains less.

2.4.4 Discussion

2.4.4.1 Construct Validity of the Scales

Bachman and Palmer (1996) define construct validity as "the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores" (p. 21). Knoch (2009) argues that to establish construct validity for a rating scale, we should not only understand the purpose and context of an assessment, but determine whether the rating scale can effectively help raters in arriving at scores.

According to Jin (2008), the intended purpose of CET-4 is two-fold: one is to provide an objective evaluation of a student' s overall English proficiency, and the other is to exert positive impact on the EFL teaching at the tertiary level in China. In view of the first aspect, the rating scale of the CET-4 writing is supposed to help evaluating the examinees' writing proficiency objectively. In other words, the rating scale should be powerful in discriminating the examinees' writing proficiency. In this study, the construct validity of the analytic scale could be evidenced on one hand by the results of MFRM analysis, and on the other by the raters' opinions elicited through the interviews. Specifically, the analytic scale resulted in higher examinee separation ratio and lower rater separation ratio, which indicated a more satisfactory scale discrimination. Besides, the raters' comments on the overall effectiveness of the two scales also bore out the superiority of the analytic scale in terms of its construct validity. According to the raters, the rating process facilitated by the analytic scale was more objective as the detailed level descriptions made the rating more valid. In other words, the raters didn't have to rely on their own intuitive judgments. As such, it is fair to say that the analytic scale is a little more advantageous than the holistic scale in terms of construct validity.

2.4.4.2 Reliability of the Scales

According to Bachman and Palmer (1996), reliability refers to consistency of measurement, and it is a necessary condition for construct validity. In this study, scale reliability was demonstrated through two major aspects: rater separation and rating variations. For the first aspect, two types of statistics were closely looked at –

rater separation ratio and rater separation reliability. The results showed that the analytic scale could help to yield moderately higher rater separation ratio as well as greater rater separation reliability, hence the scale reliability was more satisfactory.

The other aspect concerning scale reliability is raters' variation in rating which reveals the extent to which the raters' behaviors were consistent with the model expectation. Ideally, neither too much nor too little variation is desired because the two cases suggest that the raters either rate too inconsistently or over consistently (Bond & Fox, 2015). The results of MFRM indicated that compared with the holistic scale, the analytic scale lent itself more easily to control variations in rating. Hence, it can be concluded that the analytic scale is more reliable than the holistic scale in the CET-4 writing.

2.4.4.3 Impact of the Scales

According to Bachman and Palmer (1996), the impact of test use generally operates at two levels: a micro level, which concerns the impact on individuals, and a macro level, which refers to the impact on the educational system or society. As CET-4 is widely perceived to be a high-stakes test, the impact of the rating scale used in CET-4 writing should be considered both at the micro level and at the macro level. At the micro level, the scales are expected to provide useful and meaningful feedback to the test takers. In this regard, the results of the interviews could provide some insights. For instance, the raters mentioned that compared with the existing rating scale, more detailed information regarding the CET-4 writing performance was entailed in the analytic scale, such as the range of lexical use or the grammatical accuracy. If delivered to the test takers along with the score report, the information can supposedly help learners diagnose their strengths and weaknesses in EFL writing. Because the raters involved in the interviews were all experienced college English teachers, their opinions could to a large degree reflect the test takers' real needs.

In addition to the test takers, the raters as a group were also directly affected by the rating scale. During the interviews, the raters all mentioned that the analytic scale was more explicit in terms of level descriptions, enabling them to focus more intensively on the CET-4 writing scripts so as to ensure that the scripts could be assigned accurately at different levels of the analytic scale.

The impact of the rating scales at a macro level mainly concerns college English teaching. During the interviews, the advantage of the analytic scale in this respect had been echoed by the raters as they commented that the detailed information offered by the analytic scale would make the teaching of College English writing more focused than before.

In all, the above evidence indicated that the analytic scale could exert more positive impact both at micro level and at macro level.

2.4.4.4 Practicality of the Scales

According to Bachman and Palmer (1996), practicality denotes “the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities” (p. 36). In other words, the practicality of a test is constrained to some degree by the available resources. In a similar vein, the practicality of a rating scale is also limited by such kind of resources.

When being inquired about the implementation of the rating scales in CET-4 writing, the raters showed quite mixed attitudes: first, half of the interviewees exhibited great confidence in the practicality of the scale, commenting positively on the use of it in the CET-4 writing; Second, three of the interviewees expressed some concerns over the intensive labor involved in analytic scoring; Third, two of the interviewees seriously doubted the operability of the analytic scale in scoring the CET-4 writing performance. Admittedly, such mixed attitudes might mitigate the practicality of the analytic scale in the CET-4 writing to some extent. However, the interviewees’ mixed attitudes could be attributed to two possible factors: on one hand, the scale descriptors were established merely on an experimental basis at this stage, hence they still need further refinement; On the other, the interviewees were highly experienced in using the existing holistic scale, and it was somewhat difficult for them to tailor themselves to analytic scoring in such a short period. As such, retraining activities would be called for to help raters become more accustomed to the analytic scale, were it to be adopted in the operation scoring of the CET-4 writing. In addition, considering that large-scale EFL tests have increasingly resorted to automatic essay scoring in recent years, the potential of such an analytic rating scale in the CET-4 writing deserves further exploration because in automatic scoring process, the large rating cost traditionally associated with the use of analytic scales can be reduced significantly (Lee et al., 2010).

2.5 Conclusion and Limitations

Taking advantage of a mixed-methods approach, this study compared the effectiveness of two common types of rating scales in the context of scoring CET-4 writing performance. Results of the study indicated that the empirically developed analytic rating scale was more robust than the holistic scale currently adopted by CET-4 writing although its practicality needed more research.

Methodologically, this study not only offered some valuable insights into the validation of rating scales in the context of large-scale EFL writing assessments, but provided some new evidence for the comparison of the two commonly used types of scales. More practically, by validating an empirically developed analytic scale for CET-4 writing, the study had some implications for the clarification of the CET-4 writing construct and the interpretation of the CET-4 writing scores. As mentioned in the beginning, the interpretability of tests scores has become an increasing demand on behalf of test users. Jin and Yang (2018) also stress that "language assessment researchers in China should attach more importance to the interpretation of test scores, treating it as a crucial part in test validation" (p. 36).

However, this study is not without limitations. The most obvious one lies in that the evidence for the use of the two scales is only one-sided, that is, from CET-4 raters. CET-4 constructors' views are yet to be investigated. Therefore, it would be a gross overstatement to say that the analytic rating scale has been fully validated because "validity is an evolving property and validation is a continuing process" (Messick, 1989: 13). In other words, more evidence relating to the scale validity will be called for, particularly evidence demonstrating the applicability of the analytic scale to the operational scoring of the CET-4 writing as well as the potentiality of the scale in reporting the CET-4 writing scores.

Acknowledgements The author would like to thank the editors and anonymous reviewers for their feedback. The author would also like to thank all the raters who participated in the study. The author wishes to gratefully acknowledge the CET committee for giving permission of the CET-4 writing scripts.

Funding This book chapter is generously funded by Shandong provincial social science planning foundation, China [grant number 19CWZJ22] and a research foundation supported by Qingdao Agricultural University [grant number 663/1118704].

References

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371–383.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Dissertation. University of Toronto.

Berger, A. (2015). *Validating analytic rating scales: A multi-method approach to scaling descriptors for assessing academic speaking*. Peter Lang.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang.

Fei, Q., & Zhao, Y. Q. (2008). Problems in CET-4 writing rubric and scoring method. *Foreign Language Learning Theory and Practice, 4*(45–52), 93.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 6*(3), 238.

Hamp-Lyons, L. (1986). No new lamps for old yet, please. *TESOL Quarterly, 20*(4), 790–796.

Hamp-Lyons, L. (1991). Scoring procedures. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Ablex.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing, 29*, A1–A5.

Jian, Q. M., & Lu, J. P. (2005). Inadequacy of the proposition for the writing section of CET-4. *Foreign Languages and Their Teaching, 1*, 32–33.

Jin, Y. (2008). Powerful tests, powerless test designers? Challenges facing the College English Test. *CELEA Journal, 5*, 3–11.

Jin, Y., & Yang, H. Z. (2018). Developing language tests with Chinese characteristics: Implications from three decades of the College English Test. *Foreign Language World, 2*, 29–39.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Peter Lang.

Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*(3), 391–417.

Li, Q. H. (2014). *Constructing and validating a rating scale for TEM-4 writing*. Science Press.

Li, H. (2015). The effects of the use of holistic and analytic scales on the reliability of EFL essay scoring. *Foreign Languages and Their Teaching, 2*, 45–51.

Linacre, J. M. (2013). *Facets Rasch measurement computer program* (version 3.80.0). Winsteps.com.

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*(1), 23–44.

Shaw, S. D. (2004). Automated writing assessment: A review of four conceptual models. *Cambridge Research Notes, 17*, 13–18.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*(2), 163–182.

Sun, H. Y., & Han, B. C. (2013). A Comparative study of analytic and holistic rating scales for English writing. *Journal of PLA University of Foreign Languages, 6*, 48–54.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall Regents.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*, 400–409.

White, E. M. (1985). *Teaching and assessing writing*. Jossey-Bass Inc.

Wiseman, C. (2008). *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring method*. Doctoral dissertation. Columbia University.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Shaoyan Zou received her PhD degree in Shanghai Jiaotong University. She is currently Associate Professor of Qingdao Agricultural University. Her research interests mainly include language assessment and evaluation, and foreign language teaching. She has been involved in the China Standards of English project since 2014, hence particularly interested in research on the development and validation of language proficiency scales.

Chapter 3
A Comparative Study of Chinese Test Takers' Writing Performance in Integrated and Discrete Tasks: Scores and Recurrent Word Combinations in PTE Academic

Yu-Hua Chen and Ying Zheng

Abstract Despite an increasing number of studies on L1 Chinese students' L2 English writing in recent years, little research is conducted on Chinese test takers' performance in high-stakes international tests, particularly in terms of how integrated and discrete assessment may impact on their writing. This book chapter, therefore, aims to fill this gap by comparing Chinese and non-Chinese test takers' responses in different writing tasks of PTE Academic in terms of scoring and use of recurrent word combinations. As an international test of academic English, PTE Academic includes both the traditional design of independent essay writing as well as two integrated tasks of summary writing from listening or reading input.

Written samples from 500 Chinese test takers and another 500 test takers of other L1s were randomly chosen from PTE Academic and analysed. The results indicate that Chinese test takers outperformed their non-Chinese peers in the tasks of read-to-summarise and essay writing. Chinese test takers also tended to write longer responses with more occurrences of recurrent word combinations, and they used more stance expressions in the essays compared with their peers. A closer examination of highly frequent word combinations, however, suggests a relationship between task type and the language elicited, regardless of test takers' L1.

Keywords Academic writing · Task type · Integrated writing · Recurrent word combinations · PTE Academic

Y.-H. Chen (✉)
Coventry University, Coventry, UK
e-mail: yu-hua.chen@coventry.ac.uk

Y. Zheng
University of Southampton, Southampton, UK
e-mail: ying.zheng@soton.ac.uk

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_3

3.1 Introduction

As a computer-based international test of academic English, Pearson Test of English Academic (hereafter PTE Academic) contains a variety of integrated writing tasks, including listen-to-write and read-to-write items which require test takers to produce a summary based on a given passage. Traditionally, a typical writing task involves only writing skills, hence often known as discrete or independent writing assessment, whereas an integrated task uses listening or reading input (or a combination of both as in the case of TOEFL iBT) as the source of information for language production, hence also known as source-based writing (e.g., Merkel, 2020). Although an increasing number of studies in recent years have investigated the linguistic features in Chinese students' L2 English writing (e.g., Chen & Baker, 2010, 2016; Leedham & Cai, 2013), very little research is conducted on Chinese test takers' performance in integrated writing assessment in relation to the traditional design of independent writing assessment. This study, therefore, aims to fill this gap by investigating Chinese test takers' writing performance in terms of scores and use of recurrent word combinations in PTE Academic, where a task of independent essay writing is used to measure writing ability as well as two integrated summary writing tasks that also evaluate reading or listening comprehension.

Recurrent word combinations are recurring word sequences extracted by computers with a set of selection criteria. This type of formulaic sequences has been identified as "building blocks" of discourse (Biber et al., 2004), and their occurrences and use in the written samples from Chinese and non-Chinese candidates will be compared between independent and discrete tasks in the current study.

One thousand test takers in total were randomly selected (500 each for Chinese and non-Chinese candidates) from PTE Academic, and four summaries (two each for the integrated tasks of read-to-summarise and listen-to-summarise) and one essay from each of the test takers in one single sitting of the test were extracted. With Chinese test takers as the focus of this study, the non-Chinese group was used as a reference to determine whether the patterns identified was unique to the Chinese group only. By comparing Chinese test takers' scores of those writing tasks and written samples with test takers from other L1 backgrounds, we will be able to identify distinctive features in L1 Chinese learners' L2 English writing between independent and integrated writing items. The results can then be used to help us better understand issues in Chinese learners' academic writing in different task types and provide insights into the teaching, learning or test development for the area of English for Academic Purposes (EAP).

3.2 Literature Review

3.2.1 Integrated vs. Independent Writing Assessment

Traditionally, writing ability is often assessed through a timed task independent of other skills such as reading or listening. Independent tasks are perhaps still the most popular approach in assessing L2 writing ability because its design is straightforward and marking also tends to be easier in comparison to integrated assessment which involves more than one skill. However, this 'snapshot' approach has been criticized because of its limited capacity (Hamp-Lyons & Kroll, 1996) as well as the lack of authenticity or validity (e.g., Cho, 2003; Read, 1990; Weigle, 2004). One of the greatest challenges confronted with independent writing assessment is its reliance on topic familiarity or creativity to a large extent. Since writing is notoriously time-consuming, a test often includes only one or two tasks which evaluate writing skills. If a test taker, however, is not interested in or not familiar with the given topic(s), this potentially place them at a disadvantage.

Compared to independent tasks such as timed essays, integrated tasks appear to have a higher degree of authenticity. Cumming et al. (2000) argue that real-life writing often involves writers drawing on various resources available during the writing process rather than just a given topic. Similarly, Weigle (2004) and Sawaki et al. (2013) point out that in real academic settings, English learners are expected to synthesize what they read or hear into written works. This seems to suggest that academic writing should be integrated with other skills. In relevant literature in Higher Education, this is sometimes referred to as source-based writing, and the focus typically surrounds the notion of how student writers engage with source materials (e.g., Hirvela & Du, 2013; Li & Casanave, 2012; Merkel, 2020). According to a meta-analysis conducted by Cumming et al. (2016), they compared 69 empirical studies on source-based academic writing and concluded that source-based writing often imposes a challenge for students, particularly for L2 students because of different prior knowledge and experience.

Accordingly, the combination of different skills and source materials into writing tasks under the general term of integrated assessment reflects the reality of academic writing that test takers will encounter in their future studies. However, Buck (2001) points out the challenges of identifying the construct of integrated tasks. After reviewing a number of studies which report on integrated writing assessment, Cumming (2013) also concludes that integrated writing tasks "confound the measurement of writing ability with abilities to comprehend source materials". Although there has been an increasing interest in integrated writing assessment in recent years, not much research can be found regarding the inclusion of both read-to-write and listen-to-write tasks. Zheng and Mohammadi's study (2013), which explored the constructs of six writing item types in PTE Academic via the employment of exploratory factor analysis, is probably one of the few studies that have offered an insight into not only read-to-write tasks but also listen-to-write tasks. Considering the high stakes of university admission tests such as PTE Academic, it is therefore important for us to better understand the relationship between various types of integrated writing and independent writing.

3.2.2 Research into L2 English of L1 Chinese Students

The motivation of focusing on Chinese students' L2 academic English writing is initiated by an increasing number of Chinese students who choose to study in English-speaking countries or EMI (English as Medium of Instruction) institutions. Accordingly, there is also growing literature which reports Chinese students' L2 English performance in comparison with other L1 groups. For example, previous studies have explored the potential link between test takers' first languages and their performance in English tests (Abbott, 2007; Chen & Henning, 1985; Kim, 2001), and one of the earliest attempts to investigate this relationship was a study conducted by Chen & Henning (1985). By comparing the responses from native speakers of Chinese and Spanish in an EFL test of five parts: listening, reading, grammar, vocabulary, and writing error correction, they identified vocabulary items which favoured the Spanish test takers. Also comparing the performance between Chinese and Spanish test takers, Sasaki (1991) found that vocabulary items with idiomatic expressions favoured the former group while English-Spanish cognates favoured the latter. The tradition of comparing a group of a specific L1 learners lays a foundation for the current study to compare L1 Chinese test takers with their peers of other L1s.

In relation to sourced-based writing discussed in the last section, Shi (2004) compared Chinese university students' L2 English writing with an L1 English student group. The written samples were collected from two types of tasks – writing an opinion essays and writing a summary, both of which were given a source text as the prompt. The findings indicate that students in general copied more words from the source text in the summary task than in the opinion essay task. Shi (ibid.) also found that Chinese students tended to borrow words from source text without proper citations when compared with the group of L1 English students.

The examples above illustrate that Chinese learners of English appear to perform their writing with certain linguistic features specific to their L1 background. Together with these distinguishing features, the increasing number of Chinese students pursuing a degree abroad in English, hence having to take high-stakes international tests of English (MOE, 2018), highlights the necessity for studies on Chinese test takers' written performance in these "gatekeeping" tests. The current study therefore addresses this call for research and hopes to provide some pedagogical implications for EAP courses catering for Chinese students.

3.2.3 Recurrent Word Combinations in L2 Writing

Functioning as "building blocks" of discourse (Biber et al., 2004), recurrent word combinations are also known as lexical bundles. Recurring word sequences are typically retrieved using a computer tool with specified frequency or dispersion thresholds. Recent research indicates that there are significant differences in terms of occurrences and discourse functions of recurrent clusters between genres (Biber et al., 1999, 2004; Biber & Barbieri, 2007), disciplines (Hyland, 2008), between L1 and L2 writers (e.g., Ädel & Erman, 2012; Chen & Baker, 2010), or across L2 proficiency levels (Chen & Baker, 2016).

Focusing on Chinese students' L2 English writing, Chen & Baker (2016) examined the use of lexical bundles and found that at lower levels, the writing discourse from Chinese learners shared more features with that of conversation, while the discourse of more proficient writing was more similar to that of academic prose. Similarly, Ruan (2016) looked at a corpus of academic texts written at four points of time between Year 1 and Year 4 at an EMI university in China to identify frequently used lexical bundles. The findings suggest that there is a developmental pattern of lexical bundle use in terms of both structures and discourse functions.

Staple et al.'s (2013) work is one of the few studies where the use of recurrent word combination was investigated and integrated tasks were included in the writing samples from a high-stakes English test. In their study, lexical bundle use was compared across three levels in the writing tasks of TOEFL iBT: high, intermediate and low. The findings indicate that test takers at lower levels overall used more bundles and also "borrowed" more bundles from the given prompts. Lexical bundles in this study were divided into two groups: prompt and non-prompt bundles, with the former referring to those "that appeared word for word and that are clearly related to the topic or task" (ibid.: 217). Based on our discussion about source-based writing earlier, it is reasonable to assume that test takers would "borrow" more words from the integrated tasks, hence more prompt bundles, because such tasks require test takers to use the source material to produce a summary. In Staple et al. study (2013), the written samples from the integrated and discrete writing tasks, however, were not distinguished when results were presented. It is therefore impossible to see whether there was any difference in terms of prompt bundles versus non-prompt bundles between integrated and independent writing tasks.

Another issue in traditional research on second language writing (or learner corpus research in recent years) is that L2 writing from learners of a certain L1 background is often compared with L1 writing (e.g., Granger, 1998; Granger et al., 2002; Ädel & Erman, 2012) rather than L2 learners' peers from other L1 backgrounds. This might lead to a potential misconception that Chinese students' writing, for example, seem to have plenty of issues such as overuse or underuse of certain language expressions, but without a comparison with other L2 learners, we cannot possibly determine whether those patterns are universal across different L2 students or unique in Chinese students' writing.

Taking into account relevant research discussed above, the current study will therefore aim to compare Chinese test takers' scores and use of recurrent word combinations (specifically prompt and non-prompt based ones) in the integrated and discrete writing tasks with those from non-Chinese test takers.

3.3 Research Questions

Focusing on Chinese students' writing performance in the integrated and discrete tasks in PTE Academic in relation to the non-Chinese group, the research questions are formulated as the following:

RQ1. How do Chinese test takers perform in the integrated and discrete writing tasks in comparison to non-Chinese test takers?
RQ2. To what extent are there differences of overall frequencies of recurrent word combinations between the integrated and discrete writing tasks in these two test taker groups?
RQ3. Are there differences, if any, of prompt and non-prompt recurrent word combinations between integrated and discrete writing tasks in these test taker groups?
RQ4. How are the discourse functions of non-prompt recurrent word combinations different or similar between integrated and discrete writing tasks in Chinese test takers' writing in relation to non-Chinese test takers?

3.4 Data and Methodology

3.4.1 Test Takers

Focusing on the comparison between independent writing and integrated writing tasks, 1,000 test takers were randomly selected from L1 Chinese and other L1 backgrounds (500 each), and their scores and written responses were used for comparison.

The demographic backgrounds of those chosen test takers can be found in Table 3.1. As those test takers are randomly selected, it is reasonable to assume that they represent the actual test taker population to a large extent. Among the non-Chinese test takers, the largest group comes from India, which accounts for over half of the international test takers (53.6%), while the remaining test takers primarily come from Southern or South-east Asia. It is also interesting to see that 21.4% of those international test takers speak English as their home language. Among those speakers of English, 43.9% again come from India while the others cover a range of other countries such as Malaysia, Pakistan, Philippines, Singapore, South Africa or U.K.

Table 3.1 Comparison of Chinese and non-Chinese test takers

	Chinese	non-Chinese (International)
Number	500	500
Nationality	China PRC 100%	India 53.6% Pakistan 7.2% Nepal 6.6% Philippines 4.2% South Korea 2.6% Vietnam 2.6% Others 23.2%
Home language	Mandarin 100%	English 21.4% (43.9% from India) Hindi 12.4% Punjabi 10.2% Urdu 8.2% Nepalese 5.6% Telugu 5.6% Others 36.6%

Table 3.2 An overview of the three writing item types in PTE Academic

Communicative skills	Item type code	Item type	Brief description	Enabling skills
Reading & writing	RW-SUMM	Summarize written text	Summarize written text in one sentence of no more than 75 words	Content; form; grammar; vocabulary
Listening & writing	LW-SUMM	Summarize spoken text	Summarize spoken text in 50–70 words	Content; form; grammar; vocabulary; spelling
Writing	WW-ESSA	Write essay	Write an argumentative essay (200–300 words) in response to a given topic	Content; form; development, structure and coherence; grammar; general linguistic range; vocabulary; spelling

3.4.2 *Test Tasks*

An overview of the three writing tasks from PTE Academic and the skills measured can be found in Table 3.2. Each of the writing task type is coded in a similar way: the first two letters in the first part indicate the skills assessed (e.g., RW for reading and writing), and the four letters in the second part refer to the task (e.g., SUMM for summary writing). The task type code will be used in the rest of this chapter when a specific task is discussed.

In addition to an overall score, the score report that a test taker receives after completing the test also contains communicative skill scores (for writing, speaking, listening and reading) and enabling skill scores (i.e. productive subskills such as content or vocabulary for writing and speaking tasks only). The automated scoring system used in PTE Academic is trained and calibrated on the trait scores of response

Table 3.3 Number of items used for each of the Chinese and non-Chinese groups in the current study

		Chinese	Non-Chinese
Skills	Item type code	Number of responses	Number of responses
Reading & writing	RW-SUMM	1000	1000
Listening & writing	LW-SUMM	1000	1000
Writing	WW-ESSA	500	500
Total		2500	2500

samples scored by human markers, and the traits measured in PTE academic include a number of enabling skills (Pearson, 2018). Enabling skills in Table 3.2 therefore provide valuable information about the linguistic features measured for the two summary writing and the independent writing tasks in PTE Academic. As can be seen, content is scored for each of the writing tasks, and language ability is evaluated differently between summary and discrete writing tasks as the latter covers a wider range of traits.

For the group of 500 Chinese test takers, 1000 item responses for each of the integrated task types (two items each test taker for the RW-SUMM and LW-SUMM tasks) were extracted. Because each test paper only had one essay item (WW-ESSA), only 500 of them were available. This means two listen-to-write, two read-to-write summary tasks as well as one essay task for each of the test takers, amounting to 2500 responses in total (see Table 3.3). The same procedure was completed for the non-Chinese group. It has to be noted that because the test paper delivered to a test taker on a computer is randomly assigned and each paper consists of different items, it is impossible to control the prompts between the Chinese and non-Chinese test takers, which might have an impact on the statistical and linguistic analysis to some extent. Yet since the data was randomly selected from a large pool (as will be discussed later), we believe such an impact should be quite minimal.

3.4.3 *Procedures*

The data including item responses and scores (from automated scoring) were provided by Pearson. To answer RQ1, MS Excel and SPSS were used for the calculation of score averages, standard deviations and correlations among the three task types. Independent samples T-tests were also run to test significant average score differences between the two groups. For RQs 2–4, the corpus tool AntConc 3.5.8 (Anthony, 2019) was utilised to extract recurrent word sequences, and any continuous word sequences with minimum three occurrences were retrieved. Note that “type” and “token” are distinguished in this study: the former refers to different word combinations (e.g., *on the other hand* and *at the same time* counted as two types) while the latter refers to the number of occurrences (e.g., *on the other hand* occurring twice counted as two tokens).

To investigate the relationship between prompt and non-prompt word combinations in integrated and discrete writing tasks in RQ3, because of the huge amount of data, i.e. thousands of recurrent 4-word combinations with tens of thousands of instances extracted this way, only the most frequent 20 word combinations (with occurrences ranging between 10 and 40 times) from each of the task types were chosen for further analysis. This is considered comparable with Staple et al.'s (2013) study, where a cut-off point of 25 occurrences was used to retrieve lexical bundles from the writing responses of 480 test takers in TOEFL iBT but discrete and integrated tasks were not separated.

In addition to a cut-off frequency threshold, the dispersion requirement, i.e. how many texts a word sequence occurs in (typically no less than 3–5 texts), is also often adopted when determining a recurrent word combination to guard against individual idiosyncrasies (e.g., Ädel & Erman, 2012; Biber et al., 2004; Biber & Barbieri, 2007; Chen & Baker, 2016; Hyland 2008; Staples et al., 2013). A scrutiny of highly recurring clusters in the current study revealed that all of them occurred across different item responses. This was probably because test takers in PTE Academic only needed to summarise the input in one or two sentences in an integrated task (see Table 3.2), and the shorter lengths of texts plus a higher frequency threshold adopted in this study warrants that it is unlikely a word combination would occur multiple times in one single text.

In terms of RQ 4, only non-prompt word combinations were further examined because they represented test takers' writing skills in using their own words to present arguments or organize ideas rather than "borrowing" chunks of text as in prompt-based word combinations. The primary reason for excluding prompt bundles here is the concern of data sensitivity as our data comes from live test content, and any information regarding prompts needs to stay confidential to maintain the test integrity.

After removing all the prompt-based word sequences, the remaining highly frequent non-prompt clusters were then classified according to the discourse functions of referential, stance, or discourse organising, a widely used framework developed by Biber et al. (2004; Biber & Barbieri, 2007).

3.5 Results

3.5.1 Comparing Writing Task Scores

To answer RQ1, the average item scores and standard deviations (SD) from the three writing tasks and overall scores for both Chinese and non-Chinese test takers were calculated, and the results can be found in Table 3.4. The overall score is based on all four skills and all the test items from one single test sitting per test taker. Despite a lower average overall score (61.7), Chinese test takers outperformed their peers (with an overall score of 64.3) in two of the three writing tasks: WW-ESSA (7.13 vs. 6.35) and RW-SUMM (2.20 vs. 1.91). In contrast, both groups achieved

Table 3.4 Descriptive statistics of Chinese and non-Chinese groups' scores in the writing tasks of PTE Academic

		Integrated		Discrete	Overall Score
Skill		Read-to-write	Listen-to-write	Writing only	
Item type		RW-SUMM Summary	LW-SUMM Summary	WW-ESSA Essay writing	
Max. score		4	7	10	90
Chinese	Average	2.20	3.85	7.13	61.68
	SD	*0.77*	*1.64*	*1.97*	*12.59*
Non-Chinese	Average	1.91	3.87	6.35	64.33
	SD	*0.77*	*1.75*	*2.38*	*15.84*

a very similar score for the LW-SUMM task. This suggests that the Chinese students in general performed relatively well in the writing tasks in comparison with the average test takers. Based on our relevant teaching experience, Chinese students tend to be weaker in listening comprehension in comparison with reading, and it is possible that because of this reason the listen-to-write summary task (LW-SUMM) might be more challenging for Chinese test takers than the essay or read-to-write tasks when compared with non-Chinese test takers. Correlations among the three writing tasks indicate that the average scores of the three task types are all significantly correlated ($p < 0.01$), with the WW-ESSA score correlated more highly with LW-SUMM ($r = .509$), followed by the correlation between WW-ESSA and RW-SUMM ($r = .353$) and the correlation between RW-SUMM and LW-SUMM ($r = .336$).

Overall, the non-Chinese group seems to have slightly larger standard deviations (except for the RW-ESSA task type), indicating a wider spread in performances on the LW-SUMM and WW-ESSA tasks. This is unsurprising, considering that those test takers came from a much wider range of different backgrounds (cf. Table 3.1).

In Table 3.5, results from Levene's test shows that equal variances between the two test taker groups can be assumed on their RW-SUMM performance, and the two average group scores are significantly different (2.20 vs. 1.91). While on LW-SUMM and WW-ESSA writing, equal variances cannot be assumed, and the two group average scores are not significantly different for LW-SUMM (3.85 vs. 3.87), but significantly different on WW-ESSA (7.13 vs. 6.35).

3.5.2 Comparing Overall Frequencies of Recurrent Word Combinations

As mentioned earlier, recurrent word combination with minimum frequency of three times were first retrieved across the writing tasks from Chinese and non-Chinese test takers, and the type and token counts of those 4-word combinations are presented in Table 3.6. As can be seen, Chinese test takers overall used significantly more word

Table 3.5 Results of the independent samples test in comparing Chinese and non-Chinese groups' scores in the writing tasks of PTE Academic

Independent samples test									
		Levene's test for equality of variances		t-test for equality of means					
								95% confidence interval of the difference	
		F	Sig.	t	df	Sig.	Mean Difference	Lower	Upper
RW-SUMM	Equal variances assumed	0.472	0.492	6.016	998.000	0.000*	0.292	0.197	0.387
LW-SUMM	Equal variances not assumed	4.389	0.036*	−0.149	994.366	0.882	−0.016	−0.227	0.195
WW-ESSA	Equal variances not assumed	23.645	0.000*	5.645	963.920	0.000*	0.780	0.509	1.051

Note: * significant at $p < 0.05$

Table 3.6 Recurrent word combinations with no less than three occurrences

Test takers	Type/token	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Type	2374	938	2479
	Token	14,695	4555	16,689
	Ratio per 100 tokens	26.53	7.32	13.61
Non-Chinese	Type	1067	820	1528
	Token	4640	3953	9069
	Ratio per 100 tokens	12.54	6.33	7.58

Table 3.7 Total and average lengths of task responses in the Chinese and non-Chinese groups

Test takers	Lengths	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Total	55,383	62,219	122,629
	Average	55.4	62.4	245.7
Non-Chinese	Total	36,998	62,483	119,706
	Average	37.0	62.6	239.9

combinations in all of the writing tasks in terms of both types and tokens. Note that at this stage prompt and non-prompt clusters were not distinguished. Because there are various length requirements for each task (cf. Table 3.2), the occurrences are also standardised to the ratios per 100 tokens for comparison.

However, note that overall Chinese test takers appear to have produced longer responses in the RW-SUMM and WW-ESSA tasks (see Table 3.7), and it is therefore perhaps not surprising to see that more recurrent word combinations were identified in the Chinese test taker group although this is also true in terms of the ratio per 100 words (cf. Table 3.6).

3.5.3 Comparing Prompt and Non-prompt Recurrent Word Combinations

To compare the use of prompt and non-prompt recurrent word combinations, as mentioned earlier, only the most frequent word combinations (i.e. 20 for each of the tasks) were examined and divided into prompt and non-prompt categories (cf. Staples et al., 2013). The results can be found in Table 3.8. As can be seen, there is a striking difference among writing tasks in terms of the distribution of prompt and non-prompt word combinations, and the patterns of use is consistent for both the Chinese and non-Chinese test takers' groups (Figs. 3.1 and 3.2). The RW-SUMM tasks have the highest numbers of prompt-based word combinations in terms of both type and token whereas the highest numbers of non-prompt word combinations are found in the tasks of LW-SUMM or WW-ESSA.

Table 3.8 Most frequent 20 recurrent word combinations divided into prompt and non-prompt groups

Test takers	Non-prompts/ Prompts	Type/ token	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Non-prompts	Type	1	19	13
		Token	35	633	692
	Prompts	Type	19	1	7
		Token	479	18	319
Non-Chinese	Non-prompts	Type	2	19	10
		Token	27	494	392
	Prompts	Type	18	1	10
		Token	225	19	316

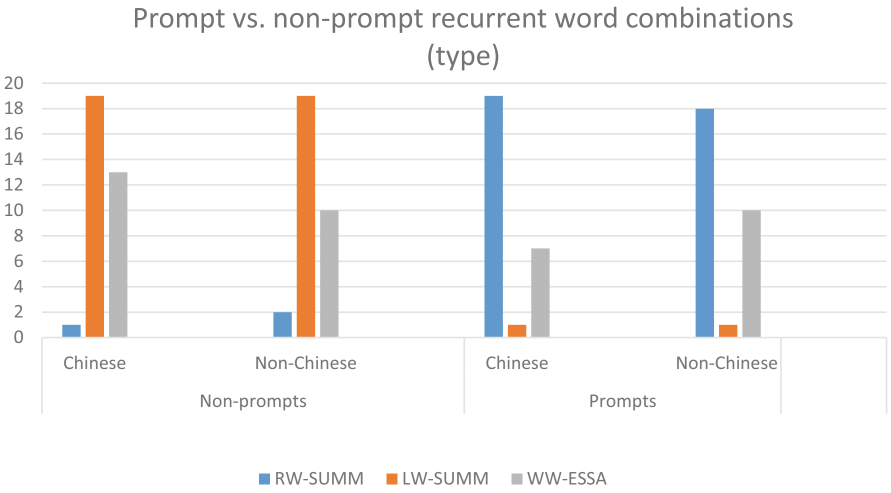


Fig. 3.1 Most frequent 20 word combinations (type) divided into prompt and non-prompt categories

3.5.4 *Non-prompt Recurrent Word Combinations*

In the last section, we have seen the writing task types appear to have an impact on the ratio of prompt and non-prompt word combinations, and prompt bundles account for a significantly large portion of written text in the RW-SUMM tasks, at least in the most frequent word clusters that we examined. As mentioned in Sect. 3.4.3, only non-prompt word combinations were further qualitatively analysed. The highly frequent non-prompt word combinations from Table 3.8 were categorised on the basis of three discourse functions: making reference, expressing stance or organising the discourse (cf. Biber et al., 2004; Biber & Barbieri, 2007). In a more qualitative

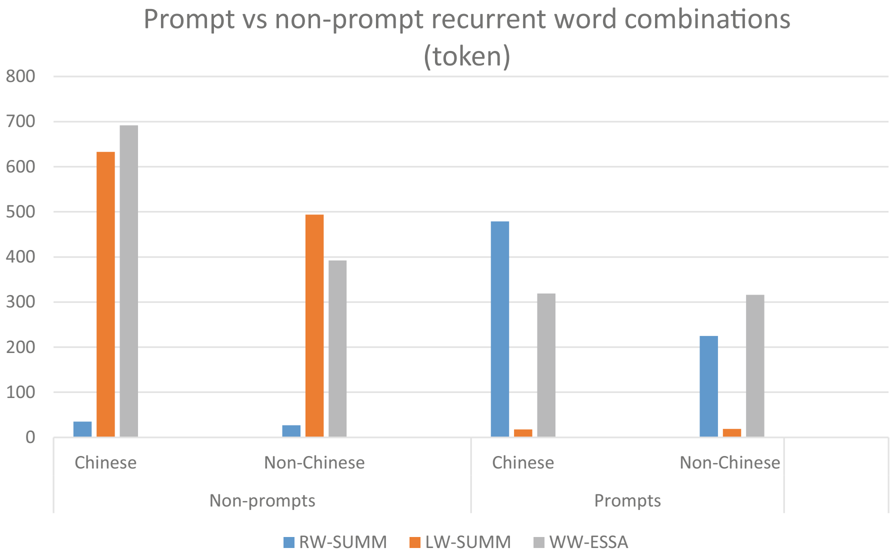


Fig. 3.2 Most frequent 20 word combinations (token) divided into prompt and non-prompt categories

examination of the data, it was then discovered that quite a few 4-word combinations were part of a longer phraseological unit. For example, the three sequences of "*this lecture mainly talks*", "*lecture mainly talks about*", and "*mainly talks about the*" actually form a longer 6-word expression of "*this lecture mainly talks about the*". To guard against inflated counts of word combination types, those clusters were therefore combined with brackets indicating the extensions such as "*this lecture mainly talks* (*about the)*" (cf. Chen & Baker, 2010). The results are presented in Table 3.9.

As can be seen, there is a striking difference between discourse functions and task types. In the RW-SUMM tasks, only "*at the same time*" and "*on the other hand*" were identified, and those are typical discourse markers found in the literature (e.g., Chen & Baker, 2010, 2016). In terms of the LW-SUMM tasks, almost all the word combinations fall into the discourse function of discourse organising, which incorporated the phrases of "*the lecture*" or "*the speaker*" to introduce main ideas of a summary. Interestingly, this type of introductory expressions were not found in the other summary writing task RW-SUMM with reading input. For the WW-ESSA tasks, this is the only task type that stance expressions occurred in the recurrent word combinations retrieved here, and Chinese test takers appeared to have used more stance bundles than their peers to present an argument or express their stance. In terms of discourse organizers in this task type, again, typical bundles that can be found in the literature such as "*on the other hand*" or "*is one of the most*" were identified.

Table 3.9 Discourse functions of non-prompt recurrent word combinations in the three writing tasks

Task	Function	Chinese	Non-Chinese
RW-SUMM	Referential	1. *at the same time* (35)[a]	1. *at the same time* (15)
	Discourse organising	–	2. *on the other hand* (12)
LW-SUMM	Referential	–	–
	Discourse organising	Referring to "*lecture*" 1. *this lecture talks about* (67) 2. *the lecture talks about (the)* (52) 3. *in this lecture the (42)* 4. *(this) lecture mainly talks about (the) (33)* 5. *the lecture is about (31)* 6. *this lecture is about (30)* 7. *lecture is talking about (28)* 8. *this lecture is mainly (19)* *Referring to "speaker"* 9. *the speaker talks about (the) (61)* 10. *the speaker mentioned that (33)* 11. *the speaker mentions that (19)* 12. *the speaker said that (17)* *Referring to both* 13. *this lecture the speaker (talks) (32)* 14. *the lecture the speaker (talks)* (23)	Referring to "*lecture*" 1. *the lecture was about (the)* (49) 2. *the lecture is about (the) (33)* 3. *in this specific lecture (17)* 4. *lecture and it comprises (that) (16)* *Referring to "speaker"* 5. *the speaker was discussing (about) (44)* 6. *according to the speaker (29)* 7. *the speaker talks about (19)* 8. *the speaker talked about (17)* 9. *speaker was talking about* (16) *Others* 10. *the talk delineates the (significance of)* (41)
WW-ESSA	Stance	1. *(while) others hold the view (that)* (74) 2. *some people believe that* (71) 3. *as far as I (am concerned)* (62) 4. *in my opinion I* (46) 5. *it has been argued (that)* (40) 6. *my point of view* (44)	1. *I would like to* (43) 2. *(above) one can conclude that* (31)
	Discourse organising	7. *on the other hand* (69) 8. *first and foremost it* (48) 9. *is the most important* (39)	3. *(is) one of the (most) (52)* 4. *on the other hand (49)* 5. *in this essay I (39)* 6. *at the outset there (are) (36)* 7. *this essay will discuss (31)*

[a] Frequency is added at the end of each word combination in brackets

3.6 Discussion and Conclusion

This study set out with a view to reviewing and comparing PTE Academic writing tasks, both integrated and independent, to compare the writing scores and use of recurrent word combinations between integrated and discrete writing tasks in the group of Chinese test takers in comparison with test takers of different L1s. Integrated writing assessment, or source-based writing, is still a fast-growing area which deserves further research. In this study, it can be seen that Chinese test takers outperformed their non-Chinese peers in the read-to-summarise and essay writing tasks despite a lower overall score in PTE Academic, and they also tended to produce longer responses in both of these two task types. This is probably somewhat unexpected for many because in the traditional second language writing research where Chinese students' writing is compared with native standards of English, usually only issues such as non-native language use are reported. This is therefore perhaps encouraging for Chinese students, particularly considering the fact that over 20% of the non-Chinese test takers in PTE Academic actually use English as their home language (see Table 3.3). For the listen-to-summarise tasks, interestingly, Chinese and non-Chinese test takers performed similarly in relation to the item scores and response lengths.

In terms of recurrent word combination use, higher numbers of occurrences in Chinese test takers' writing were identified in all of the writing tasks when compared with the non-Chinese group. This is interesting because Staples et al. (2013) reported that candidates at lower level in TOEFL iBT overall used more bundles in comparison with candidates at higher level, but in the current study, Chinese test takers overall performed significantly better than the other group in the writing tasks (except for LW-SUMM tasks) while also using more recurrent word combinations. It is possible that more content was covered in Chinese students' writing, considering that they tended to produce longer responses, and content is one of the enabling skills evaluated in PTE Academic (see Table 3.2).

It also has to be noted that there appears to be a relationship between task type and language use represented by recurrent word combinations. In the current study, prompt-based bundles dominated almost all of the most frequent word combinations in the read-to-summarise tasks, but this feature was not found in the listen-to-summarise or the essay writing tasks. This patterns also seems to hold true for both Chinese and non-Chinese groups. It is likely that for the read-to summarise tasks, where the source text can be easily accessed, test takers can just easily copy and paste chunks of text to the summary they are producing. This may be a concern for a test of academic English because discrete and integrated writing tasks contain different constructs, or labelled "task representation" by Plakans (2010). For academic writing, the ability to paraphrase as well as proper citation are both very important skills, and failing to do so can lead to plagiarism, a serious offense in Higher Education. However, this aspect of integrated writing assessment seems to have been overlooked in the design of integrated writing assessment in existing international tests of academic English, and it is perhaps an area that should be researched further.

In terms of integrated writing assessment, there appears to be mixed attitudes towards the implementation of integrated tasks in high-stake English language tests (Wei & Zheng, 2017), and one possible reason may be the lack of comprehensive L2 read-to-write or listen-to-write constructs (Plakans, 2008; Sawaki et al., 2013; Zheng & Mohammadi, 2013). However, instead of avoiding "the necessary interdependence of writing performance on reading and/or listening performance", it should be the time for language assessment researchers to take a step forward and redefine the writing construct for academic purposes, not only for the benefits of integrated tasks but also of independent tasks (Cumming, 2013: 5). We also argue that future research should look at both skills and sub-skills, for example, how to engage with the source in academic writing, which will allow us to better interpret test results. The interpretation will enable test developers to improve the tests and provide pedagogical implications to better support test takers in preparation for their tests. In terms of limitations, there are a couple of methodological issues in the current study. First of all, because our data were extracted from live test content, unfortunately we are unable to reveal much information about task prompts, but great efforts were made to ensure that the analysis still generated some meaningful results despite the lack of prompts for readers. In addition, in the qualitative investigation of recurrent word combinations, as a result of huge amounts of data, only the most frequent items were examined from each of the writing task types, and the generalizability of the results were therefore affected to some extent.

Considering the high stakes of academic English tests which are often used for visa or admission purposes, we urge that in the future more research should be conducted in the area of integrated writing assessment in relation to the traditional task of essay writing. For example, test takers' strategies, types of source input (e.g., Rukthong & Brunfaut, 2020) or students' perceptions and practices about plagiarism in source-based writing (Merkel, 2020) should be researched together with scoring and textual analysis of test taker responses to better inform test development.

Acknowledgements We are grateful for Pearson plc who kindly provided the data for us to use in this study.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7–36.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes., 31*(2), 81–92.

Anthony, L. (2019). AntConc 3.5.8 (Windows) [Computer Software]. Waseda University. Available from http://www.laurenceanthony.net/

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263–286.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30–49.

Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: lexical bundles in rated learner essays. CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849–880.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–163.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8*(3), 165–191.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1), 1–8.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL monograph series no.18). Educational Testing Services.

Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes, 23*, 47–58. https://doi.org/10.1016/j.jeap.2016.06.002

Granger, S. (Ed.). (1998). *Learner English on computer*. Longman.

Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition, and foreign language teaching*. John Benjamins.

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6*(1), 51–72.

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?" Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12*, 87–98. https://doi.org/10.1016/j.jeap.2012.11.005

Hyland, K. (2008). *As can be seen*: lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89–114.

Leedham, M., & Cai, G. (2013). *Besides… on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22*(4), 374–389.

Li, Y., & Casanave, C. P. (2012). Two first-year students' strategies for writing from sources: Patchwriting or plagiarism? *Journal of Second Language Writing, 21*(2), 165–180. https://doi.org/10.1016/j.jslw.2012.03.002

Merkel, W. (2020). A case study of undergraduate L2 writers' concerns with source-based writing and plagiarism. *TESOL Quarterly,11(3), e00503*. https://doi.org/10.1002/tesj.503

Ministry of Education of the People's Republic of China. (2018). *2017 sees increase in number of Chinese students studying abroad and returning after overseas studies*. Retrieved from http://www.moe.gov.cn

Pearson. (2018). *Pearson test of English academic score Guide*. Available on https://pearsonpte.com/organizations/why-pte-academic/understand-our-scores/

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*, 111–129.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185–194.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*(2), 109–121.

Ruan, Z. (2016). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC, 48*(3), 327–340.

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31–53.

Sasaki, M. (1991). A comparison of two methods for detecting differential I item functioning in an ESL placement test. *Language Testing, 8*(2), 95–111.

Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strength and weaknesses: Assessing performance on an integrated writing tasks. *Language Assessment Quarterly, 10*(1), 73–95.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171–200.

Staples, S., Egber, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP development: Lexical bundles in the TOEFL iBT writing section. *English for Specific Purposes, 12*(3), 214–225.

Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. *Computer Assisted Language Learning, 30*(8), 864–883.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55.

Zheng, Y., & Mohammadi, S. (2013). An investigation into the writing construct(s) measured in Pearson Test of English Academic. *Dutch Journal of Applied Linguistics, 2*(1), 108–125.

Yu-Hua Chen is interested in the use of Corpus Approaches in SLA and Language Testing. Her research has been published in international journals such as Applied Linguistics. She co-developed the Academic Collocation List (ACL) and also led on the development of the CAWSE Corpus and Transcribear (an online transcription tool).

Ying Zheng is Associate Professor from Faculty of Arts and Humanities at the University of Southampton, UK. Ying received her PhD in Cognitive Studies from Queen's University Canada, specialising in Second Language Testing and Assessment.

Chapter 4
To Whom Do I Write? Chinese EFL Test-Takers' Conceptualisation and Construction of Their Audience in the Aptis Writing Test

Ying Chen and Xiaoxian Guan

Abstract The present study explored whether, and if so how, Chinese test-takers portrayed the audience and catered to the needs of two different audiences specified in the Aptis General Writing Task 4 employing think-aloud protocols (TAPs), questionnaire survey, and semi-structured interviews. Sixty six test-takers participated in the present study, including two employees, at B2 and C level respectively, and two groups of students with different levels of English proficiency, 32 high-proficient (at B2 or C level) and 32 low-proficient (at B1 level). Six test-takers, consisting of two employees and four students, were invited to take part in the think-aloud session. Recordings of test-takers' TAPs and interviews were transcribed, then double coded and analysed on the basis of Berkenkotter's (Coll Compos Commun 32:388–399, 1981) audience awareness coding scheme. Results from both quantitative and qualitative analyses indicated that explicitly specifying audiences in the writing tasks successfully weakened the understood testing context given that test-takers wrote to the audience specified in the task rather than to rater(s) for the mere purpose of getting a high score. But there were marked differences in audience conceptualisation and adaptation strategies between high-proficient and low-proficient writers as well as between the two audience conditions. While shedding light on the construct validity of the Aptis writing component, the findings of this study have important implications for EFL writing teaching, learning and assessment.

Keywords Audience awareness · Audience conceptulisation · Audience-related strategies · Contextualised writing prompts

Y. Chen (✉)
Ocean University of China, Qingdao, China
e-mail: cy@ouc.edu.cn

X. Guan
East China Normal University, Shanghai, China
e-mail: xxguan@english.ecnu.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_4

4.1 Context of the Study

In 2018, Council of Europe updated the 2001 scale by adding a number of new descriptors, one of which is the inclusion of formal correspondence (Council of Europe, 2018). The updated scale thus consists of both personal and formal correspondences which address two different registers, informal and formal. The use of writing tasks addressing different registers, however, is not widespread in large-scale standardised tests. One exception is the Aptis General Writing Task 4 designed by the British Council, which assesses test-takers' performance in both informal and formal conditions with the inclusion of register as one important aspect in writing task design and rating rubric (O'Sullivan & Dunlea, 2015). Aptis General, launched in August 2012, is the first variant within the Aptis Testing System which aims to offer test users flexible English language assessment options (Dunlea, 2018; O'Sullivan, 2012). The theoretical model behind Aptis General is the socio-cognitive model advanced by O'Sullivan (2011), O'Sullivan and Weir (2011), and Weir (2005) which pays special attention to the interaction between context and construct, and those contextual parameters such as topic, genre and the intended audience are explicitly specified to provide contextual information to test-takers.

Aptis General Writing, built around a series of theme-related activities, consists of four tasks which range from very basic form filling to quite complex email messages. Task 4 requires test-takers to write two emails in response to a short letter/notice: an informal email to a friend or close family member and a formal email to an unknown reader specified in the prompt. To validate this task, a number of studies, which are freely available on the British Council website (https://www.britishcouncil.org/exam/aptis/research/publications), have been carried out particularly on test development, rater training, refinement of the rating rubric, and cognitive validity. Given that the task requires test-takers to write to two different audiences, a close look at how test-takers conceptualise and construct the two target audiences is necessary. The present study undertakes the project of investigating how Chinese test-takers deal with audience demands by collecting both online think-aloud data and off-line questionnaire and interview data. The findings of this study could add to the theoretical validity of the Aptis General Writing Task 4, and enhance our understanding about the provision of contextual information in writing task design.

4.2 Literature Review

4.2.1 Audience and Audience Awareness: Definitions, Debates, and Disagreements

Researchers in the field of rhetoric and composition have long recognised that audience is imperative for successful communication. Audience, however, is an elusive concept, projecting different images to different writers (Porter, 1996). Broadly speaking, the images of audience can be grouped into two categories

(Park, 1982). One category is the actual people external to a text whom the writer must accommodate and the other is the audience implied in the text, or "audience-addressed" and "audience-invoked" respectively in Ede and Lunsford's (1984) term. Worse still, exact terminology may differ. For example, researchers have differentiated 'reader' from 'audience'. Due to the limited space, the present study will not distinguish 'reader' and 'audience' and will use the widely-accepted term 'audience' here. Those who are interested in the difference between 'reader' and audience' can refer to Park (1982) and Ede and Lunsford (1984). Given the fact that audience is ill-defined, the concept of audience awareness is unavoidably a slippery term, indicating various conceptualisations and interpretations within different rhetorical and compositional epistemologies. Relevant literature indicates that conceptualisations of audience and audience awareness have shifted over time, mainly under rhetorical, cognitive, and sociocultural frameworks (Magnifico, 2010).

Although researchers approach audience and audience awareness from diverse perspectives and with a long tradition of debates and disagreements, these studies do not suggest contradictions or controversies, but rather complement each other in providing a deeper and more comprehensive understanding of the complexities of audience and audience awareness (Kirsch & Roen, 1990). For example, Willey (1990) found that some writers addressed a real audience but invoked a fictional audience at various stages of the writing processes. In their more recent publication, Lunsford and Ede (2009) emphasised that "understanding the complexity of the writing process, audience awareness, and participation calls for more specific grounded, and nuanced analysis than the binary of addressed and invoked audiences can provide" (p. 56). As such, with an attempt to enrich the current scholarship of audience research, the present study explores how Chinese EFL learners conceptualise and construct audience in a standardised writing assessment context by adopting an open and multiple perspective of audience. The working definition of audience is thus established as the person or group to whom writers seek to convey their message through a written text, whether this being the writer him/herself, the abstract or fictional reader imagined by the writer, the people specified in the writing prompts whom writers are asked to write to, the teachers who will read the written texts and give writers feedback, or the raters who will evaluate writers' performance by awarding a score. Audience awareness refers to writers' understanding of the audience's characteristics, expectations and beliefs, and adjusting their message accordingly so as to effectively communicate with the target audience.

4.2.2 Previous Empirical Studies on Audience Awareness

Recognising the importance of context on communication and writing and also prompted by the increasing demand for authentic tasks in language tests, in late 1970s and early 1980s, an increasing number of scholars and practitioners (e.g., Crowhurst & Piche, 1979; Odell, 1981) strongly recommended providing rich contextual features in writing tasks. Empirical studies on the impact of providing these features, including specifying the target audiences, have thus been undertaken but yielding conflicting results. Some researchers found that there would be a change in the syntactic complexity when students write for different audiences (Crowhurst & Piche, 1979; Smith & Swan, 1978) and that skilled writers distinguish themselves from their less skilled counterparts in their ability to recognise and address the demands of different audiences (e.g., Ransdell & Levy, 1994; Zainuddin & Moore, 2003). Other studies, however, indicated no significant differences in the holistic scores of different audience conditions (e.g., McAndrew, 1982). Still others, such as Brossell (1983), revealed inconsistent evidence of the effects of contextual features in general and specific audience conditions on test takers' written products. Two reasons can help account for these contradictory results (Chen, 2014). On the one hand, definitions of audience remain elusive, hence researchers contextualise writing tasks in different ways. On the other hand, the majority of research adopted a product paradigm and employed different rating scales to evaluate written performance.

Despite the inconsistencies and even contradictions in empirical research, many scholars and teachers regard the ability to address different audiences appropriately as one importance indicator of the development of writing ability (Camp, 2012; Oppenheimer et al., 2017) and audience awareness is taken as a trait of the scoring scales in many writing programs at U.S. universities (Dryer, 2013).

In contrary to the abundance of research in L1 setting, few research has been undertaken to explore how L2 students conceptualise the target audience and how they deal with audience demands throughout the writing process (Cheng, 2005; Wong, 2005). Still scarce is research focusing on audience in L2 writing assessment prompts (Cho & Choi, 2018). To the authors' knowledge, there are only three published research studies investigating, to some extent, the influence of audience specifications on L2 test-takers' writing performance. The study conducted by Porter and O'Sullivan (1999) explored the effect of the age of the intended addressee on the written performance of Japanese university students and found that 'there is clear evidence to support the assertion that awareness of audience is an important factor affecting the scores achieved in these writing tasks' (p. 71). Chen (2014) developed a four-point holistic scale, ranging from 0 to 3, to quantify Chinese test-takers' sense of audience in texts written on tasks with three different audience conditions and found that some proficient students showed strong awareness of audience when the prompt specified the audience, thus making their essay more impressive. In a much recent study, Cho and Choi (2018) examined the effects of audience specification in a prompt on the quality of summary texts by L2 writers in a standardised testing context in America and found that these writers were able to take the needs of audience into consideration when rhetorical constraints were made clear to them. The researchers concluded that drawing ESL test takers' attention to an audience seemed to evoke their rhetorical awareness, leading them to adapt their writing to the needs of a specified audience. In designing a valid large-scale writing test, therefore, importance should be attached to the specification of an audience and a real need of communication, as Cho and Choi (2018) suggested.

It is worth noting that the few studies which have been undertaken in L2 writing assessment settings all investigate test-takers' audience awareness from the product perspective. The process perspective is therefore needed, which will enable us to get a detailed picture of how L2 writers, faced with cultural, rhetorical and linguistic demands simultaneously, cope with different types of audiences for different purposes in a testing context: to communicate with the audience specified in the writing task and, at the same time, to get a high score. Another point noteworthy of the present study is the use of email writing prompts rather than the widely-researched persuasive writing (Midgette et al., 2008). Here we do not mean persuasive writing is not important, but that by exploring test-takers' performance in other genres, such as email writing in this chapter, we hope to add richer findings to the current literature of the audience awareness research, especially on the intricated relationship between the rhetorical context and the assessment context.

4.3 Methodology

4.3.1 Research Questions

The research questions addressed in this study are as follows:

1. How do EFL test-takers conceptualise the target audiences specified in the Aptis Writing Test? What strategies do they adopt in order to meet the needs of the audience?
2. Are there any differences in the above-mentioned aspects depending on the audience condition (informal vs. formal) and EFL test-takers' writing proficiency?

To answer these research questions, this study adopted a within-subjects and between-subjects mixed design, with the audience condition (informal and formal) as the within-subjects factor and test-takers' writing proficiency the between-subjects factor. Moreover, this study avails itself of triangulation by probing into test-takers' writing process. To be specific, the present research employed think-aloud protocols (TAPs), questionnaires, and interviews to find those moments during the writing process where Chinese test-takers consider the needs of an audience.

4.3.2 Participants

The present study recruited test-takers through posters on university campuses and consultations with human resource department of prospective companies and organisations. Due to logistical constraints, this study selected 64 student test-takers, 21 boys and 45 girls, aged 19–21, who had been learning English for 9–12 years, and two employee test-takers, aged 35 and 36 respectively, both using English as part of their working language. Among the 64 students, 32 are low-proficient and 32 high-proficient, based on their College English Test (Band 4 and Band 6; CET-4 and CET-6 for short) scores, the most well-known national test of English proficiency of non-English major college students in China, and teacher evaluation. For the low-proficient group, roughly at B1 level, their CET-4 score is below 500 or CET-6 score below 450. The high-proficient group consists of non-English majors, at B2 or C level, whose CET-4 score is above 600 or CET-6 score above 550, and English majors recommended as proficient English learners by their teachers. Two employees include one male, at B2 level, working for a corporate company and one female, at C level, working for a government organisation as a Chinese-English translator.

To carry out TAPs, this study selected four students from the two proficiency groups, one boy and one girl from each group, representing a microcosm of the entire student participants in terms of English proficiency. The two low-proficient students, Kun and Chi, majored in Engineering and Arts respectively and at their second year of college by the time of data collection. Their high-proficient counterparts, Hui and Liu, were senior English majors and among top five in their class. The two employees, Wei and Zhang, also participated in the think-aloud session. For all the test-takers, Chinese is their first language.

4.3.3 Writing Prompts

As introduced earlier, Aptis General Writing Task 4 requires test-takers to write two emails to two different audiences (see O'Sullivan & Dunlea, 2015 for the detailed specifications of Task 4). In the present study, test-takers were invited to write to a friend and the customer service team about an online language course. For the first email, the situation is set as follows,

> You have joined an online language course. You've been learning the language for a few months. Your friend is interested in learning the language, too. Write an email to your friend recommending the course and give advice on how to improve it.
>
> Write up to 50 words. You have 10 minutes.

And the second prompt goes like this:

> You don't have time to finish your language course. Write an email to the customer service team telling them why you can't continue, and say what you think about the course.
>
> Write up to 150 words. You have 20 minutes.

4.3.4 Think-Aloud Protocols and Questionnaires

TAPs can gather on-line data on individuals' underlying cognitive processes, and is particularly useful in tracking the linear unfolding of their thought processes when they perform a task. Although concerns were expressed over the extent to which an individual's process of thinking aloud actually alters the cognitive processes required to carry out the given task (Stratman & Hamp-Lyons, 1994), it is "the best research tool for teasing out the cognitive processes that reveal themselves in what we call audience awareness" (Berkenkotter, 1981: 389). TAPs are, therefore, adopted to keep track of when and how frequently the considerations about audience entered the test-taker's mind, and to what extent audience-related considerations guided his/her rhetorical, organisational, and stylistic decisions. Drawing on relevant literature, this study worked out a set of instructions to maximise the validity of this tool, including careful selection of those participants who are willing to verblise their thoughts when composing, adequate training with both mathematics problems (Ericsson & Simon, 1993) and sample email writing tasks, use of any preferable language of the participants, and setting no time limit on the whole procedure.

For the analysis of the TAPs, the coding scheme was adapted from the Berkenkotter's (1981) framework and also drawn on Wong (2005). The finalised coding scheme includes four categories. The first category, *analysing/constructing an audience*, consists of four strategies, namely, analysing audience's features (A), identifying self with audience (i.e., role-playing) (SA), identifying audience with self (i.e., projecting) (AS), and creating the rhetorical context (RC). The second category, *goal setting and planning for a specific audience*, refers to generating audience-related goals (AG) or refinements of the plan (R). The third category, *evaluating content and style with regard to anticipated audience response*, is concerned with writers' evaluation of audience's possible response to content (C) or style (ST) of the text after the audience reads it. The final category, *revising for a specific audience*, includes three strategies, i.e., reviewing the text with audience in mind, making sentence- or discourse-level changes (D), or word-level changes (W) in accordance with audience's characteristics or needs.

In addition to TAPs, this study also conducted the questionnaire survey and semi-structured interview to identify whether and to what extent test-takers showed their awareness of the audience when writing two emails. The questionnaire, designed on the basis of the TAPs coding scheme, were piloted and revised several times before being put to use. The final version, worded in Chinese in order to facilitate comprehension, consisted of 16 items and targeted specifically at whether, and if so how, test-takers took into consideration of audiences' needs and expectations. Table 4.3 presented a translated version of the questionnaire used in the study.

The semi-structured interview required test-takers to briefly recall their writing processes, and then elaborate on their considerations of audience specified in the writing prompt including such questions as *(T)o whom do they write in the first and second emails? What are the characteristics of their audiences? Did they notice the differences between these two audiences? If yes, how did they adjust their emails to the different emails?* Meanwhile, test-takers were asked whether or not they had taken raters into consideration when writing two emails. In the present study, interview was implemented in Chinese to facilitate free discussion.

4.3.5 Data Collection and Analysis

Data was collected in two phases. In Phase I, six test-takers were asked to say everything they thought about while they performed the writing task. After the completion of think-aloud, they participated in the questionnaire survey and semi-structured interview. In Phase II, 60 test-takers sat for the normal Aptis writing test in one computer room. Immediately after the Aptis Writing Test, the questionnaire survey was administered and 15 test-takers were randomly sampled to participate in the following interview administered one by one.

In order to enlist participant cooperation, test-takers were first informed of the purpose of this project and their time, support, and contribution were acknowledged and highly valued. To ensure the validity of the TAPs, six test-takers were trained one by one and allowed to think-aloud as long as they would like. Recordings were later transcribed word for word, producing between four and seven pages of transcribed text for each protocol.

Data collected in the present study included six TAPs, 132 emails, 66 questionnaires, and 21 interviews. TAPs and interviews were transcribed and double coded. Both TAPs and interviews were coded on the basis of the Audience Awareness coding scheme mentioned earlier. Given that test-takers were asked to reflect on the differences between the two audiences during the interview session, this study developed one more category when coding the interview data. The inter-coder correlation coefficient for the TAPs data was 0.87 before the resolution, and that for the interview data was 0.84. Disagreements were resolved after discussion. Finally, quantitative data was processed and summarized descriptively with SPSS. Constrained by the small sample, however, inferential statistics could not be performed to examine whether there were significant differences between low- and high-proficient writers.

4.4 Results

4.4.1 Details of Think-Aloud Protocols

Table 4.1 describes the time spent, the number and percentage of Chinese and English words used in the TAPs by six test-takers across two audience conditions, which indicates they varied greatly in every aspect.

For the time spent, five among six test-takers spent much longer in Email 2 than Email 1, which is understandable considering the required length of Email 2 is three times as long as Email 1. There is one exception, though. Hui spent less time in Email 2 than in Email 1. A closer look at his protocols revealed that he spent more than 15 min to create a full picture for an online language course. Later he realised he "thought too much about the online language course", and had to cut it down because of the length limit. Another English major, Liu, was caught in a similar

Table 4.1 Details of TAPs

Candi.	Gender	Email 1				Email 2				Total	
		Time	Chinese	English	Word	Time	Chinese	English	Word	Time	Word
Kun	M	22:17	557 (51%)	525 (49%)	1082	55:20	1734 (64%)	979 (36%)	2713	77:37	3795
Chi	F	11:58	182 (34%)	352 (66%)	534	76:20	2030 (54%)	1738 (46%)	3768	88:18	4302
Hui	M	27:04	4 (0%)	1863 (100%)	1867	23:06	347 (20%)	1415 (80%)	1762	50:10	3629
Liu	F	23:45	600 (31%)	1348 (69%)	1948	38:10	29 (1%)	2872 (99%)	2901	61:55	4849
Wei	M	22:30	2565 (91%)	239 (9%)	2804	33:32	2713 (85%)	465 (15%)	3178	56:02	5982
Zhang	F	16:07	604 (52%)	553 (48%)	1157	57:48	2476 (59%)	1698 (41%)	4174	73:55	5331

situation. She pondered over what type of language course it was and later had to delete several sentences because she wrote too much.

As to the language used, four test-takers (Kun, Chi, Wei, Zhang) code-switched during the TAPs and used both Chinese and English when brainstorming, although the percentages differed. Wei predominantly used Chinese, whereas other three participants presented a roughly even distribution. Different from these four test-takers, Hui in Email 1 and Liu in Email 2 used English in their protocols except when retrieving specific words (such as 'account' and 'irresistible') from their long-term memory. Another tendency is that compared with Email 1, test-takers used more Chinese in Email 2, with two exceptions (Liu and Wei). The possible reason might be that Email 2 is relatively more complicated and cognitively demanding.

In terms of time length and the number of words produced in the protocols, the two low-proficient test-takers, Kun and Chi, spent the most time but generated almost the least amount of protocols. Hui and Liu spent almost the least time, while the two employees, Wei and Zhang, generated the most protocols. In addition, female test-takers generally spent more time and produced more protocols than their male counterparts.

4.4.2 Audience-Related Strategies Across the Two Audience Conditions

A glance at Table 4.2 indicated test-takers did show some awareness of audience. In both emails, test-takers analysed the audience's characteristics and needs mainly with the tactics of considering the audience's name and creating the rhetorical context for the audience, and also evaluated the content and style of their emails with regard to anticipated audience response.

In terms of test-takers' use of audience-awareness protocols across the two emails, some salient patterns pop out. In Email 1 test-takers most frequently analysed their audience, whereas in Email 2 they most frequently evaluated the content and style with regard to audience needs and expectations. Also in Email 2, test-takers tended to generate more audience-related goals than in Email 1. These differences might reside in the different complexity of the two emails as well as in the different cognitive load required from the test-takers. But for both emails, test-takers seldom generated sub-goals or refinements of the plan, which means test-takers rarely changed or improved their plans once made.

Among the four categories of audience-related protocols, 'revising for a specific audience' was least used in both emails although sufficient time was allowed. It might be due to Chinese test-takers' habitual practice of one-shot draft in tests and their assignments. This result is cross-checked by the questionnaire and interview data.

A relatively larger pool of the questionnaire data (Table 4.3) indicated a similar pattern in the use of audience-related strategies as the TAPs displayed. Generally

Table 4.2 Response frequencies for audience-related strategies in TAPs (N=*6*)

	Email 1										Email 2										
	Analys.				Goal.		Evaluat.		Revis.		Analys.				Goal.		Evaluat.		Revis.		
Cand.	A	SA	AS	RC	AG	R	C	ST	D	W	A	SA	AS	RC	AG	R	C	ST	D	W	Total
Kun	1	/	/	1	/	/	/	/	/	/	/	/	/	2	/	/	/	1	/	/	5
Chi	1	/	/	1	/	/	/	/	/	/	1	/	/	1	1	/	1	2	/	/	8
Hui	1	2	/	3	/	/	2	/	/	/	/	/	/	1	/	/	2	/	/	/	11
Liu	1	/	/	3	/	/	1	4	1	/	2	/	/	2	1	1	2	2	/	/	20
Wei	1	/	/	1	/	/	/	1	/	/	/	/	/	2	1	/	1	/	/	/	7
Zhang	1	2	/	4	1	/	/	1	/	/	/	/	/	2	3	/	8	3	/	2	27
Total	6	4	0	13	1	0	3	6	1	0	3	0	0	10	6	1	14	8	0	2	78
	23				1		9		1		13				7		22		2		

Table 4.3 Response frequencies for audience-related strategies in questionnaires (N = *66*)

Statements	Completely agree	Agree	Disagree	Completely disagree
01. Consider what to write to meet the task requirements.	54	12	/	/
02. Consider the purpose of the email.	54	10	2	/
03. Consider whom to write.	46	15	5	/
04. Regard the rater as audience.	5	10	26	25
05. Analyse the features of audience.	17	33	14	2
06. Consider how to meet audience's expectations.	13	24	26	3
07. Consider audience's possible responses.	7	20	22	17
08. Consider the register.	38	21	6	1
09. Always take audience into consideration while writing.	14	34	16	2
10. First write out my thoughts and then revise.	8	27	25	6
11. Check whether the writing purpose has been accomplished.	38	25	3	/
12. Review/revise the email from audience's perspective.	3	17	28	18
13. Evaluate the content.	2	12	29	23
14. Evaluate the register.	10	22	24	10
15. Make sentence- or discourse-level changes.	2	11	29	24
16. Make word-level changes.	25	32	9	/

speaking, almost all the test-takers tended to consider what to write to meet the task requirement, the purpose, and the audience of the email. Few regarded the rater as their audience. Most test-takers analysed the features (e.g., characteristics and needs) of their audience, considered how to meet their audience's expectations, and decided on the register to be used. But less than one thirds of test-takers took into consideration the audience's possible responses after reading their emails. While writing, most test-takers took their audience into consideration, although the degree varied from person to person. Only one third of test-takers reported they first wrote out their thoughts and then revised their emails. After writing, almost all the test-takers checked whether or not the purpose had been accomplished. But less than one third reviewed or revised their emails from the audience's perspective. Even less evaluated the content or the register. Almost all the test-takers made word-level changes rather than sentence- or discourse-level changes.

The in-depth interview data offered us the opportunity to capture rich data about test-takers' use of audience-related strategies, revealing that candidates' analysis of the audience varied in the two audience conditions. In Email 1, 13 out of 15 interviewees reported a comparatively rich analysis of the features of their friend, and constructed this friend as an imaginary close friend on the basis of the writer's real friend. As for the other two interviewees, one regarded himself as the friend and the

Table 4.4 Response frequencies for low and high-proficient writers' audience-related strategies in questionnaires (N = *12*)

	Low-proficient		High-proficient	
Statements	AGREE	DISAGREE	AGREE	DISAGREE
01. Consider what to write to meet the requirements.	12	/	12	/
02. Consider the purpose of the email.	11	1	12	/
03. Consider whom to write.	12	/	12	/
04. Regard the rater as audience.	4	8	/	12
05. Analyse the features of audience.	2	10	12	/
06. Consider how to meet audience's expectations.	/	12	10	2
07. Consider audience's possible response.	/	12	9	3
08. Consider the register.	5	7	12	/
09. Always take audience into consideration while writing.	2	10	12	/
10. First write out my thoughts and then revise.	/	12	/	12
11. Check whether the writing purpose has been accomplished.	/	12	12	/
12. Review/revise the email from audience's perspective.	/	12	6	6
13. Evaluate the content.	/	12	9	3
14. Evaluate the register.	/	12	10	2
15. Make sentence- or discourse-level changes.	/	12	8	4
16. Make word-level changes.	4	8	10	2

Note: 'AGREE' here refers to both 'Completely agree' and 'Agree' in the questionnaire, by combining the two options together. The same is true for "DISAGREE"

other considered the rater as the friend. But when writing Email 2, most interviewees took the customer service team for granted but seldom analysed their features.

4.4.3 *Effects of Writing Proficiency on Audience-Related Strategies*

To probe into whether or not low-proficient and high-proficient writers analysed and constructed audience differently, the present study further investigated 12 low-proficient writers whose score, ranging from 1.5 to 2.5, ranked the lowest among 66 test-takers, and their 12 counterparts whose score, ranging from 5 to 5.5, ranked the highest. Table 4.4 lists perceived audience-related activities by the low-proficient and high-proficient writers. A glance at Table 4.4 reveals that low-proficient and high-proficient writers analysed and/or constructed audience in similar but also dramatically different ways. Almost all the low-proficient and high-proficient writers considered what (content), who (audience) and why (purpose) before writing, despite one low-proficient writer reported he did not consider the purpose of the email. On the other hand, neither group admitted to first writing out their thoughts and then revising according to audience's characteristics or needs.

Aside from these similarities, low- and high-proficient writers exhibited distinctive differences in their audience analysis and construction strategies. To be specific, low-proficient writers did not analyse audience's features or characteristics but just wrote out their own thoughts. After finishing their draft, low-proficient writers did not check whether the writing purpose had been accomplished. Nor did they review or revise the email from the audience's perspective. Although one third of low-proficient writers reported that they made word-level changes in accordance with the audience's needs and characteristics, their revision was mainly concerned with spelling. They did not make any major changes at the syntactic or discourse level. As a matter of fact, low-proficient writers either struggled to generate contents or tried hard to translate their thoughts into English, hardly thinking beyond the content of their emails. For instance, when Kun spoke aloud his writing processes, he experienced great trouble in generating contents especially in the second email by laboriously searching for "what to write next". There was no surprise then when low-proficient writers paid scarce attention to the audience specified in the writing prompts.

Different from low-proficient writers, high-proficient writers were at great ease when translating their thoughts into English and many even planned their writing in English. All of them analysed the audience's features, and the overwhelming majority endeavoured to meet audience's expectations and evaluate the content with regard to anticipated audience response. When working on the second email, some high-proficient test-takers took into consideration the customer service team's response and expectations in order to make their writing more convincing. For example, when Zhang brainstormed the excuses for not having time to finish the language course, she evaluated the persuasiveness of her excuses.

> Why can't I finish this course? I am working in one government institution. I usually have a lot of work by the end of the year and have to stay up late. So I do not have time to learn online. This reason seems convincing enough.

Once the emails were drafted, high-proficient writers considered audience's possible response and checked whether or not the writing purpose had been accomplished. One half reviewed or revised the email from audience's perspective and the majority made word-level changes in accordance with audiences' characteristics and needs. Some even made sentence- or discourse-level changes such as resequencing the order of the content especially when test-takers compiled their reasons for lacking the time to finish their language course in the second email, deleting some content which test-takers regarded as irrelevant or inappropriate, etc.

For example, the TAPs data indicated that different from low-proficient writers, Zhang, a high-proficient writer, exhibited the greatest frequency and widest distribution of audience-related strategies. One remarkable feature of Zhang's audience-related protocols is that she frequently evaluated audience response to her content, which accounts for almost one third of the total instances.

The above differences between low-proficient and high-proficient writers in their audience analysis and construction strategies can also be echoed by the evidence gained from the in-depth interview data, as illustrated by the following excerpts.

> I just wrote to a friend, quite close friend. I didn't write her name, either. Actually, I didn't analyse the specific features of the audience. (Test-taker 17, low-proficient)

> After I finished my writing, I usually check whether or not the writing purpose has been fulfilled. In this case, for the first email, I'll ask myself "did I recommend the course to my friend?". And for the second email, I'll look at whether the audience can understand what I've written and how will they react after reading my email. (Test-taker 26, high-proficient)

To further investigate how the low-proficient writers and the high-proficient writers dealt with both the communicative context presented in the writing task and the testing context embedded in this study, we looked specifically at test-takers' responses in Statements 04 and 05 and also explored the reasons behind. Surprisingly, two thirds of the low-proficient writers and all the high-proficient claimed they did not regard the rater as audience even though their performance would be rated by professional raters. The reason is that the writing task used in the present study clearly set the communicative context for the writers, that is, the test-taker joined an online language course and needed to write one email to their friend recommending the course and write another email to the customer service team to discontinue this course. Thanks to the authenticity of the context, the test-takers easily entered into and submerged themselves in this context and communicated with the two audiences like they did in real life. For example, Test-taker 17, high-proficient, stated during the interview, "No, I did not take the rater into consideration. In fact, I totally forgot this was a test!" Meanwhile, the high-proficient writers tended to analyse the features of audience such as "my friend Sara likes English very much and she dreams to improve it" (Test-taker 49, high-proficient).

The majority of the low-proficient writers also put themselves in the communicative context specified in the writing task. However, four among 12 low-proficient writers still wrote to the rater and ten did not analyse the features of the audience. There are two reasons which help account for this result. First, low-proficient writers seldom use English to communicate with their friends. Their English writing is mostly the argumentative writing with the only purpose of fulfilling the homework or testing requirements, as the questionnaire data showed, where audience is always the teacher or the rater. Second, low-proficient writers are characterised by "knowledge-telling" (Bereiter & Scardamalia, 1987) and could not decenter from his or her own perceptions of reality to consider the needs of the audience. As a matter of fact, some text-takers projected themselves as the audience. For example, Test-taker 37 explicitly reported "I imagined myself as my friend". Therefore, even the low-proficient writers were asked to write to their friend or the custom service team, they failed to analyse the features of the audience and just wrote what they thought.

It should be noted that although the test-takers claimed they did not write to the rater, most of them did take the rater into consideration and tended to use safe expressions and structures as the interview data indicated. For instance, Test-taker 57, low-proficient, remarked “after writing I checked whether there were typos or grammatical mistakes” and Test-taker 42, high-proficient, stated “logic and accuracy were my top priorities”.

4.5 Discussion

4.5.1 EFL Test-Takers’ Audience Awareness Across the Audience Conditions

Different from Chen (2014), most test-takers in the present study paid great attention to the audience specified in the Aptis Writing Test. Before writing, they tended to think about both the purpose and the audience of their writing, and also created a rhetorical context for their writing. Meanwhile, they tend to analyse the audience’s features and adapt their writing to different audience conditions, as both the questionnaire data and the displayed. When writing to their friend (the informal audience), test-takers more frequently analysed the characteristics of the audience compared with their writing to the customer service team (the formal audience). This might be because an identifiable audience, such as a friend, makes it easier to keep that audience in mind (Rubin & O’Looney, 1990). A remote and distant audience, the customer service team in this case, however, is just a far-away existence and makes the dialogue between the writer and the audience less accessible (Cohen & Riel, 1989). So many test-takers resorted to the traditional way of writing to express their own views or opinions regardless of the needs and/or expectations of the customer service team.

Another finding is that few test-takers regarded raters or teachers as the audience. Specifying the audience in the Aptis Writing Test, therefore, does not create a dilemma for Chinese test-takers, despite the fact that the writing test itself entails a kind of context, in which test-takers are aware that they are asked to write a text to be evaluated later by a certain rater in a certain way. Nevertheless, the influence of the testing context cannot be overlooked. Although most test-takers did not write to markers or teachers, they indeed tended to use ‘safe’ phrases or structures for the sake of accuracy and make word-level rather than sentence- or discourse-level changes even with the convenience of the computer facility. This holds true for both informal and formal audience conditions. To strengthen the positive influence of the communicative context presented in the writing task, test developers need to emphasise the communicative adequacy in their rating rubric.

4.5.2 Effects of Test-Takers' Writing Proficiency on Their Audience Awareness

In line with the previous research (Black, 1989; Kirsch, 1991; Wong, 2005), the present study also found that audience awareness is a robust indicator of writing proficiency and differentiates between high and low achievers. High-proficient writers employed relatively rich and varied strategies to conceptualise their audience and cater to their needs, such as analysing the needs of the audience and evaluating content and style with regard to anticipated audience responses. Their low-proficient counterparts, however, focused almost only on the content of their emails and rarely considered the needs and expectations of the audience. Such finding is understandable and reasonable because low-proficient writers face more struggle when transforming the inner language to written text (Gregg et al., 1996).

Although test-takers in the present study took into consideration of the needs and expectations of the audiences, low- and high-proficient writers conceptualised and constructed their audiences in markedly different ways. Generally speaking, low-proficient writers had great trouble in completing the task and rarely thought beyond the content of their emails. Compared with low-proficient test-takers' topic-bound protocols, their high-proficient counterparts adopted rich and varied audience conceptualisation and adaptation strategies. Among these strategies, the appropriate use of the evaluation strategy could help distinguish experienced writers from poor writers.

Another finding common in the testing setting is that for both audience conditions, test-takers seldom generated sub-goals or refinements of the plan, which means they rarely changed or improved their plans once made. This might be due to the often-used one-shot writing test which deprives test-takers of the opportunity to revise their essays due to test time limit and the pen-and-paper test format. Test-takers are thus accustomed to the first draft practice for the sake of cleanness of their written scripts. Even when writing in computer-delivered mode in the present study, they still hardly make major changes beyond the sentence level. Such practice goes contrary to some research suggesting writers should get their ideas down first before they can be expected to revise toward audience needs and expectations (Frank, 1992; Roen & Willey, 1988; Rubin & O'Looney, 1990) and advocating for considering the audience needs during the process of revision (Fitzgerald & Stamm, 1990; Midgette et al., 2008). Further research is, therefore, needed to explore whether or not test-takers who are familiar with the process writing approach, if encouraged to, revise their written scripts in accordance with the different audiences.

4.5.3 Contextulisation of EFL Writing Learning, Teaching, and Assessment Prompts

Writing is a communicative act. Just as Hamp-Lyons and Kroll (1997: 8) put it, writing is "an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience". Writing prompts should, therefore, better reflect such social nature. But the importance of the context is not well implemented in many large-scale writing assessments as well as classroom instructions, both in L1 (Behizadeh & Pang, 2016; Olinghouse et al., 2012) and L2 (Al-Mohammadi & Derbel, 2015; Chen, 2014) settings. This de-contexulisation of writing prompts confine writers to a cognitive and social vacuum and thus greatly diminishes audience awareness (Cohen & Riel, 1989). To make sure test-takers better display their writing performance, it is advisable and imperative, therefore, to contextulise writing prompts in both writing instruction and assessment.

Aside from the provision of contextualised writing prompts, another caveat both scholars and practitioners should bear in mind is that test-takers' performances should be evaluated in accordance with the desirable criteria in the context specified in the writing prompts. Otherwise, the whole endeavor might be jeopardized. For instance, working with 360 undergraduate education majors at two universities in America, Brossell (1983) tested the hypothesis that writing tasks with full contextual features would elicit higher quality essays than other less specified ones. Instead of evaluating the communicative effect from the target audience's perspective, the researcher informed student writers that their performance would be judged by trained raters, which might reinforce the typical assessment situation and thus weaken the effects of the contextualised writing task, thereby blurring the results. Currently, audience awareness is beginning to be used as an important trait on scoring rubrics for native (Oppenheimer et al., 2017) as well as nonnative writers (Cho & Choi, 2018). Although unestablished in terms of both validity and reliability in the implementation stage, this trait, undoubtedly worthwhile, calls for more in-depth research in the near future so as to broaden the construct of language proficiency and provide a valuable source of hypotheses for understanding an interactionalist approach to construct definition (Jin, 2017).

4.6 Conclusion

Both the online think-aloud data and off-line questionnaire and interview data collected in this study showed that awareness of the audience's needs and expectations is an important factor affecting test-takers' decisions on how and what to present when they compose their writing. Although whether, and if so how, using different audience-related strategies affects the quality of writing is still an open question, the findings of the present study can tentatively lend support to the necessity of the inclusion of both informal and formal registers in the writing test given that test-takers adopted different audience conceptualisation and construction strategies when writing to different audiences. Meanwhile, specifying the contextual features engaged by the test-takers successfully weakens the understood testing context given that test-takers write to the audience specified in the writing prompts rather than write to raters for the mere purpose of getting a high score. The contextualisation of writing prompts is, therefore, of significant importance in task design which can have the potential to promote positive washback in teaching and learning.

Although the present study employed multiple research methods to triangulate the data, certain limitations were inevitable from the outset. Interpreting the results, therefore, must acknowledge the following limitations. First and foremost, except for two employees, 64 student test-takers were recruited from one university due to practical limitations of time and manpower. Moreover, many critical factors which might influence the process or product of EFL writing were not included, such as candidates' motivation for writing, L1 competence, etc. A large sample of candidates from a diverse of L1 backgrounds is needed to further explore the strategies used by test-takers to adapt to different audiences specified in the writing prompts.

Second, TAPs were used to investigate test-takers' audience conceptualisation and adaptation strategies. Although TAPs are very informative, they could not be exempt from the main criticisms, i.e., veridicality and reactivity (Ericsson & Simon, 1993; Stratman & Hamp-Lyons, 1994). Future research can adopt other methods with the help of modern technology, such as key board tracking and eye-tracking to better track test-takers' writing processes, and event-related potentials (ERPs) to measure how test-takers' brains response differently when writing to two different audiences.

Acknowledgements We thank the editors Professor Liz Hamp-Lyons and Professor Yan Jin, and external reviewers for their insightful comments on an earlier version of this chapter. We also deeply appreciate the generous help of the British Council throughout the project. The views expressed in this chapter do not necessarily reflect those of the British Council. The authors are responsible for all the statements and any errors.

Funding This research was funded by the Aptis Research Grant (East Asia) and supported by the National Planning Office for Philosophy and Social Sciences of the People's Republic of China [20BYY108].

References

Al-Mohammadi, S. A., & Derbel, E. (2015). To whom do we write? Audience in EFL composition classes. In R. Al-Mahrooqi (Ed.), *Methodologies for effective writing instruction in EFL and ESL classrooms* (pp. 197–208). IGI Global.

Behizadeh, N., & Pang, M. E. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing, 29*, 25–41.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Erlbaum.

Berkenkotter, C. (1981). Understanding a writer's awareness of audience. *College Composition and Communication, 32*, 388–399.

Black, K. (1989). Audience analysis and persuasive writing at the college level. *Research in the Teaching of English, 23*(3), 231–253.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English, 45*(2), 165–173.

Camp, H. (2012). The psychology of writing development – And its implications for assessment. *Assessing Writing, 17*, 92–105.

Chen, Y. (2014). *An investigation into the context validity of EFL writing tests*. China Renmin University Press.

Cheng, F. (2005). Audience strategies used by EFL college writers. *Journal of Pan-Pacific Association of Applied Linguistics (PAAL), 9*(2), 209–225.

Cho, Y., & Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing, 37*, 25–38.

Cohen, M., & Riel, M. (1989). The effect of distant audiences on students' writing. *American Educational Research Journal, 26*(2), 143–159.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Available online: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Crowhurst, M., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English, 13*(2), 101–109.

Dryer, D. B. (2013). Scaling writing ability: A corpus-driven inquiry. *Written Communication, 30*(1), 3–35.

Dunlea, J. (2018). Aptis. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). Wiley.

Ede, L., & Lunsford, A. (1984). Audience address/audience invoked: the role of audience in composition theory and pedagogy. *College Composition and Communication, 35*, 155–171.

Ericsson, K. A., & Simon, H. A. 1993. *Protocol analysis: Verbal report as data* (Revised ed.). The Massachusetts Institute of Technology.

Fitzgerald, J., & Stamm, C. (1990). Effects of group conferences on first graders' revision in writing. *Written communication, 7*(1), 96–135.

Frank, L. A. (1992). Writing to be read: Young writers' ability to demonstrate audience awareness when evaluated by their readers. *Research in the Teaching of English, 26*(3), 277–298.

Gregg, N., Sigalas, S. A., Hoy, C., Wisenbaker, J., & McKinley, C. (1996). Sense of audience and the adult writer: A study across competence levels. *Reading and Writing, 8*(1), 121–137.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community, and assessment* (TOEFL monograph series report no. 5). Educational Testing Service.

Jin, Y. (2017). Construct and content in context: Implications for language learning, teaching and assessment in China. *Language Testing in Asia, 7*(1), 12.

Kirsch, G. (1991). Writing up and down the social ladder: A study of experienced writers composing for contrasting audiences. *Research in the Teaching of English, 25*(1), 33–53.

Kirsch, G., & Roen, D. (1990). Introduction: Theories and research on audience in written communication. In G. Kirsh & D. H. Roen (Eds.), *A sense of audience in written communication* (pp. 25–39). Sage.

Lunsford, A. A., & Ede, L. (2009). Among the audience: On audience in an age of new literacies. In M. E. Weiser, B. Fehler, & A. M. González (Eds.), *Engaging audience: Writing in an age of new literacies* (pp. 42–69). National Council of Teachers.

Magnifico, A. M. (2010). Writing for whom? Cognition, motivation, and a writer's audience. *Educational Psychologist, 45*(3), 167–184.

McAndrew, D. A. (1982). *The effects of an assigned rhetorical context on the syntax and holistic quality of the writing of first year college students*. Unpublished dissertation at State University of New York.

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and the audience awareness goals for revision on the persuasive essays of fifth and eighth grade students. *Reading and Writing, 21*, 131–151.

O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *Routledge handbook of applied linguistics*. Routledge.

O'Sullivan, B. (2012). *Aptis test development approach* (Aptis technical report (ATR-1)). British Council.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis General technical manual version 1.0* (Technical Report TR/2015/005). British Council.

O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language testing: Theory and practice*. Palgrave.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), *The nature and measurement of competency in English* (pp. 95–138). National Council of Teachers of English.

Olinghouse, N. G., Zheng, J., & Morlock, L. (2012). State writing assessment: Inclusion of motivational factors in writing tasks. *Reading & Writing Quarterly, 28*(1), 97–119.

Oppenheimer, D., Zaromb, F., Pomerantz, J. R., Williams, J. C., & Park, Y. S. (2017). Improvement of writing skills during college: A multi-year cross-sectional and longitudinal study of undergraduate writing performance. *Assessing Writing, 32*, 12–27.

Park, D. B. (1982). The meanings of "audience". *College English, 44*(3), 247–257.

Porter, J. E. (1996). Audience. In T. Enos (Ed.), *Encyclopedia of rhetoric and composition: Communication from ancient times to the information age* (pp. 42–49). Garland.

Porter, D., & O'Sullivan, B. (1999). The effect of audience age on measured written performance. *System, 27*(1), 65–77.

Ransdell, S. E., & Levy, C. M. (1994). Writing as process and product: The impact of tool, genre, audience knowledge, and writer expertise. *Computers in Human Behavior, 10*(4), 511–527.

Roen, D. H., & Willey, R. J. (1988). The effects of audience awareness on drafting and revising. *Research in the Teaching of English, 22*(1), 75–88.

Rubin, D. L., & O'Looney, J. (1990). Facilitation of audience awareness: Revision processes of basic writers. In G. Kirsh & D. H. Roen (Eds.), *A sense of audience in written communication* (pp. 280–292). Sage.

Smith, W. L., & Swan, M. B. (1978). Adjusting syntactic structures to varied levels of audience. *Journal of Experimental Education, 46*, 29–34.

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89–111). Sage.

Weir, C. J. (2005). *Language Testing and Validation: An evidence-based approach*. Palgrave.

Willey, R. J. (1990). Pre-classical roots of the addressed/invoked dichotomy of audience. In G. Kirsh & D. H. Roen (Eds.), *A sense of audience in written communication* (pp. 25–39). Sage.

Wong, A. T. Y. (2005). Writers' mental representations of the intended audience and of the rhetorical purpose for writing and the strategies that they employed when they compose. *System, 33*, 29–47.

Zainuddin, H., & Moore, R. A. (2003). Audience awareness in L1 and L2 composing of bilingual writers. *TESL-EJ, 7*(1), A-2.

Ying Chen is Associate Professor in Applied Linguistics at Ocean University of China. She obtained her PhD from Shanghai Jiao Tong University, China. Her research interests are in writing assessment, vocabulary acquisition, pragmatics assessment, and second language development. She has led several funded projects on second language assessment.

Xiaoxian Guan, senior lecturer at East China Normal University, has acquired her doctorate degree in Linguistics and Applied Linguistics from Shanghai Jiao Tong University. Her research interests include language testing and assessment. She is currently in charge of the school-based English proficiency test of her university.

Chapter 5
Strategy Use and Performance in EFL Writing of Taiwanese Learners

Naihsin Li

Abstract The competence to write in English has become an essential ability especially as English has gained the status of a lingua franca. However, many EFL learners in Taiwan struggle to write a "well-organized and generally coherent essay that demonstrates sufficient control of vocabulary and sentence structures" as described in the GEPT High-Intermediate Writing Test Rating Scale. Drawing on a pool of 470 EFL learners who took the GEPT High-Intermediate Writing Test (roughly equivalent to the CEFR B2 level), this study seeks to examine learning strategies employed by Taiwanese EFL learners, their overall writing performance, and specific areas of difficulty they face. Their learning strategies were investigated by use of a writing strategy questionnaire based on Oxford's (Language learning strategies: what every teacher should know. Heinle & Heinle, Boston, 1990) taxonomy of learning strategies. Statistical analyses were employed to explore the relationships between strategy use and writing performance, and a comparison was made between the successful and the unsuccessful candidates of the writing test. Furthermore, we characterized Taiwanese EFL learners' writing difficulties, as reflected by their writing errors, by analyzing a sample of writing scripts randomly selected from both the successful candidate and unsuccessful candidate groups. The writing difficulties were linked to the learning strategies on which EFL writing instruction should focus. Findings of this study are of pedagogical significance to writing instructors.

Keywords EFL writing · GEPT · Learning strategy use · Error analysis

N. Li (✉)
The Language Training and Testing Center, Taipei City, Chinese Taipei
e-mail: naihsinli@lttc.ntu.edu.tw

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_5

5.1 Introduction

The ability to communicate, either orally or in writing, through English is highly prized in this globalized world. However, EFL learners in Taiwan have been struggling specifically with writing in English. The annual score data summaries of the General English Proficiency Test (GEPT), a level-based criterion-referenced English testing system tailored to Taiwanese EFL learners, have revealed that Taiwanese EFL learners consistently score higher in the speaking component than in the writing component (The Language Training & Testing Center [LTTC], n.d.). Statistics from the Advanced Subjects Test (AST), one of the college entrance examinations administered by the College Entrance Examination Center in Taiwan, show that in the past 10 years, median scores for English composition lie between 6 and 8, out of a total score of 20. Data from 2018 to 2019 further revealed that around 10% of all test-takers received a score of 0 (College Entrance Examination Center [CEEC], n.d.). To help EFL learners in Taiwan improve their ability to write in English and overcome their writing difficulties, efforts should be taken to understand learners' learning processes and their effects on writing performance.

Language learning processes can be understood partly through an investigation into the learning strategies language learners employ to acquire particular language skills. Learning strategies, as defined in Oxford (1990), are "specifications taken by the learners to make learning easier, faster, more enjoyable, more self-directed, more effective, and more transferable to new situations" (p. 8). Moreover, learning strategy use is closely associated with performance in second language learning as the use of strategies has been recognized as one important source of individual variation in language learning outcome. In particular, it is believed that competent learners are effective because of special learner techniques or strategies (Bai et al., 2014; Naiman et al., 1978; O'Malley & Chamot, 1990; Rubin, 1981; Wong & Nunan, 2011; Wu, 2008).

The association between the use of learning strategies and performance in EFL writing has been explored in a number of studies (De Silva, 2015; Huang & Chen, 2006; Lei, 2016; Victori, 1999; Yang, 2013). The investigation into learning strategy use benefits from adopting a taxonomic analysis of learning strategies, which can reflect the underlying cognitive processes that are involved in a learning task. While a number of taxonomic systems have been proposed in the literature (e.g., O'Malley & Chamot, 1990; Oxford, 1990), their focus on the general processes of foreign language learning may not adequately account for the real processes involved in a specific language skill or task (Hsiao & Oxford, 2002). This emphasizes the need for a skill-based investigation of learning strategies, which can more readily contribute to our understanding of the relevant learning processes.

Kao and Reynolds (2017) may have been the first to investigate the taxonomy of learning strategies involved in EFL writing as a prerequisite to establishing the link between strategy use and writing performance. Their study identified four types of strategy use: cognitive/preparation, compensation/supporting, affective, and social/textual interaction. It further showed that cognitive/preparation strategy use was positively correlated with writing ability and negatively correlated with writing difficulty. However, some questions arise from their research methods and findings. First of all, the absence of metacognitive strategy use is worthy of note because previous studies have shown that metacognitive strategy use is an important facilitator of learning performance (Anderson, 2005; O'Malley & Chamot, 1990; Victori, 1999). Secondly, the authors suggested that the results might be biased partly by the homogeneity of the subject pool. Furthermore, it should be noted that the participants' writing ability and writing difficulties were self-rated; therefore, the measures were likely to be the reflection of students' self-confidence, rather than their real performance.

This study aims to examine Taiwanese EFL learner's use of learning strategies for EFL writing and its link with their EFL writing performance in a high-stakes writing test (i.e., the GEPT High-Intermediate Writing Test) using a larger subject pool and by taking both quantitative and qualitative approaches. The EFL learners' writing performance is assessed on a more objective measure of writing ability—the score they received in the writing test. In addition to using the quantitative measure as an indicator of writing performance, this study also examines Taiwanese EFL learners' writing difficulties by conducting error analyses on a sample of writing scripts. Comparisons have been made between more proficient writers (i.e., successful candidates of the writing test) and less proficient writers (i.e., unsuccessful candidates of the test) in terms of learning strategy use and writing difficulties. It is expected that a synthesis of the findings will have implications for writing instruction.

The following research questions were addressed in this study:

1. What are the learning strategies used by EFL learners in Taiwan to learn to write in English? Are there any differences between the more proficient and the less proficient EFL writers?
2. What are the relationships between learners' learning strategy use and writing performance? Are there any differences between the more proficient and the less proficient EFL writers?
3. What are the writing difficulties of the less proficient EFL writers in comparison with the more proficient ones?

5.2 Literature Review

5.2.1 Taxonomic Systems of Learning Strategies

Learning strategies can be classified into several broad categories, and different taxonomies have been proposed in the literature. For example, O'Malley and Chamot (1990) differentiate three types of learning strategies in their model: metacognitive, cognitive, and socio-affective strategy uses. Cognitive strategies involve practicing the language to be learned, while metacognitive strategies are higher-order executive skills that help learners regulate their learning processes by planning, monitoring and evaluating their learning. The socio-affective strategies concern the interaction with others or the skills necessary to regulate personal emotions, such as anxiety or self-confidence. Another widely-cited model is that proposed by Oxford (1990), in which there are six strategy groups. In addition to the core strategy types proposed in O'Malley and Chamot (1990), Oxford distinguishes two additional strategy types from cognitive strategies: memory and compensatory strategy use. Memory strategy use involves remembering and retrieving new information, while compensatory strategy use involves using the language despite knowledge gaps. Furthermore, she also recognizes the individual impacts of social factors and affective factors on language learning and regards them as independent primary strategy groups. The validity of these different taxonomic systems has been evaluated in Hsiao and Oxford (2002); nevertheless, the results suggest that a skill-based investigation of learning strategy use can better reflect the underlying learning processes. While Kao and Reynolds (2017) may have been the first to investigate the classification of learning strategy use in relation to EFL writing, the lack of metacognitive strategy in their findings warrants further exploration into this issue.

5.2.2 Strategy Use and Writing Performance

The link between strategy use and performance in EFL writing has been established in several studies, which compare strategy use between EFL writers with different levels of writing skills. For example, by using think-aloud and interview data, Yang (2002) and Victori (1999) found considerable differences between skilled writers and less skilled writers in their application of text planning, monitoring and evaluation strategies. Skilled writers were found to spend more time planning their writing and tended to revise and enhance their writing at the global level, targeting such aspects as the coherence and unity of their texts. In addition to differences in the writing processes, the two groups showed attitudinal differences toward the writing activity (Lei, 2016; Victori, 1999; Wong & Nunan, 2011). Specifically, skilled writers tended to be more committed to the task and displayed a greater degree of autonomy. For example, Lei (2016) found that less skilled writers often regarded themselves as task-doers and the writing task as a task to be fulfilled, while skilled writers viewed writing as a way of communication and themselves as both a language learner and an author. The two groups also differed remarkably in the aspects of noticing language use, imitating good writers (e.g., teachers or skilled peers) in their strategy use, maximizing their chances of practicing English writing, and setting goals related to developing writing ability.

Other studies explore the issue by means of quantitative approaches, and the discussion of strategy use is grounded in a taxonomic system of learning strategies, which provides additional insights into the interrelationships of different strategy types. For example, Yang (2013) and Huang and Chen (2006) both employed O'Malley and Chamot's model to examine the use of learning strategies among Chinese-speaking EFL students. Yang (2013) found that EFL novice writers at the secondary level of education most often applied affective strategies and memorization, but least often applied strategies such as revision, social interaction and planning. Examining a different cohort, Huang and Chen (2006) demonstrated that Chinese college students most often applied cognitive strategies, such as use of synonyms and checking errors, but least often applied socio-affective strategies, such as asking for the teacher's help or asking for comments from their peers. A path analysis further showed that while all three strategy types (i.e., metacognitive, cognitive and socio-affective strategy uses) had a direct impact on writing, writing performance was particularly related to cognitive and metacognitive strategy use. Similar findings were observed in Kao and Reynolds (2017), which demonstrated a significant link between cognitive strategy use and self-rated measures of writing performance.

As many studies have already explored writing performance in terms of test scores or self-reporting, this study aims to add granularity by further characterizing EFL learners' writing difficulties through examination of EFL errors, which potentially reflect their knowledge of the language system and their difficulties with using the language. Tseng (2016) conducted a similar study examining the writing errors of Taiwanese EFL learners who took the GEPT Intermediate Writing Test (roughly equivalent to CEFR B1 level). Writing errors were examined and coded at different linguistic and textual levels, including word choice, within-sentence grammaticality, text coherence and unity, and rhetorical structure. Her study has shown that less proficient writers demonstrate specific error patterns that are less observed among the more proficient writers. Such findings may instruct writing instructors as to which aspects of writing to focus on when they are helping less proficient writers improve their writing skills. Adapting Tseng's (2016) error coding framework, this study similarly examines the writing difficulties of our target population.

5.2.3 The GEPT Writing Test

The General English Proficiency Test (GEPT), developed and administered by the Language Training and Testing Center (LTTC) in Taiwan, is a criterion-referenced EFL testing system targeting English learners in Taiwan from junior high school upwards. The test is offered at five levels: Elementary, Intermediate, High-Intermediate, Advanced and Superior, which are roughly equivalent to CEFR A2, B1, B2, C1 and C2, respectively. Test-takers can register for any level that fits their needs. GEPT scores are used for a variety of purposes including job applications, university admission, placement and graduation (Kunnan & Wu, 2009), and it also has considerable impacts on English language teaching and testing in Taiwan (Wu, 2012).

Writing is one component included in all levels of the GEPT, though the test is in different formats depending on the writing ability targeted at each specific level. For example, the writing tasks at the Intermediate and High-Intermediate levels contain Chinese-English translation and guided writing, but those at the higher proficiency levels involve integrated writing. The alignments between the different levels of the GEPT writing test and the CEFR have been validated in Knoch and Frost (2016). In addition, the validity of the GEPT writing tests has been examined and established in a number of studies (Chan et al., 2014; Kunnan & Carr, 2015; LTTC, 2003; Qian, 2014; Weir et al., 2013; Yu & Lin, 2014) and they have been used extensively to collect and explore Taiwanese EFL learners' writing processes (Lin, 2019), use of English language (Cheung et al., 2010) or their writing difficulties (Kuo, 2005; Sun, 2018; Tseng, 2016; Wu, 2016).

This study aims to examine which types of strategies are employed by Taiwanese EFL learners to learn to write in English and how their use of learning strategies is related to their EFL writing performance, as reflected by the writing test scores and writing errors. A synthesis of the findings will have implications for the learning strategies on which EFL writing instruction should focus.

5.3 Method

5.3.1 Participants

Participants were 479 test-takers of the GEPT High-Intermediate Writing Test sampled from four test sites in different regions of Taiwan. The participants were asked to fill out a learning strategy questionnaire immediately after the writing test. They were informed that the purpose of the questionnaire was to understand their use of strategies to learn to write in English and that it was only for research purposes. They were also guaranteed that their willingness to respond to the questionnaire would not affect their test score.

Among the respondents, 2 did not make any responses and 7 did not complete substantial portions of the questionnaire, and thus they were excluded from the subject pool. In all, questionnaires from 470 respondents were considered in the subsequent analyses. The sample demonstrated a representative sample of the population who took part in this specific GEPT test in terms of sex and status. Around 60% ($n = 280$) of respondents were female. A great majority of the respondents were high school students ($n = 290$, 61.70%), 24.04% ($n = 113$) were college and graduate students, while the remaining 14.26% ($n = 67$) were non-students.

5.3.2 Instruments

5.3.2.1 The GEPT High-Intermediate Writing Test-Guided Writing

The GEPT High-Intermediate Writing Test is composed of two subtests: Chinese-English translation and Guided Writing. Due to our interest in essay writing performance, this study looked only at the Guided Writing subtest. EFL learners at this level are characterized as being able to write about topics related to daily life and express their personal viewpoints on current events. Therefore, in this subtest, the test-takers are required to write a 150 to 180-word essay to express opinions about a specific issue and support their ideas with examples.

In terms of rating, each writing script was double rated holistically based on a 6-point rating scale, ranging from 0 to 5. In cases where a discrepancy between the two marks was greater than 2 band scores, a third rater determined the final score. A well-organized piece of writing which adequately addressed the topic and task and demonstrated sufficient control of vocabulary and sentence structure received a score of 4 or above and reached the passing standard of this subtest. However, pieces which displayed less satisfactory coherence and more limited and incorrect use of vocabulary and structures received a score of 3 or below. A sample test and sample essays for different band scores can be found at https://www.lttc.ntu.edu.tw/geptscoreremark/hicomposition.pdf. It should be noted that the writing score the test-takers finally received was a composite of the scores from the Chinese-English translation part (40%) and the Guided Writing part (60%). However, due to our focus on Taiwanese English learners' ability to compose an essay, this study only considered the scores the respondents received in the Guided Writing part.

The average writing subtest score of the 470 respondents was 3.35 (SD = 0.51), with most of the scores ($n = 453$) falling within the range between 3 and 4. The writing score distribution of our sample was similar to that of the overall population who took this test. Among our sample, 107 test-takers (22.77%) received a score at or above 4, and thus passed the Guided Writing subtest, while the remaining 363 test-takers (77.23%) obtained a score below 4 and were considered unsuccessful on this subtest.

5.3.2.2 The Learning Strategies for EFL Writing (LS-Writing) Questionnaire

The LS-Writing Questionnaire was devised to investigate the strategies EFL learners in Taiwan usually apply to develop their ability to write in English (see Appendix I). The questionnaire was constructed with reference to the suggested learning skills relevant to writing provided in Oxford (1990). The questionnaire consists of 40 items, with responses to be made on a 5-point Likert scale (1 = never or almost never true of me; 5 = always or almost always true of me). To ensure the respondents' understanding of the questions, the questionnaire was translated into Chinese. The Chinese version was piloted on a group of college students to collect suggestions regarding the clarity of question items and revisions were made accordingly. Further revisions were also made based on comments from two language experts.

After collecting the data of the respondents, the underlying constructs of the LS-Writing questionnaire were investigated by using exploratory factor analysis

(EFA). The data yielded a Kaiser-Meyer-Olkin (KMO) value of .912 and a Bartlett's statistical significance with a probability below .001, suggesting that it was suitable for factor analysis. Then, we used the factor extraction method with the principal axis factor and the promax rotation and removed items which did not meet qualifications (a factor loading lower than .4 or items with cross-loading) and finally derived 5 factors and 24 items. The five factors were termed *cognitive strategy use*, *affective strategy use*, *seeking practice opportunities*, *planning and evaluation*, and *self-regulation*, respectively. For more details about these constructs, please refer to Appendix II.

5.3.3 Data Analysis

5.3.3.1 Quantitative Analyses

In addition to applying basic statistical analyses to gain a general overview of strategy use, we also adopted the structural equation modeling (SEM) approach to explore the interrelationships of different learning strategy types and their links to writing performance.

5.3.3.2 Error Analysis

To characterize EFL learners' writing difficulties, a sample of 30 scripts were randomly selected from the successful candidate group and 30 from the unsuccessful candidate group. As the comparison between the scripts receiving a score of 3 and those receiving a score of 4 is of particular interest to us since most of the test-takers' writing score fell within the range of 3–4, we selected only scripts with either of these two grades.

Our error coding framework was adapted from the frameworks in James (1998) and Tseng (2016). It covers 4 broad categories: mechanic errors, lexical use errors, grammatical errors, and textual level errors. In each category, there are specific error types, making a total of 34 error types (see Appendix III). To evaluate the reliability of the coding, 10% of the sampled scripts (6 from the successful candidate group and 6 from the unsuccessful candidate group) were re-coded by a second coder who had expertise in Applied Linguistics and experience teaching EFL writing. The inter-coder reliability reached 90.70% of agreement (Range = 82.5–100%). In the following analysis, the successful candidate group was compared with the unsuccessful candidate group in terms of the error types and error frequencies in order to ascertain which aspects could be targeted to improve overall writing effectiveness.

5.4 Results

5.4.1 The Use of Learning Strategies for EFL Writing

5.4.1.1 Relationships Among the Strategy Types

We examined the interrelationships between the five strategy types emerging from the EFA and their taxonomic structure by using the SEM approach. The maximum likelihood estimation (MLE) was applied to calibrate the parameters, and the hypothesized models were tested by using a series of chi-square difference tests. The resulting model is displayed in Fig. 5.1. This model revealed a significant chi-square value [$\chi^2(222) = 557.49$, $p < .001$], which could be due to the large sample size. Despite this, the model demonstrated a satisfying goodness-of-fit: χ^2/DF = 2.51, GFI = .91, CFI = .91, RMSEA = .057, SRMR = .057, indicating that it provided an acceptable representation of the sample data.

In this model, there was a higher-order factor of metacognitive strategy use governing the constructs of seeking practice opportunities, self-regulation and planning and evaluating, with factor loadings ranging between .62 and .86 (*p*s < .05). On the other hand, the metacognitive strategy use held an executive control over the use of cognitive strategies and affective strategies, with factor loadings of .62 and .64 (*p*s < .05), respectively. Cognitive strategy use and affective strategy use were found to be significantly correlated at a factor loading of −.25, suggesting a negative relationship between the use of these two strategy types.

5.4.1.2 EFL Learners' Strategy Use

Based on the taxonomy of learning strategies presented in the analyses above, we examined the EFL learners' strategy use. It was found that EFL learners frequently

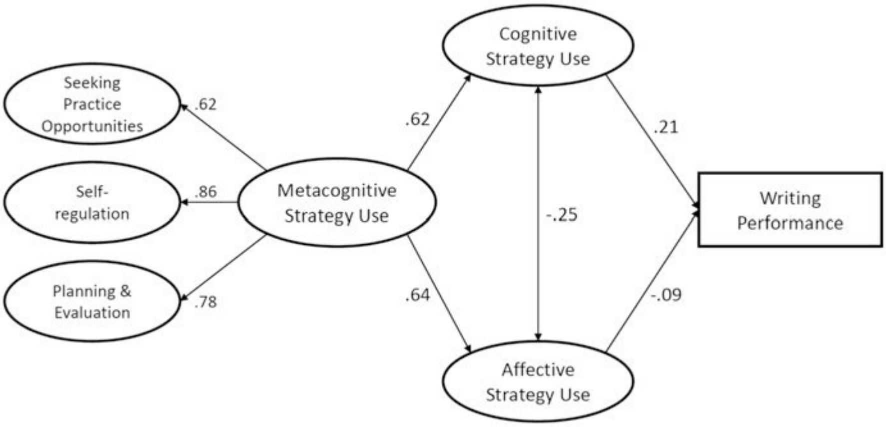


Fig. 5.1 Relationships between strategy use and writing performance

employed cognitive strategies [$M = 3.81$, SD $= .68$, $t(469) = 25.63$, $p < .001$], but significantly less frequently applied metacognitive strategies [$M = 2.84$, SD $= 0.69$, $t(469) = -5.13$, $p <.001$] and affective strategies [$M = 2.65$, SD $= .89$, $t(469) = -8.89$, $p <.001$]. Despite the finding on metacognitive strategy use, an examination of its three sub-factors revealed that EFL learners often applied self-regulation strategies [$M = 3.10$, SD $= .84$, $t(469) = 2.47$, $p = .014$], though they less often employed strategy types relating to seeking practice opportunities [$M = 2.50$, SD $= .93$, $t(469) = -11.70$, $p < .001$] and planning and evaluation [$M = 2.92$, SD $= .83$, $t(469) = -2.19$, $p = .029$].

5.4.2 Relationships Between Strategy Use and EFL Writing Performance

Correlation analyses demonstrated that EFL writing performance, as represented by the GEPT writing scores, demonstrated significant positive correlation with cognitive strategy use ($r = .14$, $p < .01$) and the construct of self-regulation of the metacognitive strategy use ($r = .11$, $p < .05$). Though the effect size was small, which could be due to the small range of writing score variation, it revealed the tendency that writing performance correlates with use of self-regulation behaviors and efforts devoted to writing practice.

The SEM approach was further applied to explore the interrelationships between strategy use and writing performance. The results showed that cognitive strategy use had a significantly positive and direct effect on writing performance, with a factor loading of .21. Though metacognitive strategy use showed no direct effect on writing performance, it did exert an indirect effect through cognitive strategy use. On the other hand, affective strategy use was found to be negatively associated with writing performance, yet the link did not reach significance. The full latent variable model is presented in Fig. 5.1.

5.4.3 Comparisons Between the Successful Candidate Group and the Unsuccessful Candidate Group

5.4.3.1 Patterns of Strategy Use

The successful candidate group and the unsuccessful candidate group were compared in terms of the frequency the candidates applied individual strategies. It was found that the unsuccessful candidates more often applied the affective strategy of regulating their feelings of tension during writing (LS16: $Z = -2.80$, $p < .01$), but significantly less often applied the metacognitive strategy of self-evaluation (LS37: $Z = 2.14$, $p < .05$) and the cognitive strategies of practicing using phrase patterns (LS18: $Z = 2.33$, $p < .05$) and recombining sentences (LS19: $Z = 2.18$, $p < .05$).

Table 5.1 The path coefficients in the successful candidate group model and the unsuccessful candidate group model

	Successful Candidates		Unsuccessful Candidates		
	Estimate	ß	Estimate	ß	z-score
Metacognitive Strategy Use→ Cognitive Strategy Use	1.60	.78**	0.92	.58***	−1.23
Metacognitive Strategy Use→ Affective Strategy Use	1.17	.59*	0.94	.66***	−0.45
Cognitive Strategy Use → Writing performance	−0.02	−.09	0.10	.19**	3.05**
Affective Strategy Use → Writing performance	0.02	.12	0.004	.01	−0.45

*$p < .05$; **$p < .01$; ***$p < .001$

A multi-group SEM was further conducted to examine and compare the models of strategy use and writing performance of the successful candidate group and that of the unsuccessful candidate group. The results showed that taxonomic structure of strategy use applied to both the successful candidate group and the unsuccessful candidate group; however, the two groups differed in the path coefficients between cognitive strategy use and writing performance ($Z = 3.05$, $p < .01$) (see Table 5.1). Specifically, the unsuccessful candidates demonstrated a stronger positive link between cognitive strategy and writing performance (ß = .19, $p < .005$), but the link failed to reach significance in the successful candidate group (ß = −.09, $p >$ .05). Moreover, it was observed that metacognitive strategy use showed a stronger link with cognitive strategy use in the successful candidate group (ß$_{\text{Meta\_Cog}}$ = .78, $p < .005$), but a stronger link with affective strategy use in the unsuccessful candidate group (ß$_{\text{Meta\_Aff}}$ = .66, $p < .001$), suggesting different patterns of learning strategy use among the two groups.

5.4.3.2 Distinctive Error Patterns of the Unsuccessful Candidate Group

Following the findings that the two groups differed significantly in their learning strategy use, we further investigated Taiwanese EFL learners' writing difficulties in order to link to the learning strategies on which EFL writing instruction should focus. A sample of writing scripts was examined in terms of the writing errors, and comparisons were made between the successful candidate group and the unsuccessful candidate group in terms of error types and frequencies (see Table 5.2). Overall, the unsuccessful candidates committed significantly more types of errors and made significantly more errors. In terms of the group difference within each category, the unsuccessful candidate group made significantly more errors at all levels except for the use of mechanics; however, the greatest group difference was observed at the grammatical level, which might be linked to their less frequent practice using language at the phrasal or sentential level.

Table 5.2 Error patterns of the successful candidates and the unsuccessful candidates

	Successful candidates	Unsuccessful candidates	t-test result
No. of scripts	30	30	
Overall performance			
Error type	13.30 (3.91)	16.97 (3.43)	$t(58) = 3.86$***
Error frequency	24.50 (10.68)	35.10 (10.97)	$t(58) = 3.79$***
Error categories			
Mechanics	3.33 (2.22)	3.57 (2.36)	$t(58) = 0.40$
Typ[a]	2.10 (1.54)	2.37 (1.52)	$t(58) = 0.68$
Fmt	1.03 (1.16)	0.93 (1.02)	$t(58) = -0.36$
Redn	0.20 (0.41)	0.27 (0.58)	$t(58) = 0.51$
Lexical use	6.03 (3.50)	8.37 (3.52)	$t(58) = 2.58$*
LexForm	0.73 (1.02)	1.40 (1.35)	$t(58) = 2.16$*
WW	1.60 (1.63)	1.37 (1.27)	$t(58) = -0.62$
WP	0.57 (0.73)	0.50 (0.63)	$t(58) = -0.38$
Ambg	0.97 (1.40)	1.87 (1.53)	$t(58) = 2.38$*
GenW	0.27 (0.58)	0.43 (0.68)	$t(58) = 1.02$
Col	1.53 (1.28)	1.90 (1.73)	$t(58) = 0.93$
SmAn	0.37 (0.72)	0.90 (1.21)	$t(58) = 2.07$*
Grammar	12.37 (5.89)	18.47 (7.00)	$t(58) = 3.65$***
Det	2.57 (1.91)	3.37 (2.16)	$t(58) = 1.52$
Num	2.00 (1.88)	2.90 (2.16)	$t(58) = 1.73$
S-V Agm	0.53 (0.78)	1.13 (1.36)	$t(58) = 2.10$*
VbFm	2.33 (2.32)	3.33 (2.09)	$t(58) = 1.75$
Comp	0.13 (0.35)	0.33 (0.61)	$t(58) = 1.57$
S-V Mis	0.10 (0.31)	0.33 (0.71)	$t(58) = 1.65$
Arg Mis	0.40 (0.62)	0.33 (0.61)	$t(58) = -0.42$
MulV	0.10 (0.31)	0.57 (0.73)	$t(58) = 3.24$**
RProE	0.17 (0.46)	0.17 (0.38)	$t(58) = 0.00$
PE	1.50 (1.55)	1.67 (1.30)	$t(58) = 0.45$
Frag	0.30 (0.65)	1.10 (1.27)	$t(58) = 3.07$**
Clau	0 (0)	0.07 (0.25)	$t(58) = 1.44$
Run-on	0.53 (1.17)	1.63 (2.21)	$t(58) = 2.42$*
ModE	0.40 (0.86)	0.33 (0.55)	$t(58) = -0.36$
WConj	0.80 (0.85)	0.80 (1.03)	$t(58) = 0.00$
WOE	0.50 (0.68)	0.40 (0.62)	$t(58) = -0.59$
Textual level	2.77 (2.18)	4.70 (3.23)	$t(58) = 2.72$**
ProE	0.80 (0.96)	1.53 (1.59)	$t(58) = 2.16$*
OGL	0.03 (0.18)	0.03 (0.18)	$t(58) = 0.00$
IRL	0.20 (0.41)	0.60 (0.72)	$t(58) = 2.64$*
NoUnity	0.20 (0.48)	0.43 (0.57)	$t(58) = 1.71$
NoCohen	0.47 (0.68)	0.77 (0.86)	$t(58) = 1.50$
NoConet	0.60 (0.86)	0.63 (0.77)	$t(58) = 0.16$
WConet	0.43 (0.86)	0.37 (0.72)	$t(58) = -0.33$
LogFal	0.03 (0.18)	0.33 (0.61)	$t(58) = 2.59$*

*$p < .05$; **$p < .01$; ***$p < .001$

[a] For the details of the error categories, please refer to Appendix III

In terms of grammatical errors, the unsuccessful candidate group showed a greater proportion of errors not only with the morpho-syntactic marking of subject-verb agreement but also with syntactic structures such as multiple verb clauses, sentence fragments, and run-on sentences. The group difference was qualitative, as well as quantitative. For example, a closer inspection of the subject-verb agreement errors revealed that the successful candidate group committed this type of error mostly in complex syntactic structures like extended nominal groups (e.g., "This kind of employees are more productive . . .") or relative clauses (e.g., "'Learning by doing' is what a man who want to be successful should believe in"), while the unsuccessful candidate group showed this type of error even in basic sentences (e.g., "My father have a wierd sick in his childhood"). Syntactic level errors revealed the unsuccessful candidates' lesser mastery of basic syntactic rules of English, such as an independent clause requiring a subject and a verb, as well as insufficient knowledge of the syntactic behavior of some lexical units, such as subordinators and coordinators. This led them to produce multiple verb clauses such as "So the higher education guarantees greater success in life is not totally true" or sentence fragments such as "If you want to pursue further study in order to achieve their career and personal goals." Run-on sentences involving comma splices or fused sentences, were also commonly observed among the unsuccessful candidates.

In addition to difficulties with grammar, the unsuccessful candidates' error patterns at other levels provide further insight into their writing difficulties. For example, the unsuccessful candidates made significantly more lexical form errors and incomprehensible strings of words, which either generated ambiguous meanings or were otherwise difficult to interpret. These misuses suggest that the learners might lack vocabulary or adequate understanding of the words in terms of their senses. The unsuccessful candidates also revealed more prominent difficulties at the textual level. In particular, they tended to use pronouns inconsistently, include irrelevant sentences, and commit logical fallacies.

To summarize, the findings showed significant group differences at all linguistic and textual levels, yet with the greatest group difference manifesting itself in grammatical performance. Moreover, the errors of the unsuccessful candidate group were qualitatively different from the successful candidate group, demonstrating their poorer mastery of basic syntactic rules. These findings will be discussed in conjunction with findings on the unsuccessful candidates' learning strategy use in the Discussion session.

5.5 Discussion

In this study, we first explored the taxonomy of learning strategies in relation to EFL writing and further examined the link between strategy use and writing performance, as based on the scores EFL learners received in a standardized writing test. Furthermore, we compared the successful and unsuccessful candidates of the writing test in terms of their learning strategy use and their writing errors in order to discover which learning strategies writing instruction should focus on to help learners reach the required level of writing ability.

The investigation on the taxonomy of and the interrelationship between the learning strategies was conducted with the aim of understanding the learning processes EFL learners were involved in. The results revealed that EFL learners employed metacognitive strategy use, cognitive strategy use, and affective strategy use. The cognitive strategy use was related to practicing the use of the English language and was the most often applied strategy among the EFL learners. On the other hand, affective strategy use, which involved the regulation of emotion, was found to be less often applied. Furthermore, we found a role for metacognitive strategy use, which was missing in the model proposed by Kao and Reynolds (2017). Interestingly, metacognitive strategy use appeared to be a multifaceted construct, which included not only a focus on learners' planning and evaluating their own writing, but also the acts of regulating their own learning processes with the goal of improving their writing and practicing writing in authentic contexts outside of the classroom. The metacognitive strategy use played an executive role, which determined cognitive strategy use and affective strategy use.

The findings further demonstrated that the coordination of strategy use affects EFL learners' performance in writing. While the study found that cognitive strategy use, i.e., actively using English, had a direct and positive effect on EFL writing, metacognitive strategy use actually exerted an indirect effect on writing performance via cognitive strategy use. The findings suggest that practicing using English will lead to improvement in EFL writing. However, consistent with the proposal of Brown and Palincsar (1982), it was found that learning can be more effective if the EFL learners can maximize their use of metacognitive strategies, such as seeking opportunities to write in English in real situations (e.g., chatting with friends on social media), strengthening their focus on the planning and evaluation of the text, and purposefully striving towards the goal of improving writing by consulting with more proficient writers or monitoring their own progress. Use of these strategies creates opportunities for language use and may enhance the quality of writing products.

The relationship between strategy use and EFL writing performance was further verified in the comparison of strategy use patterns between the successful candidates and the unsuccessful candidates in the writing test. The multi-group SEM results revealed that the unsuccessful candidate group showed positive association between cognitive strategy use and writing performance, though the successful candidate group did not. This suggests that even among the unsuccessful candidates, those who often applied cognitive strategies tended to have better writing performance. Although the association was not observed in the successful candidate group, this might result from the smaller group size and a narrower score variation in this group.

In addition, the two groups demonstrated different patterns of coordination among the strategy types. Specifically, the successful candidate group demonstrated a stronger link between metacognitive strategy use and cognitive strategy use, while the unsuccessful candidate group showed a stronger link between metacognitive strategy use and affective strategy use. This finding provides insight into the role of metacognitive strategy use, which may be guided by the EFL learners' knowledge about writing and about how they can improve their writing (Victori, 1999). Our findings that the unsuccessful candidates less often applied metacognitive strategies could be attributed to their insufficient knowledge of the writing task, inclusive of text structure and processes required. This undermines their ability to regulate themselves and to plan and evaluate their progress. This may explain why they more often concentrate on regulating their anxiety toward the task.

When the unsuccessful candidates were compared with the successful counterparts in terms of their application of individual strategies, it was found that the unsuccessful candidates less often practiced using phrase patterns or stringing together known expressions into longer sentences in writing. Their lack of practice using English at phrasal and sentential levels is consistent with the findings of the error analysis, which demonstrated candidates' difficulties at the grammatical level, particularly their less than satisfactory mastery of basic syntactic rules or the syntactic behavior of some lexical items. In addition to that, they also showed significantly more errors at lexical and textual levels.

A synthesis of the findings has implications for writing instruction. Though developing the ability to write in English has been included in Taiwan's education curriculum, there is considerable variation regarding how writing instruction has been incorporated and implemented in the classroom. However, the development of writing ability requires specific guidance and practices. Therefore, it is necessary for teachers to develop a writing curriculum which incorporates instruction not only on basic mechanics and writing conventions, but also of how a text should be structured. In addition, discussion on elements of good writing should be included, thereby providing learners with the ability to perform self-evaluation. In addition to this kind of textual instruction, the curriculum should also present the writing process to EFL learners, including—but not limited to—brainstorming, planning, organizing, evaluating, and resourcing. When guided through these processes, they should be able to approach the writing task more effectively.

It is also suggested that writing should be constructed as a communicative activity, not only in the process of writing, but also in the process of learning to write. Writing activities in a language classroom are often constructed in the form of one-way communication. That is, EFL learners are assigned a topic to write about, and the teacher is often the only audience. There is little substantial dialogue between the writer and the reader, and the writing activity is just another assignment to be dealt with. However, writing may be greatly improved if it is conducted for communicative purposes, such as persuading or sharing an idea with others. When EFL learners learn that what they write and how they write matters to the kind of message they are to deliver, they pay closer attention to their writing products. To make writing a communicative activity, the teachers need to create situations in which students write for a purpose. In addition, the teachers can also involve peers as one of the audiences, allowing students to provide each other with feedback. These arrangements can help EFL learners view themselves not only as language learners but also as authors and to develop a sense of ownership of their writing, which further increases their motivation to improve (Lei, 2016).

The findings of the error analysis also suggest the need to incorporate grammar instruction in EFL writing courses. However, there is no need to cover all writing errors in class, but rather to focus on those aspects which distinguish the more proficient from the less proficient writers (see also Tseng, 2016). In the case of this study, priority attention should be given to the use of subject-verb agreement and mastery of basic syntactic rules. We also suggest a functional-oriented approach of grammar instruction, so that EFL learners can learn grammar and usage in a meaningful context.

Last but not least, we would argue that strategy instruction should also be included in the writing curriculum. In addition to cognitive strategy use, which has been found to have direct effect on EFL writing performance (Huang & Chen, 2006; Kao & Reynolds, 2017), efforts need to be taken to increase EFL learners' metacognitive strategy use, focusing on self-evaluation, goal setting, and seeking practice opportunities, all of which can maximize practice opportunities and enhance motivation for learning, and thus lead to greater chances of improvement.

5.6 Conclusion and Implication

The findings of this study have both theoretical and pedagogical implications. On the theoretical side, our study presents a taxonomy of learning strategies in relation to EFL writing and further establishes the link between learning strategy use and writing performance, which can contribute to EFL writing pedagogies. In addition, the findings that the unsuccessful candidates in the writing test showed distinctive error patterns and greater writing difficulties compared with the successful candidates provides evidence for the scoring validity of the GEPT Writing Test at the High-Intermediate level. Despite its contributions, this study still has limitations. For example, our subject pool demonstrated only a small range of score variation, thus making statistical significance difficult to detect for some phenomena. Moreover, this study examined EFL learners who took an English writing test at a single proficiency level (i.e., the GEPT High-Intermediate Level); caution should be employed if the findings are extended to other populations, specifically those at a lower proficiency level. In addition, GEPT test-takers can register for any level that fits their needs; therefore, we cannot know the exact proficiency level of all of the test-takers. Finally, the link between learning strategy use and writing performance can be best verified empirically; therefore, further empirical studies on the link between strategy use and writing performance should be encouraged.

Acknowledgements I would like to thank the colleagues in LTTC who helped with data collection and the participants who took part in this study. Special thanks to Dr. Jessica Wu for her suggestions on an early draft of the work. Finally, I thank the editors and the anonymous reviewers for their constructive comments.

Appendices

Appendix I

The Learning Strategies for EFL Writing (LS-Writing) Questionnaire

Below are statements about learning to write in English. Please read each statement and mark the response that tells how true the statement is in terms of what you actually do when you learn or practice to write in English.

1 = never or almost never true of me
2 = usually not true of me
3 = sometimes true of me
4 = usually true of me
5 = always or almost always true of me

Question Items
1. I imitate native speakers' writing (e.g., use of sentence patterns) to learn English writing.
2. In class, I use note-taking technique to practice writing.
3. In class, I highlight different types of information relevant to English writing (e.g., vocabulary, grammar points, cultural concepts, etc.) in the textbook or handouts.
4. I write a summary for a longer passage to practice writing.
5. While learning writing, I analyze elements in an article (e.g., indicating topic sentence).
6. I practice writing by writing in a real situation (e.g., writing a letter in English to friends).
7. I organize my learning by practicing writing on a regular basis, or keeping a language learning notebook to write down new target language expressions or structures.
8. I seek opportunities to practice writing (e.g., chatting with friends on LINE or other social media by typing English).
9. I write language learning diaries to understand and to keep track of my thoughts, attitudes, and language learning strategies.
10. I discuss my feelings and needs about writing with someone else.
11. I share my writing with my classmates and ask for feedback and comments.
12. To improve my writing, I ask my teacher to mark my writing errors and I correct them on my own.
13. Before writing, I examine the requirements of the task and my language ability to determine the need for further aids (e.g., seeking the teachers' help with or checking the grammar book about the use of conditional clauses when this unfamiliar sentence pattern is required in the writing task).
14. Before writing, I search for some information to gain the background knowledge or cultural background relevant to the writing topic in advance.
15. I pay attention to my feelings (e.g., tension) before writing.

16. When I feed nervous before or during writing, I try to reduce my tension by use of music, deep breathing, or laughter.
17. While writing, I use words I recently learn in my writing.
18. While writing, I use phrase pattern in my writing.
19. I string together two or more known expressions into writing.
20. I apply previous knowledge or experience to produce my writing (e.g., use knowledge of Chinese *bei*-construction to produce passive clauses in English).
21. I use resources (e.g., dictionary, grammar book or something related to the topic I will write) to write.
22. I use translation skills to help me produce my writing.
23. While writing, I use a synonym to convey the intended meaning.
24. When I encounter a word I do not know how to express in my writing, I make up my own word to gain meaning.
25. When I cannot write difficult sentences, I use simpler, less precise, or slightly different ones.
26. While writing, I select the topic I can handle well to write.
27. I brainstorm to generate ideas for writing to bring out my own existing ideas and start expanding them as preparation for the future writing task.
28. While writing, I decide in advance which aspects of the writing (e.g., grammar, vocabulary, sentence structure) to focus on at given time.
29. While writing, I plan writing by taking into consideration the purpose of the task (including the type of written format and the needs of the potential audience).
30. In order to make my writing better, I use checklists to monitor my writing.
31. I am aware of my readers' thoughts and feeling while writing.
32. During my writing, I am not afraid of using different sentence patterns or vocabulary regardless of the possibility of making mistakes.
33. Before, during and after writing, I make positive statement to encourage myself to be confident.
34. While writing, I set a deadline and expect to reach some writing achievement in the period of time (e.g., finish writing an essay within 3 days).
35. While writing, I consult with proficient writers to enhance my writing.
36. After writing, I reread my writing to find out whether there is an inappropriate construction or vocabulary and revise it.
37. I review samples of my writing, note the style and content and assess progress over time.
38. After writing, I review my writing at regular intervals and make necessary revisions.
39. After writing, I actively find ways to help me learn writing, e.g., reading books about writing skills, talking about my writing problems with others, and share ideas with each other about effective strategies
40. I reward myself after completing a writing task.

Appendix II

Item, factor loading, explained variance, and Cronbach's α

		Descriptive stats		Factors				
	Items	M	SD	Cog[1]	Aff[2]	SPO[3]	P&E[4]	SR[5]
1	I string together two or more known expressions into writing. (LS19)	3.91	0.91	.76				
2	While writing, I use a synonym to convey the intended meaning. (LS23)	3.85	0.88	70				
3	While writing, I use phrase pattern in my writing. (LS18)	3.76	0.93	.66				
4	While writing, I use words I recently learn in my writing. (LS17)	3.69	0.95	.64				
5	I apply previous knowledge or experience to produce my writing (e.g., use knowledge of Chinese *bei*-construction to produce passive clauses in English). (LS20)	3.57	1.12	.59				
6	When I cannot write difficult sentences, I use simpler, less precise, or slightly different ones. (LS25)	4.08	0.79	.59				
1	When I feed nervous before or during writing, I try to reduce my tension by use of music, deep breathing, or laughter. (LS16)	2.84	1.23		.67			
2	Before, during and after writing, I make positive statement to encourage myself to be confident. (LS33)	2.58	1.22		.65			
3	I pay attention to my feelings (e.g., tension) before writing. (LS15)	2.79	1.12		.59			
4	I reward myself after completing a writing task. (LS40)	2.38	1.19		.48			
1	I seek opportunities to practice writing (e.g., chatting with friends on LINE or other social media by typing English). (LS8)	2.79	1.20			.67		
2	I write language learning diaries to understand and to keep track of my thoughts, attitudes, and language learning strategies. (LS9)	2.11	1.05			.65		
3	I practice writing by writing in a real situation (e.g., writing a letter in English to friends). (LS6)	2.59	1.21			.62		
1	While writing, I decide in advance which aspects of the writing (e.g.,	3.03	1.14				.89	

(continued)

		Descriptive stats		Factors				
	Items	M	SD	Cog[1]	Aff[2]	SPO[3]	P&E[4]	SR[5]
	grammar, vocabulary, sentence structure) to focus on at given time. (LS28)							
2	I brainstorm to generate ideas for writing to bring out my own existing ideas and start expanding them as preparation for the future writing task. (LS27)	2.91	1.14				.51	
3	While writing, I plan writing by taking into consideration the purpose of the task (including the type of written format and the needs of the potential audience). (LS29)	3.50	1.06				.50	
4	In order to make my writing better, I use checklists to monitor my writing. (LS30)	2.23	1.14				.47	
1	While writing, I consult with proficient writers to enhance my writing. (LS35)	3.45	1.17					.84
2	I discuss my feelings and needs about writing with someone else. (LS10)	2.85	1.15					.73
3	To improve my writing, I ask my teacher to mark my writing errors and I correct them on my own. (LS12)	3.51	1.18					.70
4	I share my writing with my classmates and ask for feedback and comments. (LS11)	2.46	1.07					.67
5	After writing, I actively find ways to help me learn writing, e.g., reading books about writing skills, talking about my writing problems with others, and share ideas with each other about effective strategies. (LS39)	3.31	1.12					.58
6	While writing, I set a deadline and expect to reach some writing achievement in the period of time (e.g., finish writing an essay within 3 days). (LS34)	2.70	1.20					.56
7	I review samples of my writing, note the style and content and assess progress over time. (LS37)	3.39	1.10					.54
	Explained variance			8.33	4.31	3.13	3.56	26.99
	Cumulative explained variance			8.33	12.64	15.77	19.33	46.32
	Cronbach's α			.85	.69	.72	.70	.86

(continued)

		Descriptive stats		Factors				
	Items	M	SD	Cog[1]	Aff[2]	SPO[3]	P&E[4]	SR[5]
	Construct reliability (CR)			.83	.69	.72	.74	.83
	Average variance extracted (AVE)			.45	.36	.47	.42	.46

Notes: Cog[1] = Cognitive strategy use; Aff[2] = Affective strategy use; SPO[3] = Seeking practice opportunities; P&E[4] = Planning and evaluation; SR[5] = Self-regulation

Appendix III

Error Coding System

Code	Definition
Mechanics	
Typ	Inaccurate word forms, including (1) misspellings, (2) a lack of or additional space in a compound, or (3) inaccurate tense or aspect markings.
Fmt	Format errors, including (1) no indentation, (2) wrong punctuation (e.g., non-restrictive relative clauses), (3) lower-case and upper-case errors, (4) improper paragraphing
Redn	Redundancy—redundant words or phrases
Lexical level	
LexForm	Correct spelling, yet they are erroneous due to similarity in forms or similarity in pronunciation, or incorrect part of speech
WW	Wrong word—incorrect use of words
WP	Wrong phrase—incorrect use of phrases (correct in form but semantically incongruous in context)
Ambg	Ambiguity—words/phrases that are ambiguous in meaning and cause incomprehensibility of the sentence
GenW	Generic words—words that are too general, not specific in meaning
Col	Collocation errors, including (1) violation of collocations, (2) incomplete phrasal verbs with no preposition or adverb
SmAn	Semantic anomaly—incorrect combination of words that leads to semantically anomalous or incomprehensible strings
Grammatical level	
Det	Determiner errors—incorrect use of determiners and articles
Num	Number errors—incorrect use of plural or singular forms of nouns
S-V Agm	Violation of subject-verb agreement
VbFm	Incorrect use of verb form, including wrong use of (1) tense, (2) aspect, (3) voice, (4) mood, (5) gerund or infinitive
ModE	Incorrect use of modals or auxiliaries
Comp	Errors in comparative forms
S-V Mis	Mismatch of subject and verb, including (1) incongruence between subject and verb, (2) lack of subject, (3) more than one subject

(continued)

Code	Definition
Arg Mis	Lack of argument
MulV	Multiple verbs—more than one verb in one sentence
RProE	Incorrect use of relative pronouns—(1) using wrong relative pronouns, (2) lack of relative pronouns
PE	Incorrect use of prepositions
Frag	Sentence fragments, including (1) incomplete clauses, (2) phrases with no verb
Clau	Dangling clauses that start with (1) a coordinating conjunction, or (2) a subordinating conjunction
Run-on	Including comma splices (i.e., overuse of commas to link independent clauses) and fused sentences (i.e., there is barely any comma between independent clauses)
WConj	Incorrect use of conjunctions, including using wrong conjunctions or lacking conjunctions
WOE	Word order errors—incorrect word orders
Textual level	
ProE	Incorrect use of pronouns, including (1) pronoun inconsistency, (2) improper use/omission of pronouns
OGL	Overgeneralization—sentences that involve overgeneralization
IRL	Sentences that are irrelevant to the development of the main idea
NoUnity	A supporting idea that is irrelevant to the topic
NoCohen	Inappropriate development of ideas in relation to the topic
NoConet	No connector—lack of connectors to combine sentences
WConet	Wrong connector— incorrect use of connectors between sentences
LogFal	Logical fallacy—(1) violation of logic, (2) incorrect cause-effect relation, (3) fail to represent ideas logically

References

Anderson, N. J. (2005). L2 strategy research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 757–772). Lawrence Erlbaum Associates.

Bai, R., Hu, G., & Gu, P. Y. (2014). The relationship between use of writing strategies and English proficiency in Singapore primary schools. *The Asia-Pacific Education Researcher, 23*, 355–365.

Brown, A. L., & Palincsar, A. S. (1982). Inducing strategies learning from texts by means of informed, self-control training. *Topics in Learning and Learning Disabilities, 2*, 1–17.

Chan, S. H. C., Wu, R. Y. F., & Weir, C. J. (2014). *Examining the context and cognitive validity of the GEPT Advanced Writing Task 1: A comparison with real-life academic writing tasks*. LTTC-CRELLA Collaboration Project.

Cheung, H., Chung, S. F., & Skoufaki, S. (2010). Indexing second language vocabulary in the Intermediate GEPT. In *Proceedings of the twelfth academic forum on English Language Testing in Asia: Continuity, innovation and synergy* (pp. 118–136). Language Training and Testing Center.

College Entrance Examination Center [CEEC]. (n.d.). *Advanced Subjects Test (AST)—English*. Retrieved December 9, 2019, from https://www.ceec.edu.tw/en/xmdoc/cont?xsmsid=0J180520414679660023

De Silva, R. (2015). Writing strategy instruction: Its impact on writing in a second language for academic purposes. *Language Teaching Research, 19*, 301–323.

Hsiao, T.-Y., & Oxford, R. L. (2002). Comparing theories of language learning strategies: A confirmatory factor analysis. *The Modern Language Journal, 86*, 368–383.

Huang, Y., & Chen, J. (2006). An empirical study on university students' extracurricular English writing strategies. *Foreign Language World, 112*, 35–40.

James, C. (1998). *Errors in language learning and use: Exploring error analysis*. Longman.

Kao, C.-W., & Reynolds, B. L. (2017). A study on the relationship between Taiwanese college students' EFL writing strategy use, writing ability and writing difficulty. *English Teaching & Learning, 41*, 31–67.

Knoch, U., & Frost, K. (*2016*). *Linking the GEPT writing sub-test to the Common European Framework of Reference (CEFR)*. LTTC-GEPT Research Reports RG-08.

Kunnan, A. J., & Carr, N. (2015). *A comparability study between the General English Proficiency Test-Advanced and the Internet-Based Test of English as a Foreign Language (iBT TOEFL)*. LTTC-GEPT Research Reports.

Kunnan, A. J., & Wu, J. R. W. (2009). The Language Training and Testing Center, Taiwan: Past, present and future. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 77–91). Routledge.

Kuo, G. (2005). *A preliminary corpus study on EFL test takers' writing proficiency*. Paper presented at the eighth international conference on English Language Testing in Asia.

Language Training & Testing Center [LTTC]. (2003). *Concurrent validity studies of the GEPT intermediate level, GEPT high-intermediate level, CBT TOEFL, CET-6, and the English test of the R.O.C. College Entrance Examination*. LTTC.

Language Training & Testing Center [LTTC]. (n.d.). *GEPT score reports*. https://www.lttc.ntu.edu.tw/results.htm

Lei, X. (2016). Understanding writing strategy use from a sociocultural perspective: The case of skilled and less skilled writers. *System, 60*, 105–116.

Lin, M. H. (2019). Correlations of English writing complexity, accuracy, fluency, and test scores: A case study of standardized writing test samples. *The Journal of Teaching English for Specific and Academic Purposes, 7*, 199–201.

Naiman, N., Frohlich, M., Stern, H. H., & Todesco, A. (1978). *The good language learner*. Ontario Institute for Studies in Education.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge University Press.

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Heinle & Heinle.

Qian, D. (2014). *A register analysis of advanced GEPT examinees' written production*. LTTC-GEPT Research Reports.

Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics, 11*, 117–131.

Sun, C-Y. (2018). *EFL learners' use of lexical chunks on the GEPT Writing Test*. Unpublished master's thesis. Tunghai University.

Tseng, C. C. (2016). Subsumable relationship among error types of EFL writers: A learner corpus-based study of expository writing at the intermediate level. *English Teaching and Learning, 40*, 114–151.

Victori, M. (1999). An analysis of writing knowledge in EFL composing: A case study of two effective and two less effective writers. *System, 27*, 537–555.

Weir, C. J., Chan, S. H. C., & Nakatsuhara, F. (2013). *Examining the criterion-related validity of the GEPT Advanced Reading and Writing Tests: Comparing GEPT with IELTS and real-life academic performance.* LTTC-GEPT Research Reports.

Wong, L. L. C., & Nunan, D. (2011). The learning styles and strategies of effective language learners. *System, 39*, 144–163.

Wu, J. R. W. (2008). An investigation of the relationships between strategy use and GEPT test performance. *English Teaching & Learning, 32*, 35–69.

Wu, J. R. W. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly, 9*, 11–25.

Wu, C-Y. (2016). *Thematic progression and cohesion in higher-rated and lower-rated essays in High-Intermediate GEPT Writing Test*. Unpublished master's thesis. National Taiwan Normal University.

Yang, S. (2002). A comparison of writing strategies employed by successful and unsuccessful EFL writers. *Foreign Language World, 3*, 57–64.

Yang, C. (2013). How Chinese beginning writers learn English writing: A survey of writing strategies. *Journal of Educational and Social Research, 3*, 9–18.

Yu, G., & Lin, S-W. (2014). *A comparability study on the cognitive processes of taking graph-based GEPT-Advanced and IELTS-Academic Writing Tasks.* LTTC-GEPT Research Reports.

Naihsin Li received her PhD degree in National Taiwan University and currently is a postdoctoral research fellow in the Language Training and Testing Center (LTTC), Taiwan. Her research interests lie in exploring the underlying processes of language learning and examining the factors affecting the development of first language and second language.

Chapter 6
How Chinese College EFL Learners Use Source Material Across Different Integrated Writing Tasks: A Comparative Study

Yan Zhou and Ke Bin

Abstract Although integrated writing tasks have received increasing attention in language testing research over the past two decades, the processes related to how students use source material in integrated tasks remain largely unknown. The present study probed into the volume and the patterns of source use, if any, across three integrated writing tasks, namely reading–writing, listening–writing, and reading–listening–writing. Source material use variations among students by proficiency levels were also examined. The results showed that although both the more and the less proficient students used source material in their writing, their source use frequency and source type differed in terms of comprehension, the borrowing of ideas, and gaining language and organizational support. From this, implications for the design of and further inquiries into integrated tasks were drawn.

Keywords Integrated writing task · Source material use · Proficiency level

6.1 Introduction

As representative tasks of great authenticity in the context of multimodality and communicative language ability, integrated writing tasks have aroused increasing interest and attention in language testing research, especially given the dissemination of the view of writing as a cognitive process (Bereiter & Scardamalia, 1987; Flower & Hayes, 1984). However, researchers have different opinions on the extensive use of integrated writing in various large-scale high-stake tests such as the TOEFL iBT, the IELTS, the TestDaF, the internet-based CET-4/6, the new HSK, etc. Supporters hold that integrated writing tasks have significant advantages over traditional independent writing tasks, including greater similarity to the real academic writing process (Cumming et al., 2005; Plakans, 2008; Serviss & Rodrigue, 2010); an ability to improve language accuracy, fluency, and complexity (Ruiz-Funes, 2015; Xu & Gao, 2007; Zhang & Zhou, 2014, 2016); and enhanced test fairness (Plakans, 2008;

Y. Zhou (✉) · K. Bin
Southern Medical University, Guangzhou, China

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_6

Weigle, 2004). However, the skeptics question the construct relevance of integrated tasks, especially those related to source use, such as source properties (He & Sun, 2015; Plakans & Gebril, 2013; Weigle & Parker, 2012). Although Plakans (2015) pointed out the necessity of defining what role source material plays in integrated writing tasks, few studies have investigated participants' source material use processes in these tasks.

The present study addressed the above gap by examining how Chinese college EFL learners at different English-language proficiency levels used source materials differently across the reading–writing, listening–writing, and reading–listening–writing tasks. Gebril and Plakans (2009) found that students' use of sources in integrated writing tasks was affected by their writing proficiency, and by the channels/ modalities through which source information was delivered. Therefore, the present study, a comparative one on different participants' source use across various integrated writing tasks, can enrich validity evidence and provide more comprehensive information for research on integrated writing tasks.

6.2 Literature Review

Studies on integrated writing tasks have generally been conducted from either the writing product approach or the writing process approach. The studies that adopt the writing product approach usually make comparisons between/across different independent and/or integrated writing tasks from the perspective of test score validity (Ohta et al., 2018; Plakans & Gebril, 2017; Wang et al., 2017) and features of written products (Cumming et al., 2005; Ruiz-Funes, 2015; Yang, 2009; Zhang & Zhou, 2014), while the studies from the writing process approach give priority to investigating test takers' writing strategy use and task performance in the three overlapping non-linear writing stages, namely the pre-writing, during-writing, and post-writing stages (Plakans, 2008; Yang, 2009; Yang & Plakans, 2012; Zhang & Zhou, 2016).

Source material input is one of the main concerns that integrated writing tasks encounter. Existing studies on source material use cover the following areas. Plakans (2008) was the very first to undertake comparative research on high- and low-proficiency participants' source use, finding no significant difference. Results also showed that source material in the reading–writing task functioned as content and organization model support for both proficiency groups, thus reducing their initial planning time (Plakans, 2008). Furthermore, Plakans and Gebril (2012) and Wu (2014) explored the relationship between participants' language proficiency and how they used source material, detecting no significant differences. However, they found significant differences between high- and low-proficiency participants in terms of their understanding of the source material. Later, Plakans and Gebril (2013) worked on participants' source use to predict an individual's score in a reading–listening–writing task. Their results indicated that the three areas of source text use (i.e., the importance of source text ideas that writers included in their summary, the use of ideas from a reading source text and a listening text, and the borrowing of exact wording from the source texts or verbatim source use) could explain over 50% of the variance in participants' scores in a reading–listening–writing task, indicating that source use plays a vital role in writing-from-source tasks. Moreover, the influence of different source material characteristics constitutes its own research focus. He and Sun (2015) and Liu and Stapleton (2018) explored how different source material input influences the textual features of participants' written products. Weigle and Parker (2012) analyzed the extent to which (i.e., how much) students borrowed source text language in a reading–writing task. Their dataset comprised 63 essays on two topics, with two groups of students at four proficiency levels. Their results suggested that only a small percentage of students borrowed extensively from the source texts and that there were only minor differences in borrowing patterns across topics, student groups, and proficiency levels (Weigle & Parker, 2012). Being different from the above three, Homayounzadeh et al. (2019) discussed participants' performance on two TOEFL iBT reading–listening–writing tasks that entailed source materials on different topics and with different structural organization and lexical and conceptual overlaps; they observed significant differences in test takers' overall scores, their comprehension of the source information, the type and quality of source borrowing strategies, and their tendency to either use their original ideas and structures or rely more on the sources.

To effectively define the construct of integrated writing tasks, it is imperative to collect empirical evidence about how writers compose from sources (Chan, 2017) and investigate writers' cognitive processing in writing tasks. Zhu et al. (2016) used correlation, joint factor analysis, and regression analysis to examine the construct of integrated writing tasks in Chinese-language examinations in Hong Kong and found small but significant correlations between students' performances in the independent listening and listening–reading–writing tasks. However, joint factor analysis revealed that there were no common factors between the two tasks. Chan (2017) employed keystroke logging and discovered that the nature of writers' reading-into-writing processes might have a major influence on final performance. Chan (2018) used keystroke logging to investigate both the contextual features of a range of real-life academic writing and reading-into-writing test tasks and the cognitive processes required to complete the integrated task type successfully. In addition, Cheong et al. (2018) used a multifaceted approach to study the relationship between three Chinese-language assessments. They discovered that reading cognitive skills contributed more to integrated writing task performance than listening cognitive skills did; furthermore, the interaction between the relationships of reading and listening to integrated writing performance was significant (Cheong et al., 2018). Rukthong and Brunfaut (2020) used stimulated recalls and questionnaires to collect task takers' processing and strategy use as well as their perceptions of processing difficult sources. Their study revealed that the source passage's linguistic difficulty partly resulted in variations in the use of listening processes and strategies in tasks with different listening inputs (Rukthong & Brunfaut, 2020). Michel et al. (2020) examined the extent to which task type influenced both group of writers' behaviors and their cognitive processes through a mixed-methods approach that employed keystroke logging, eye-tracking, and verbal protocols. They reported that compared with the independent task, the integrated task elicited more dynamic and varied behaviors and cognitive processes across writing stages (Michel et al., 2020). From the above review, it is evident that research investigating writers' cognitive processing in integrated writing tasks is still in its infancy; meanwhile, recent efforts have introduced many methodological innovations, including the use of stimulated recalls, eye-tracking, keystroke logging, etc., that will significantly expand the scope and depth of research in this domain.

Despite the great contributions the above studies achieved, a research gap remains in that no study has combined the following: (1) more and less proficient Chinese college EFL learners; (2) the reading–writing, listening–writing, and reading–listening–writing tasks; and (3) the extent to which (the amount) and how participants use source material (the pattern of use). Therefore, the present study has explored how language proficiency (in more and less proficient groups) and writing task (i.e., the reading–writing, listening–writing, and reading–listening–writing tasks) influence participants' source use in terms of amount and pattern, represented by the following two research questions:

1. What differences, if any, exist in terms of source use amount across the three integrated tasks by different proficiency groups of Chinese college EFL learners?
2. What differences, if any, exist in terms of source use patterns across the three integrated tasks by different proficiency groups of Chinese college EFL learners?

6.3 Methodology

6.3.1 Materials and Instruments

The instruments adopted in the present study were three writing tasks (the reading–writing, listening–writing, and reading–listening–writing tasks) and three questionnaires corresponding to each task. All three writing tasks used *drug testing* as the topic. The reading–writing task required the students to read a 300-word English passage (in about 5 min) before writing an argumentative essay of 200–300 words in 30 min. The listening–writing task required the students to listen to an English video clip of about 300 words (approximately 128 s) and then write an argumentative essay of 200–300 words in 30 min. The reading–listening–writing task asked the students to first read a 300-word English passage(in about 5 min), listen to an English video clip of about 300 words (approximately 128 s), and then write an argumentative essay of 200–300 words in 30 min. The reading material in the reading–listening–writing task was the same as that in the reading–writing task, and the listening material in the reading–listening–writing task was the same as that in the listening–writing task. Most of the words and phrases in the reading and listening materials were on the College English Test Band 4 (CET-4) vocabulary list; in cases where a word was not on the list, a Chinese equivalent was provided. The required reading speed was about 55 words per minute and the broadcasting speed of the listening material was around 125 words per minute. Both speeds were lower than those

Table 6.1 Source use framework in each questionnaire

Source use type	Questionnaire items
Reading for comprehension	1. I could understand most of the words in the reading passage and/or the listening record. 2. I could understand most of the ideas in the reading passage and/or the listening record. 3. I often reread the reading passage and/or replay the listening record while I was writing.
Gaining ideas	4. I used the reading passage and/or the listening record to help me get ideas on the topic. 5. I used only my own ideas in my writing, nothing from the reading passage and/or the listening record.
Shaping opinion	6. The reading passage and/or the listening record helped me choose an opinion on the issue.
Supporting opinion	7. I used some of the ideas from the reading passage and/or the listening record in my essay. 8. I used examples and ideas from the reading passage and/or the listening record to support my argument in my essay.
Language support	9. I used some words from the reading passage and/or the listening record when I wrote. 10. The reading passage and/or the listening record helped me write better.
Reading for organization	11. I used the reading passage and/or the listening record to help me organize my essay.

described in the CET-4 requirements. The reading passage, writing task requirements, and a brief rating criterion were attached to the test paper for each of the three writing tasks.

Three questionnaires that were adapted from Plakans and Gebril (2012) were used to investigate test takers' source use information in each of the above three writing tasks. Each questionnaire contained eleven 5-point Likert scale items and five semi-open questions. Table 6.1 reports the constructs of the questionnaires as well as the corresponding items and content for each type of source use. The five semi-open questions were designed to gather information about test takers' feelings and the difficulties they may have faced during writing.

6.3.2 *Participants*

Pilot tests were conducted before the field tests. For each writing task, ten junior students majoring in English and ten freshmen of medicine-related majors took part in the pilot test. To provide some operational suggestions for the field tests, their feedback on each writing task and their corresponding questionnaires were collected, including information about the difficulty of the reading and/or listening material, time allocation, directions, etc., and questionnaire readability.

In the field tests, each participant was required to complete the questionnaire after finishing the corresponding writing task. At the very beginning, 120 grade-1 freshmen of medicine-related majors who had not taken part in the CET-4 were recruited as members of the less proficient group and 120 grade-3 English majors who had passed the Test for English Majors Band 4 (TEM-4) were recruited as members of the more proficient group. The differentiation between the less and the more proficient groups was partly based on the length of time the participants had been learning English. The participants in the more proficient group had been learning English for 2 years longer than the less proficient group. The less/more proficient separation was also based on the differences in the overall English-language proficiency descriptors between CET-4 and TEM-4. CET-4 measures non-English majors' comprehensive ability of English use according to the teaching objectives set in the College English Curricular Requirements (CECR). CET-4 aligns to the "basic objectives" in CECR and allows participation of university students of grade two or above. Test for English Majors (TEM) measures English majors' comprehensive ability to use English according to The English Teaching Syllabus for English Majors (ETSEM). Participation in TEM-4 is limited to English sophomores who have completed the courses specified in the elementary stage aims of the ETSEM. For example, the CECR requires that students who pass CET-4 should be able to read at about 100 wpm and understand listening video clips broadcasted at about 120 wpm, while The ETSEM requires that students who pass TEM-4 should be able to read at about 120–180 wpm with an accuracy of no less than 70% and understand listening video clips broadcasted at about 120 wpm with an error rate of less than 8%. The contrast in the descriptors about skills and ability requirements shows that the level of language ability defined in the "basic objectives" of the CECR is lower than that defined in the elementary stage aims of the ETSEM. Therefore, it is considered reasonable to recruit those college freshmen in medicine-related majors who hadn't taken part in CET-4 as members of the less proficient group and those English junior students who had passed TEM-4 as members of the more proficient group.

In the less proficient group, which consisted of first-year students of medicine-related majors, there were 67 males (55.83%) and 53 females (44.17%) aged from 18 to 20 with an average length of 7.39 years of English-language learning. In this group, 101 students (84.17%) claimed to have been learning English for 7 years and 11 (9.17%) claimed to have been learning English for 10 or more years. In the more proficient group, which consisted of grade-3 English junior students who have passed TEM-4, there were 31 males (25.83) and 89 females (74.17%) aged from 18 to 22 with an average length of 9.71 years of English-language learning. In this group, 33 students (27.5%) claimed to have been learning English for 10 or more years and the remaining 87(72.5%) claimed to have been learning English for 9 years. This participant sampling ensured the homogeneity of the same language proficiency group.

6.3.3 Data Collection

Each proficiency group was asked to take part in the three writing tasks in three different classes simultaneously. That is to say, all the six tests (2 proficiency groups ×3 writing tasks) were conducted at the same time at different classrooms to avoid test sequence bias. It was later discovered that some students did not finish the writing task or missed some of the items in the questionnaire. Therefore, 30 writing samples and their equivalent questionnaires were used for either proficiency group in each writing task. As a result, 180 samples (30 × 2 × 3) were used in the final data analysis.

To ensure accuracy, the second writer entered the students' writing samples and questionnaire data into Microsoft Word and a Microsoft Excel sheet separately; and the first writer was responsible for checking and proofreading.

6.3.4 Data Analysis

Statistical analyses were conducted through SPSS 21.0. Before the formal data analysis, the reliability of the reading–writing, listening–writing, and reading–listening–writing questionnaires was tested. Their corresponding Cronbach's α values were 0.83, 0.81, and 0.79, respectively. Exploratory factor analyses were also carried out for the 5-point Likert scale items in each questionnaire to collect data on their validity. The value of KMO and the significance of Bartlett's test of sphericity in the reading–writing task were .755 (>.05) and .00 (<.05), respectively; in the listening–writing task, the values were .763 (>.05) and .00 (<.05), respectively; and in the reading–listening–writing task, they were .798 (>.05) and .00 (<.05), respectively. These values confirmed the distinction of six source use types, as presented in Table 6.1. The semi-open questions in each questionnaire were approved by two language testing experts (Zhang et al., 2015).

6.4 Results

This section presents the results of the analysis of the more and the less proficient participant groups' source use across the reading–listening–writing, reading–writing, and listening–writing tasks. The results will be reported in line with the two research questions concerning the participants' reported amount of source use and their source use pattern.

6.4.1 Participants' Reported Amount of Source Use

The second set of semi-open questions gathered data on whether and how participants used the source material. In the more proficient group, no significant difference could be detected across the three writing tasks. Only two participants reported not using the reading and/or listening material in either the reading–listening–writing (student ID: 3171204005) or the reading–writing task (student ID: 3171201031). However, detailed observation of their performances in the 5-scale Likert items and/or their written products contradicted their claims because both students' performances indicated that they used the reading and/or listening material to formulate opinions and gain organizational support for their writing. Meanwhile, two other participants mentioned that they did not use the listening material in the listening–writing task because they were unable to understand it. In the less proficient group, the number of participants who reported not using the source material was much greater in the listening–writing task than in the reading–listening–writing task, with the reading–writing task in between. One participant in the reading–listening–writing task and four in the reading–writing task reported not using the source material (s). Among these four reading–writing task participants, three pointed out that they could not understand the reading material, and one claimed that he did not know the words in the material. One-third of the participants (ten) reported not using the listening material in the listening–writing task because they could not understand it at all.

6.4.2 Participants' Source Use Pattern

6.4.2.1 Data Presentation

Results of the descriptive statistics and the inferential statistics are listed in Tables 6.2 and 6.3.

6.4.2.2 Using Source Material for Comprehension

From the perspective of overall language proficiency, the results of Item 1 (I could understand most of the words in the reading passage and/or the listening record), Item 2 (I could understand most of the ideas in the reading passage and/or the listening record), and Item 3 (I often reread the reading passage and/or replay the listening record while I was writing) in Table 6.2 show that the percentage of the more proficient group who reported using the source material was higher than that of the less proficient group across all three writing tasks, the only exception being the listening-writing task of Item 3. Table 6.2 also shows that from the perspective of task type, the percentages of both the more and the less proficient participants who

Table 6.2 Descriptive statistics between more and less proficient participants in items 1–11 across the three writing tasks

Task type		Reading-listening-writing		Reading-writing		Listening-writing	
Proficiency level		More	Less	More	Less	More	Less
Item 1	yes(%)	30(100)	22(73.3)	28(93.3)	22(73.3)	23(76.7)	14(46.7)
	no(%)	0	8(26.7)	2(6.7)	8(26.7)	7(23.3)	16(53.3)
Item 2	yes(%)	30(100)	23(76.7)	27(90)	23(76.7)	24(80)	10(33.3)
	no(%)	0	7(23.3)	3(10)	7(23.3)	6(20)	20(66.7)
Item 3	yes(%)	26(86.7)	18(60)	26(86.7)	21(70)	22(73.3)	22(73.3)
	no(%)	4(13.3)	12(40)	4(13.3)	9(30)	8(26.7)	8(26.7)
Item 4	yes(%)	29(96.7)	21(70)	26(86.7)	23(76.7)	23(76.7)	11(36.7)
	no(%)	1(3.3)	9(30)	4(13.3)	7(23.3)	7(23.3)	19(63.3)
Item 5	yes(%)	9(30)	10(33.3)	8(26.7)	11(36.7)	8(26.7)	10(33.3)
	no(%)	21(70)	20(66.7)	22(73.3)	19(63.3)	22(73.3)	20(66.7)
Item 6	yes(%)	29(96.7)	20(66.7)	24(80)	22(73.3)	25(83.3)	18(60)
	no(%)	1(3.3)	10(33.3)	6(20)	8(26.7)	5(16.7)	12(40)
Item 7	yes(%)	28(93.3)	22(73.3)	26(86.7)	21(70)	24(80)	16(53.3)
	no(%)	2(6.7)	8(26.7)	4(13.3)	9(30)	6(20)	14(46.7)
Item 8	yes(%)	26(86.7)	15(50)	18(60)	17(56.7)	18(60)	11(36.7)
	no(%)	4(13.3)	15(50)	12(40)	13(43.3)	12(40)	19(63.3)
Item 9	yes(%)	24(80)	23(76.7)	23(76.7)	20(66.7)	24(80)	16(53.3)
	no(%)	6(20)	7(23.3)	7(23.3)	10(33.3)	6(20)	14(46.7)
Item 10	yes(%)	28(93.3)	24(80)	25(83.3)	25(83.3)	22(73.3)	10(33.3)
	no(%)	2(6.7)	6(20)	5(16.7)	5(16.7)	8(26.7)	20(66.7)
Item 11	yes(%)	26(86.7)	23(76.7)	24(80)	22(73.3)	23(76.7)	12(40)
	no(%)	4(13.3)	7(23.3)	6(20)	8(26.7)	7(23.3)	18(60)

Note: yes: 1–3; no: 4–5; 1 = never; 2 = seldom; 3 = occasionally; 4 = often; 5 = always

reported knowing the words (Item 1) and understanding the content (Item 2) were much higher in the reading–writing and the reading–listening–writing tasks than in the listening–writing task, while the percentage for Item 3 (reread/relisten the source material) did not vary dramatically across the three tasks.

Multivariate analysis of ANOVA of the data in Table 6.2 was carried out to determine whether the above descriptive statistics were statistically significant. Levene's test of equality of error variances ($F_{\text{item } 1} = 1.46$, $p > .05$; $F_{\text{item } 2} = 2.08$, $p > .05$; $F_{\text{item } 3} = 1.95$, $p > .05$) showed that the statistics were suitable for multifactor variance analysis. Firstly, the significant differences between the more and the less proficient participants on all the three items ($F_{\text{item } 1} = 60.68$, $p < 05$; $F_{\text{item } 2} = 60.34$, $p < .05$; $F_{\text{item } 3} = 7.77$, $p < .05$) confirmed that writing tasks exerted significant influences on their reported source use in all three aspects of using reading and/or listening material for comprehension, namely, knowing the words (Item 1), understanding the content (Item 2), rereading/relistening the source material (Item 3). Secondly, the two proficiency groups reported significant differences in knowing the words ($F_{\text{item1}} = 17.73$, $p < 05$) and understanding the source material content ($F_{\text{item2}} = 17.56$, $p < 05$) across the three writing tasks, but not in rereading

Table 6.3 Inferential statistics between more and less proficient participants in items 1–11 across the three writing tasks

Source	Item 1 (F/Sig.)	Item 2 (F/Sig.)	Item 3 (F/Sig.)	Item 4 (F/Sig.)	Item 5 (F/Sig.)	Item 6 (F/Sig.)	Item 7 (F/Sig.)	Item 8 (F/Sig.)	Item 9 (F/Sig.)	Item 10 (F/Sig.)	Item 11 (F/Sig.)
Proficiency level	60.68/ .00	60.34/ .00	7.77/ .01	26.24/ .00	2.75/ .10	22.84/ .00	17.32/ .00	8.22/ .01	12.50/ .00	14.41/ .00	12.80/ .00
Task type	17.63/ .00	17.56/ .00	.25/ .78	8.59/ .00	.94/ .39	.10/ .91	2.52/ .08	1.62/ .20	2.32/ .10	12.76/ .00	4.73/ .01
Proficiency level*task type	.45/ .64	1.32/ .27	.29/ .75	1.94/ .15	.41/ .66	2.75/ .07	.45/ .64	2.47/ .09	2.35/ .10	3.73/ .03	3.11/ .05

and/or relistening to the source material ($F_{item3} = .25, >05$). No interactive effects were detected between the three tasks and the two proficiency groups on each of the three items (see Table 6.3).

6.4.2.3 Using Source Material to Gather Ideas

In Table 6.2, we can see that for all three writing tasks, the differences in terms of the use of the source material for gathering ideas between the more and the less proficient groups were much larger on Item 4 (*I used the reading passage and/or the listening record to help me get ideas on the topic*) than on Item 5 (*I used only my own ideas in my writing, nothing from the reading passage and/or the listening record*). For Item 4, the more proficient participants reported using the source material to help them get ideas more than the less proficient group did across all three writing tasks. Besides, both proficiency groups reported using the source material more often in the reading-listening-writing task and the reading-writing task than in the listening-writing task. However, for Item 5, the variations between the two proficiency groups across the three writing tasks were not dramatic in terms of the descriptive statistics, and no more than 40% of the participants reported using only their own ideas in their writing across the three writing tasks. Inferential statistics indicated that such descriptive discrepancies between the two proficiency groups in relation to Item 4 were statistically significant (see Table 6.3). That is to say, more proficient participants were more likely to get ideas from the source material than the less proficient participants across the three writing tasks ($F = 26.24$, $p < .05$) and both the more and less proficient participants reported getting more ideas from the source material in the reading-listening-writing task and the reading-writing task than in the listening-reading material ($F = 8.59$, $p < .05$), which were similar to those in Items 1–3. However, no significant differences were detected for Item 5.

6.4.2.4 Using Source Material to Shape Opinions

For Item 6 (*The reading passage and/or the listening record helped me choose an opinion on the issue*), the more proficient group had more participants who tended to select opinions from the source material to help in their writing than the less proficient group across the reading–writing task (more: 24(80%), less: 22(73.4%)), the listening–writing task (more: 25(83.4%), less: 18(60%)), and the reading–listening–writing task (more: 29(96.7%), less: 20(67.7%)). Results from inferential statistical analysis showed that such differences between the two proficiency groups across the three writing tasks were statistically significant ($F = 22.84$, $p < .05$). However, the variations for both proficiency groups across the three writing tasks were not apparent and no significant difference was detected across the three writing tasks ($F = .10$, $p > .05$). No interactive effects were detected between the language proficiency groups and the writing tasks ($F = 2.75$, $p > .05$).

6.4.2.5 Using Source Material for Supporting Opinions

Items 7 (*I used some of the ideas from the reading passage and/or the listening record in my essay*) and 8 (*I used examples and ideas from the reading passage and/or the listening record to support my argument in my essay*) concern the use of source material to support opinions in writing. Table 6.2 shows that most participants reported using the ideas and examples from the source material in their essays across the three writing tasks, and no prominent variation tendency could be detected between the proficiency groups across the three writing tasks. On the other hand, being similar to Items 1–4, the more proficient participants reported using the examples and ideas from the source material to support their writing more than the less proficient group, and such differences were also found to be statistically significant between the two proficiency levels ($F_{item\ 7} = 17.32$, $p < .05$; $F_{item\ 8} = 8.22$, $p < .05$), but not across the three writing tasks (see Table 6.3). Again, no interactive effects between the language proficiency groups and the writing tasks were found on these two items ($F_{item\ 7} = .45$, $p > .05$; $F_{item\ 8} = 2.47$, $p > .05$; see Table 6.3).

6.4.2.6 Using Source Material for Language Support

When it comes to Item 9 *(I used some words from the reading passage and/or the listening record when I wrote)*, results similar to Items 7 and 8 were detected. To be specific, most participants held that they used the words in the source material across the three writing tasks, except for the less proficient participants' reports on the listening-writing task. For one thing, no prominent task type effects ($F = 2.32$, $p > .05$) or interactive effects between task type and proficiency group ($F = 2.35$, $p > .05$) could be detected in terms of item 9. For another thing, the more proficient participants reported using the words from the source material in their writing more than the less proficient group across the three writing tasks ($F = 12.50$, $p < .05$). As for Item 10 *(The reading passage and/or the listening record helped me write better)*, the more proficient participants' reports declined from the reading-listening-writing task (28(93.3%)) to the listening-writing task (25(83.3%)), with the reading-writing task (22(73.3%)) in between. For the less proficient group, the percentage in the listening-writing task was much lower than in the reading-listening-writing task and the reading-writing task. Only one-third of the less proficient participants (10 (33.3%)) reported that input from the listening material helped their writing in the listening–writing task. In addition, the two proficiency groups differed greatly in the reading-listening-writing task (93.3%−80% = 13.3%) and the listening-writing task (73.3%−33.3% = 40%), but not in the reading-writing task (83.3%−83.3% = 0).

On the one hand, multivariate analyses of ANOVA indicated that the more proficient participants used more words from the source material than the less proficient participants in the reading–writing task ($F_{item9} = 12.50$, $p < .05$). On the other hand, both proficiency level and task type exerted the main as well as the interactive effects on Item 10 (see Table 6.3). Together with further post hoc analysis, it was found that participants in both proficiency groups declared that the reading material in the reading–writing task and the reading and listening materials in the reading–listening–writing task could help them write better compared with the listening material in the listening–writing task. Besides, the differences between the two proficiency groups lay mainly in the reading-listening-writing task and the listening-writing task.

6.4.2.7 Using Source Material for Modeling Organization

For Item 11 (*I used the reading passage and/or the listening record to help me organize my essay*), the more proficient group reports dropped slightly from the reading-listening-writing task (26(86.7%)) to the listening-writing task (24(80%)), with the reading-writing task (23(76.7%)) in between. For the less proficient group, the percentage in the listening-writing task was much lower than that in the reading-listening-writing task and that in the reading-writing task. Only 40% of the less proficient participants held that the listening material in the listening–writing task helped them organize better. Besides, the two proficiency groups differed greatly in the reading-listening-writing task (86.7%−76.7% = 10%) and the listening-writing task (76.7%−40% = 36.7%), but not in the reading-writing task (80%−73.3% = 6.7%). All these variations are similar to those on Item 10.

Significant differences were detected between the two proficiency levels (F = 12.8, p < .05) across the three writing tasks (F = 4.73, p < .05), as well as their marginal interactive effects (F = 3.11, p = .05), concerning whether the source material functioned in terms of modeling organization. The marginal interactive effect (F = 3.11, p = .05) and the post hoc analysis demonstrated that the differences mainly came from the two proficiency groups' performances between the reading–writing and listening–writing tasks (F = −.50, p < .05) and between the reading–listening–writing and listening–writing tasks (F = −.62, p < .05).

6.5 Discussion and Conclusions

6.5.1 Source Use Amount

Firstly, almost all of the more proficient participants reported having used the source material, demonstrating that to some extent they had the ability and the awareness to deal with the material. Since such an ability and awareness are inseparable in the process of internalized academic writing (Serviss & Rodrigue, 2010), the source-related activities for this writing form can be advocated in English-language teaching to prepare students for academic writing. Secondly, the great differences among less proficient participants in reporting using the source material across the three writing tasks, on the one hand, indicated that the listening–writing task was much too difficult for this group, which might be attributed to their overall low English-language listening capability. In other words, the less proficient participants did not reach the threshold of understanding the listening material as required. Therefore, the application of integrated writing tasks, especially those involving listening material, should be considered carefully when it comes to the less proficient group. On the other hand, consistent with Zhang and Zhou (2014) and Zhang et al. (2015), the bimodal input via two different channels (in this case, the visual channel for the reading material and the auditory channel for the listening material) could help participants' writing compared with the monomodal input in the reading–writing and the listening–writing tasks. However, the more and the less proficient participants' high overall proportion of reported source use amount is not in line with Weigle and Parker's (2012) claim that only a small percentage of students borrowed extensively from source materials.

6.5.2 Source Use Pattern

Compared with their low-proficiency counterparts, the more proficient participants reported higher overall use of the source material for comprehension in terms of knowing the words and understanding and rereading the source material content across the three writing tasks, which is consistent with Plakans and Gebril (2012), Wu (2014) and Homayounzadeh et al. (2019). This indicated that the more proficient participants' overall high English-language proficiency, as well as their awareness of using source material with appropriate strategies, helped their writing. Moreover, the significant differences between the two proficiency groups in relation to Items 1 (*I could understand most of the words in the reading passage and/or the listening record*) and 2*(I could understand most of the ideas in the reading passage and/or the listening record*) between the reading–writing and reading–listening–writing tasks and between the listening–writing and reading–listening–writing tasks (specifically, more reported using the source material in the reading–writing and reading–listening–writing tasks than in the listening–writing task) showed that both proficiency groups' listening ability required improvement and, compared with the monomodal input in the listening–writing task, the bimodal input in the reading–listening–writing task could facilitate participants' writing. These findings correspond with those of Zhang et al. (2015). Meanwhile, the non-significant difference for Item 3 (*I often reread the reading passage and/or replay the listening record while I was writing*) across the three writing tasks showed that both proficiency groups tended to reread and/or relisten to the source material in the process of writing, which is consistent with some previous studies' results (Plakans & Gebril, 2012; Zhang & Zhou; 2016; Zhang et al., 2015).

As for using the source material to gain ideas, for one thing, the significant differences in Item 4 (*I used the reading passage and/or the listening record to help me get ideas on the topic*) between the two proficiency groups across the three writing tasks indicated that compared with the less proficient participants, the more proficient participants had a stronger intention to use the source material in the reading–listening–writing and listening–writing tasks, which might be attributed to the fact that the more proficient participants were aware of utilizing the source material and possessed the ability to deal with multimodal information while writing. Meanwhile, the less proficient participants' overall low listening comprehension capability limited their ability to deal with multimodal information (visual and auditory) simultaneously, which might also have led to the above results. Therefore, on the one hand, source material attributes (e.g., vocabulary, the degree of difficulty, broadcast speed, etc.) should be considered when such writing tasks are used for less proficient participants. On the other hand, low-proficiency participants need more frequent practice, since this has been shown to enhance participants' overall performance (Zhou, 2004). Results from the subsequent four semi-open questions reinforced the implications of the Likert scale data, explaining that low-proficiency participants' listening difficulties mainly lay in overall comprehension of the listening material, lack of lexical knowledge, challenges related to catching up with the broadcasting speed and focusing during the process, and interchanging properly between English and Chinese. Therefore, it is imperative to pay more attention to and enhance teaching and practicing with respect to participants' English-language listening ability.

> LRW(id: 3182012045): I can't understand the listening material. I couldn't understand it even though I knew the words.
>
> LRW(id: 3182012031): It was broadcasted too quickly; I couldn't follow it.
>
> HRLW(id: 3171204018): I could not get focused during the whole listening broadcasting.

For another thing, the low proportion of data on Item 5 (*I used only my own ideas in my writing, nothing from the reading passage and/or the listening record*), as well as the non-significant differences between the more and less proficient participants across the three writing tasks, indicated that all participants were aware of using the source material.

Although most participants in both proficiency groups used source material to shape (Item 6) and support their own opinions (Items 7 and 8) and to gain language support in terms of using the words from the source material (Item 9), the more proficient group had more participants that used the source material to help them choose and support their opinions and used the words from the source materials across the three writing tasks. This may reflect that many less proficient participants might be unable to deal with the various source material(s) in such a complicated writing process, which requires the ability to manage dual-modal input while writing due to their limited overall listening proficiency. Meanwhile, the significant differences between the more and the less proficient groups in terms of whether the source material helped them write better in the reading-listening-writing task and the listening-writing task as well as their non-significant differences in the reading-writing task might also resulted from the input of the listening material. These were also revealed in their responses to the semi-open questions.

> LLW(id:3182012034): I don't like it. The listening material makes it more difficult.
>
> LRLW(id: 3180110037): It was too difficult, especially the listening material.

The two proficiency groups held that the source materials helped them write better in the reading-listening-writing tasks than the other two integrated writing tasks. This claim is in line with that of Zhang et al. (2015), that is, the bimodal input in the reading–listening–writing task could facilitate participants' writing.

Although participants' overall high reported proportion of using the source material to help their organization mirrored Plakans' findings (2008), the more proficient group still outperformed the less proficient group in using the source material to help their organization. In addition, the two proficiency groups showed great differences between the reading–writing and the listening–writing tasks and between the reading–listening–writing and the listening–writing tasks in using the source material(s) for organization modeling. These differences indicate that the monomodal auditory input in the listening–writing task increases the degree of difficulty for both proficiency levels compared with the visual modal input in the reading–writing task and the bimodal input (visual as well as auditory) in the reading–listening–writing task. Data from the fourth semi-open question (*Do you like this kind of writing task with listening material input? Why?*) in the listening–writing task showed that almost all of the less proficient participants reported that they could not understand the meaning of the listening material. They stated that the listening material input hindered their thinking to some extent, thus leading to direct plagiarism of its words and phrases. Zhang and Zhou (2016) had similar findings. Therefore, attention should be paid to source material attributes.

6.5.3 Conclusions

This study explored two different groups of Chinese EFL learners' source use across three different integrated writing tasks in terms of the volume and pattern of source use. Almost all participants confirmed using the source material, indicating that they had the intention to use and the awareness of using source materials to help their writing. However, due to their different listening ability thresholds (Plakans & Gebril, 2012) and multimodality processing abilities, the more proficient participants differed in their ways of utilizing the source material across the reading–listening–writing, reading–writing, and listening–writing tasks with respect to comprehension, gathering ideas and opinions, and gaining language and organizational support. These results highlight the implications of integrated writing tasks in English-language teaching and testing environments, and also suggest that attention should be paid to source material attributes such as input modality and level of difficulty to ensure that these are best suited to different proficiency levels.

Although the present study explored how different Chinese EFL learners used source materials across the reading–listening–writing, reading–writing, and listening–writing tasks, it did not encompass the other integrated writing tasks, for example, the story continuation writing (Ye & Ren, 2019) and summary writing tasks. Moreover, like most previous writing process studies, the present study adopted a questionnaire and semi-open questions to collect evidence. However, if employed to investigate aspects of integrated task performance, some recent research methodology innovations, for instance, the use of stimulated recalls, eye-tracking, and keystroke logging, could help yield more meaningful research findings. Furthermore, this study only addressed college-level participants. Participants who are at other language learning phases, such as the master's or doctorate level, etc., might provide additional inspiring data.

Funding This research was funded by Project of Educational Commission of Guangdong Province of China (2017WQNCX017).

References

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia, 7*(10), 1–27.

Chan, S. (2018). *Defining integrated reading-into-writing constructs: Evidence at the B2 C1 interface* (English profile series studies 08). Cambridge University Press.

Cheong, C. M., Zhu, X., & Liao, X. (2018). Differences between the relationship of l1 learners' performance in integrated writing with both independent listening and independent reading cognitive skills. *Reading & Writing, 31*, 779–811.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation of TOEFL. *Assessing Writing, 10*(1), 5–43.

Flower, L. S., & Hayes, J. R. (1984). A cognitive process theory of writing. *College Composition and Communication, 4*, 365–387.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Working Papers in Second or Foreign Language Assessment, 7*, 47–84.

He, L. Z., & Sun, Y. X. (2015). Investigating the effects of prompt characteristics on Chinese test-takers' integrated writing performance. *Foreign Language Teaching and Research, 47*(2), 237–250.

Homayounzadeh, M., Saadat, M., & Ahmadi, A. (2019). Investigating the effect of source characteristics on task comparability in integrated writing tasks. *Assessing Writing, 41*(1), 25–46.

Liu, F. L., & Stapleton, P. (2018). Connecting writing assessment with critical thinking: an exploratory study of alternative rhetorical functions and objects of enquiry in writing prompts. *Assessing Writing, 38*(1), 10–20.

Michel, M., Révész, A., Lu, X., Kourtali, N., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research, 36*(3), 1–28.

Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38*(1), 21–36.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129.

Plakans, L. (2015). Integrated second language writing assessment: Why? What? How? *Language and Linguistics Compass, 9*(4), 159–167.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*(1), 18–34.

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217–230.

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*(1), 98–112.

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing, 28*, 1–19.

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31–53.

Serviss, H. R., & Rodrigue, T. K. (2010). Writing from sources, writing from sentences. *Writing & Pedagogy, 2*(2), 177–192.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*(1), 36–47.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*(2), 118–133.

Wu, Y. (2014). Explore test-takers' source use strategies in an EFL reading-to-write integrated task. *Computer-assisted Foreign Language Education, 159*, 63–69.

Xu, H., & Gao, C. F. (2007). An empirical study on integrating reading with writing in EFL teaching. *Modern Foreign Languages, 30*(2), 184–190; 220.

Yang, H. C. (2009). *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches.* Unpublished doctoral dissertation. The University of Texas.

Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly, 46*(1), 80–103.

Ye, W., & Ren, W. (2019). Source use in the story continuation writing task. *Assessing Writing, 39*(1), 39–49.

Zhang, X. L., & Zhou, Y. (2014). The effect of task type on Chinese EFL learners' writing performance. *Modern Foreign Language, 37*(4), 548–558.

Zhang, X. L., & Zhou, Y. (2016). The influence of task type on different levels of Chinese EFL learners' English writing process. *Journal of PLA University of Foreign Languages, 39*(1), 87–95.

Zhang, X. L., Zhou, Y., & Zhang, S. Y. (2015). Chinese EFL learners' writing strategy use in reading-to-write and reading-listening-writing tasks. *TESOL International Journal, 10*(2), 97–109.

Zhou, D. D. (2004). Effects of task frequency on story retelling. *Journal of PLA University of Foreign Languages, 27*(5), 41–45.

Zhu, X., Li, X., Yu, G., Cheong, C. M., & Liao, X. (2016). Exploring the relationships between independent listening and listening-reading writing tasks in Chinese language testing: Toward a better understanding of the construct underlying integrated writing tasks. *Language Assessment Quarterly, 13*(3), 167–185.

Dr. Yan Zhou works at Southern Medical University. Her research interest is language assessment and language testing, especially the development and validation of integrated assessment tasks. She has been involved in several research projects, including the development of China's Standards of English (Speaking Scales).

Ke Bin is a Ph. D candidate at Guangdong University of Foreign Studies. He is interested in the research of language testing and assessment of vocational English competence. He has been actively involved in a number of research projects in the field of language testing and assessment, including the construction of vocational English proficiency scales.

Chapter 7
The Writing Sub-scales of the China's Standards of English Language Ability: Construction and Application in Writing Assessments in China

Mingwei Pan

Abstract This chapter is composed of two parts. The first part introduces the conceptualization, development and validation of the China's Standards of English language ability (CSE), with a focus on the sub-scales of English writing in the Chinese EFL context. Based on the Communicative Language Ability model (Bachman, Fundamental considerations in language testing. Oxford University Press, Oxford, 1990; Bachman and Palmer, Language testing in practice: designing and developing useful language tests. Oxford University Press, Oxford, 1996), the construct of writing ability in the CSE integrates organizational knowledge, pragmatic knowledge, text typology knowledge and writing strategies. As such, the sub-scales include sets of "can-do" statements that describe what English learners at different levels can do. In addition, how the writing sub-scales were validated is also briefly introduced in this part. The second part is concerned with the praxes of how the writing sub-scales can be applied to assessments of English writing in the Chinese EFL context. On the one hand, the self-assessment descriptors can be localized to enrich the feedback of writing tests to test takers. On the other hand, the writing sub-scales can also have a key role to play in formative assessment, such as peer assessment. Examples and tentative guidelines of such applications are provided in this chapter for better illustration.

Keywords China's Standards of English Language Ability · Writing assessments · Writing sub-scales

M. Pan (✉)
Shanghai International Studies University, Shanghai, China
e-mail: mwpan@shisu.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_7

7.1 Introduction

China may have the largest population of English learners in the world. As is estimated, as of 2000, among 415 million Chinese foreign language learners, approximately 390 million of them had learnt English (Wei & Su, 2012). However, given an imbalance in educational resources in China and inconsistencies between different educational phases (mainly elementary, secondary and tertiary education), the huge number of English learners does not necessarily mean a large number of successful English communicators. What is more, most locally developed English tests, if not all, have not completed any benchmarking or standard setting processes. As a result, different test results cannot be mapped onto a common scale. Although language proficiency scales, such as Common European Framework of Reference for Languages (CEFR), can serve as a good reference for standard setting, considering the Chinese EFL context, existing scales may not be directly referred to. As such, the China's Standards of English Language Ability (thereafter CSE, China's Ministry of Education and State Language Work Committee, 2018), were constructed through joint efforts of applied linguists and research students in and outside China, and were officially released in 2018.

Since the implementation of the CSE nationwide, how the CSE can be effectively and vigorously applied in assessment scenarios still remains to be explored. With the CSE writing sub-scales as a point of departure, this chapter is devoted to the English writing assessments in China, with a particular view to the application of the CSE writing sub-scales (CSE-W) to the assessments of English writing. Structurally, this chapter first introduces the rationale and construction of the CSE-W sub-scales. It then turns to the applications of the writing sub-scales into writing assessments with examples and guidelines.

7.2 Construction of the CSE-W

Figure 7.1 outlines the CSE-W construction procedure. As illustrated, the CSE-W construction involved four phases. In the first phase, the construct of CSE-W was defined based on the extant literature. The second phase primarily dealt with descriptor collection, from existing proficiency scales, test specifications, curriculums, etc. At the end of this phase, around 1300 descriptors of writing ability were pooled together. In the third phase, the CSE-W developers, based on the expert and teacher judgments, conducted screening and refinement of the descriptors. During this process, the developers removed the duplicates and revised the descriptors. At the end of this phase, about 350 descriptors survived.

The fourth phase was scaling. All the descriptors were spread into different sets of questionnaires, which were administered to teachers in different geographical locations and of various educational phases in China. They were supposed to rate the

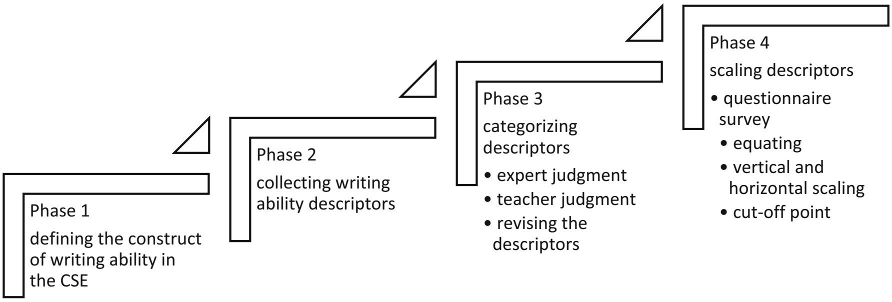


Fig. 7.1 The CSE-W development procedure

extent to which their students can perform in relation to each descriptor provided in the questionnaires. Equating was then conducted to ensure the comparability of different questionnaires, and further statistical analyses were conducted to determine the cut-off points (equated logit scores) of each level. In the final version, there are totally 299 CSE-W descriptors.

In general, the above phases parallel the development of the CEFR. Due to limited space, below are the details of Phase 1 and Phase 4 of the CSE-W construction (see Liu & Pan, 2019; Pan & Zou, 2020 for more details on the other two phases).

7.2.1 Defining the Construct

As the CSE-W sub-scales constitute an integral part of the CSE, the scale developers needed to be clear about the construct of language ability of the CSE. China's Ministry of Education and State Language Work Committee (2018: 1) defines it as "the ability to interpret and express intended meanings that learners and users of English exhibit when they perform language use tasks in a certain context or situation by applying their linguistic and nonlinguistic knowledge and communicative strategies". Therefore, the CSE is informed by the Communicative Language Ability (CLA) model (Bachman, 1990; Bachman & Palmer, 1996) to a great extent (see Liu & Wu, 2019 for more details). Following the above definition of language ability, the scale developers reviewed the related literature with a view to defining the construct of writing ability for the sub-scales. Two strands of literature were reviewed: (1) FL/L2 writing ability; and (2) existing proficiency scales of English writing.

7.2.1.1 FL/L2 Writing Ability

As part of the CSE, CSE-W also needs to fit into the definition of language ability of the CSE and also the CLA model in a broader sense. Mainly the following lines of inquiry were reviewed: (1) writing ability development, (2) writing quality, (3) cognitive models of writing, and (4) written text typology.

First, writing ability development is one concern of the CSE-W working definition. This is because the scaling of writing ability is perceivably consistent with not only different educational phases (or age ranges) in China but also the research findings in L2 writing development. For example, Schoonen et al. (2003) find L2 learners' writing ability develops in proportion to their exposure to the English language, and that they tend to transfer their knowledge of writing in the native language to their L2 writing. Similarly, Bazerman et al. (2017) provide an insightful discussion on writing development across the lifespan, where principles were proposed from different disciplinary perspectives. However, what characterizes such a development, such as development rates and salient features of each development phase, still needs further exploration.

Second, the measurement of the quality of FL/L2 writing is dependent on the text parameters, that is, what aspects/dimensions should be taken into account. Cumming et al. (2000), after depicting a rather complex picture of writing quality, suggest at least two dimensions of observations: (1) organization and expressiveness at a macro level, and (2) syntax and lexis at a micro level. In this regard, studies are more concerned with what to assess or observe, which is related to scoring criteria. For instance, Cumming et al. (2001) proposed 11 indicators and integrated them into (1) structure and organization, (2) content and idea, and (3) accuracy and fluency. These parameters are important indices of measurement regarding Chinese EFL learners' writing ability. In addition, at different levels of the CSE-W sub-scales, the writing quality foci may be shifted. For example, in the case of the higher level descriptors, assuming a high degree of accuracy of expression, may be more concerned with other parameters, such as appropriateness, in written production.

Third, cognitive models of writing (e.g., Alamartgot & Fayol, 2009; Grabe & Kaplan, 1996; Hayes, 1996, 2012; Kellogg, 1988, 1990) also drew much attention in the CSE-W construction. While these models mainly focus on the cognitive processing of writing, or writing strategy, they can be broadly classified in light of different writing stages: *planning* at pre-writing stage (Galbraith, 2009; Glynn et al., 1982), *executing* at while-writing stage (Hayes & Flower, 1980) and *proofreading and editing* at post-writing stage (Chanquoy, 2009; Fitzgerald, 1987). As this line of inquiry is directly related to writing strategies, it was also taken into account in the scale development and reflected in the CSE-W sub-scales.

The last strand of literature on writing ability relates to text typology, especially knowledge of various text functions. This can be more evident when we refer back to the definition of language ability of the CSE, where "language use tasks" are performed in various texts of functions. Therefore, in the case of writing ability, meanings of different types of texts are instantiated in written production. The systemic functional approach (Halliday, 1973, 1976; Halliday & Hasan, 1989; Halliday & Matthiessen, 2004) provides a taxonomy of text functions, ranging from narration, exposition, argumentation, description, interaction to instruction. As this classification is also applauded in various studies of text typology (e.g., Biber, 2006; Hatim & Mason, 1990), it was embedded into the construct of writing ability in the CSE-W.

7.2.1.2 Proficiency Scales of English Writing

At the onset of the CSE-W development, prevailing language proficiency scales outside China, such as CEFR (Council of Europe, 2001) were reviewed. Amongst the scales, it was found that the CEFR stood out to be most influential for the development and validation of the CSE-W. First, the CEFR can be regarded as one of the most influential existing language proficiency scales because it moves far beyond testing and promotes effective communications across cultures. For one thing, many other language proficiency scales are also claimed to be an adaption or a (sub)-branch from the CEFR, such as CEFR-Japan. For the other, international testing batteries or organizations spare no effort in aligning their tests or proficiency scales with the CEFR. Second, the CEFR is innovative in that it incorporates collaborative co-construction of meaning, and plurilingual and pluricultural competence (North & Panthier, 2016).

Nevertheless, specific to writing ability, the CEFR is not without limitations. For instance, its construct of writing ability seems more reflected in interactive written communication, thus understating the fact that writing may not be as interactive as oral production. In addition, its descriptors take "insufficient account of how variations in terms of contextual parameters may affect performances by raising or lowering the actual difficulty level of carrying out the target 'can-do' statement" (Weir, 2005: 281). Furthermore, while the CEFR claims to cover both proficiency and development in its six levels, it was found to be inconsistent (Alderson et al., 2006; Hulstijn, 2011; Norris, 2005). Despite all the above, it has to be admitted that the CEFR is pioneering and applicable in many contexts of proficiency scale construction. Although it cannot be readily used in China, where the context of teaching, learning and assessment seems quite different, it provides a good reference for the development of the CSE-W, particularly the construction procedure.

Based on the above literature review and the overall definition of language ability of the CSE, the construct of the CSE-W is defined as "in a repertoire of contexts, by adopting writing strategies and applying language knowledge, the ability to generate, construe or integrate information in written forms of texts across different functions for effective communication" (see Pan, 2017, 2018, 2019; Pan & Zou, 2020). In line with the general structure of the CSE, the CSE-W structure can be illustrated in Fig. 7.2. In the center lies *writing ability*, which is composed of *text typology knowledge*, *writing strategies* and *language knowledge*. Language knowledge, comprising *organizational knowledge* and *pragmatic knowledge*, is drawn from the CLA model; writing strategies (cognitive models of writing) and text typology knowledge are informed by the reviewed literature. It should also be

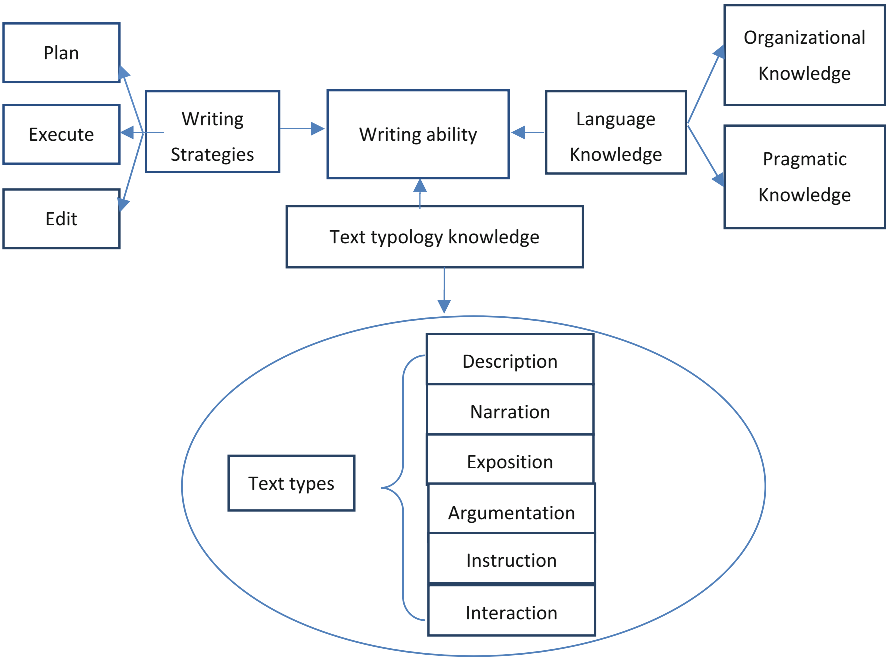


Fig. 7.2 The structure of the CSE-W

Table 7.1 The CSE-W subscales

Writing ability	Ability to write different text types	Written description
		Written narration
		Written exposition
		Written argumentation
		Written instruction
		Written interaction
	Writing strategies	Planning
		Executing
		Editing
	Self-assessment	

noted that writing ability development and writing quality are embedded in the scaling and descriptor content respectively.

Following the structure in Fig. 7.2, Table 7.1 shows all the components of the CSE-W. There are two sets of sub-scales: one covering text types, where *description*, *narration*, *exposition*, *argumentation*, *instruction* and *interaction* are described respectively, and the other covering writing as a process, including *plan*, *execute* and *edit*. Apart from those, there are also two summary scales – one for overall description and the other for self-assessment across different proficiency levels (see China's Ministry of Education and State Language Work Committee, 2018).

7.2.2 Scaling the CSE-W Descriptors

After the collection and revision of the descriptors, the CSE-W sub-scales were scaled into different proficiency levels.

All the descriptors were first categorized into their corresponding working levels. This was tentative because each descriptor should be labelled with a targeted proficiency level for the validation of scaling, or further calibration if deemed necessary. Therefore, in line with the CSE, the CSE-W descriptors were tentatively labelled with corresponding levels in relation to different educational phases in the Chinese EFL context. Altogether there are 9 levels, ranging from Level 1 at the lowest proficiency to Level 9 at the highest. Roughly speaking, CSE1 corresponds to Grade 3 elementary schoolers (a point where English learners in Mainland China start formal EFL instruction), CSE2 to elementary school leavers, CSE3 to junior high school graduates, CSE4 to senior high school graduates, CSE5 to non-English-major second-year students, CSE6 to non-English-major undergraduates, CSE7 to English major undergraduates, CSE8 to English major postgraduates, and CSE9 to professional language users such as professional translators and interpreters. It should be noted that the descriptors at lower proficiency levels are assumedly included in higher proficiency ones.

Then the 350 descriptors were spread into questionnaires for teacher rating. As it was impractical to request teachers to respond to one questionnaire containing so many items, namely the CSE-W descriptors, the CSE-W developers split the descriptors into different questionnaires with 50–70 items (CSE-W descriptors) in each. In order to equate responses from the same proficiency level across different questionnaires in the data analysis, about 20% items in each questionnaire were common anchors. It should also be noted that the anchor items were spread both horizontally and vertically. By horizontally, it means there are a number of "sound" items (by expert judgment) that are shared in all the questionnaire sets within a particular level. By vertically, it means a number of descriptors of the adjacent lower and higher levels were also embedded into one set. For instance, the CSE-W developers anchored some items of the CSE1-Set 2 into the CSE2- Set 1, some items of the CSE2-Set 2 into the CSE3-Set 1... so that all the questionnaire sets were horizontally and vertically connected.

All the questionnaires followed a 5-point Likert scale: 0, 1, 2, 3, 4. Teacher participants were supposed to assign a score to one of their observed students, who generally fall into the range of the intermediate level of that particular working level. For example, if teacher participants were junior high school EFL teachers, then they were supposed to aim at one of their students who is within the middle level of the observed cohort. On both ends of the Likert scale, 0 means "the student cannot perform what a descriptor says under whatever circumstances", whilst 4 means "the student can do that in whatever conditions". The score of 1 represents a marginal pass, where favorable conditions, such as "with the help of teachers" and "with sufficient preparations", should exist in assisting student's performance. In comparison, the score of 3 means satisfactory performance, which means student can do that in some unfavorable conditions, such as topic unfamiliarity. The score of 2 stands in the middle, indicating the student can do what is described in normal situations.

Table 7.2 Scaling results of the CSE-W descriptors

Levels	Means	Ranges
CSE-W1	−1.8088	~−2.39
CSE-W2	−1.5043	−2.39~−1.65 (0.74)
CSE-W3	−.9443	−1.65~−0.95 (0.70)
CSE-W4	−.2726	−0.95~−0.27 (0.68)
CSE-W5	.0643	−0.27~0.40 (0.67)
CSE-W6	.4464	0.40~1.08 (0.68)
CSE-W7	.6901	1.08~1.78 (0.70)
CSE-W8	1.3525	1.78~2.52 (0.74)
CSE-W9	1.7202	2.52~

After equating the CSE-W descriptors (see Liu & Pan, 2019 for more details on the methods), all the descriptors had equated difficulty estimates (logit scores). The CSE-W developers tried and compared three models of scaling the CSE descriptors: Model 1 is symmetrical scaling with equal intervals of logits between adjacent levels; Model 2 is symmetrically scaled but with slightly unequal intervals of logits between adjacent levels; Model 3 is an asymmetrical scaling with unequal intervals. As for a sound model, the scaling results should be able to accommodate the largest possible number of survived CSE-W descriptors, and they should be interpretable to a maximum extent. In addition, different CSE-W levels should be characterized with salient features that distinguish adjacent levels from each other.

As such, the CSE-W developers proposed a model that resembles Model 2 above, namely symmetrical scaling with slightly unequal intervals between adjacent levels. Table 7.2 lists the scaling results, which means a descriptor with a certain difficulty estimate falls into a range of the corresponding level. As is shown in Table 7.2, the zero logit takes place at the CSE-W5. The equated difficulty estimates spread the descriptors along a continuum, with approximately 0.7 logit as an interval. However, it should also be noted that at certain levels, the intervals are either more or less than 0.7 (e.g., 0.74 for CSE-W2 and CSE-W8). This model of scaling not only ensured comparatively equal intervals between adjacent levels, but also accommodated the existing descriptors to the best possible extent. After scaling, the number of the CSE-W descriptors was 299, and the tentative nine working levels were also maintained in the final version of the CSE-W.

7.3 Application of the CSE-W

As is stated, one of the educational intentions of the CSE is to provide references for various contexts of English language learning, teaching and assessment in China (Liu & Wu, 2019). Therefore, instead of being compulsory standards for learners of different educational phases, the CSE is just a point of reference. Therefore, when applying the CSE to assessment, users are encouraged to localize the descriptors. This section provides two references of application: applying the CSE-W sub-scales to self assessment and peer assessment respectively.

7.3.1 Applying to Self Assessment

In writing assessments, it is usually desirable that test developers provide detailed and individualized feedback for test takers. Nevertheless, most feedback for writing assessments, if not all, derives from the scoring criteria. Feedback related to whether, and if so how, test takers employ writing strategies is also far from sufficient. As a result, test takers might be roughly informed of their written output in some observable aspects, such as grammaticality and coherence. However, if test takers assess themselves both before and after a writing test, not only the quality of their written productions can be measured (for example via automatic scoring) and reported but whether they employ appropriate writing strategies may also be recorded. To a great extent, this can raise test taker's awareness of using appropriate writing strategies for a better enhancement of their writing. As students' learning moves forward, they may also trace their progress longitudinally. Below is an example of applying the CSE-W descriptors to an online writing test (see Pan et al., 2019 for more details).

The online writing test is aimed at providing students at the tertiary level in China with an access to assessing their own writing ability. It is expected that not only how well test takers write as judged by a self-developed in-house rating engine (automatic scoring), but also how they assess themselves regarding writing ability should be included.

As illustrated in Fig. 7.3, each time test takers experience four steps. The first step is a pre-test self assessment, where descriptors regarding easily observable writing ability are presented for self rating. The second step is a writing test. Test takers respond to the writing prompt in a timed test setting. The third step takes place immediately after the writing test. Test takers are supposed to score their own writing based on a number of easy-to-follow descriptors (see Appendix for an example of self rating descriptors) and also rate themselves regarding whether or not they use some strategies. Based on the steps above, a feedback report is generated, which provides a profile of the merits and demerits of test taker's writing quality as well as (in)consistency between their self-rating results and the de facto results as judged by automatic scoring. In addition, relevant learning materials are also recommended.

Therefore, in developing the above writing test, particularly the pre- and post-test self-assessment grids, the test developers referred to the CSE-W descriptors. For pre-test self-assessment grid, the following steps were followed. As it is mainly for writing ability, the descriptors should (1) be easy to understand for test takers, and (2) include features very observable in their writing. Take the following descriptor as an example. The original descriptor *can use various linking devices to make an argumentative piece of writing coherent across paragraphs* might not be fully understandable to test takers. Thus, localization and modification are very necessary. Test developers revised it as *I can use various methods to link paragraphs (e.g., sequential markers "First, Second..." or other possible markers "As explained in the previous paragraph...") to make my argumentation develop smoothly in paragraph transition.* It can be seen that the revised descriptor is more approachable and

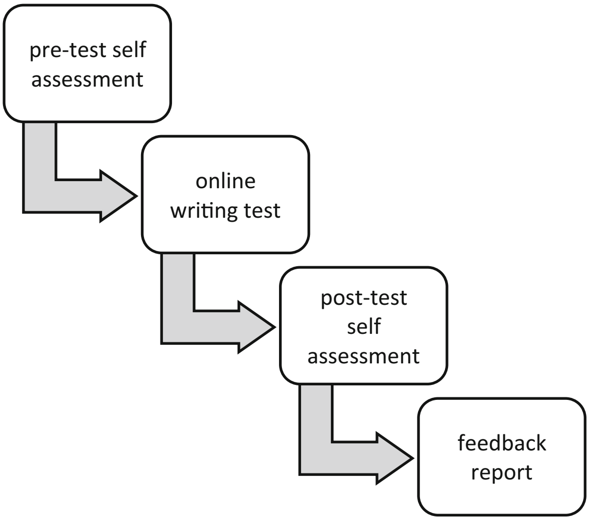


Fig. 7.3 Procedure of the writing test

test takers can also refer to the examples in the brackets to help assess themselves. For pre-test self assessment, test takers are supposed to rate themselves on a scale of *completely agree*, *agree*, *disagree* and *completely disagree*.

When it comes to post-test self-assessment grid, the descriptors should (1) be writing task related, and (2) include items of writing strategies. For example, in case of a writing task intended for a piece of argumentative writing, test developers referred to the *argumentation* descriptors (mainly concerning language quality, structure and coherence, content and idea, to fit in with the rating dimension of automatic scoring) and strategy descriptors. The former is provided for test takers to rate their own writing based on the descriptors, such as *I used various methods to link paragraphs in my argumentation*. For this part of the post-test self assessment, test takers are supposed to rate themselves on a scale of *completely agree*, *agree*, *disagree* and *completely disagree*. The latter is intended to check whether test takers employ appropriate writing strategies. Thus, the scale consists of *yes*, *no*, and *can't remember*. Naturally, localization and modification wherever necessary are also essential in developing this grid.

Feedback report consists of four aspects. The first aspect is a profile of test taker's pre-test self-assessment results. The second aspect is the results of automatic scoring (mainly concerning language quality, structure and coherence, content and idea) and test taker's comparative position among the same cohort of test takers. The third is a profile of test taker's self rating of his/her writing based on post-test self-assessment as well as the in(consistency) between self assessment and automatic scoring. The last aspect includes a summary of test taker's writing ability and some recommendations for further learning.

After several rounds of trial use, the self-assessment scales have been found to be reliable, valid and the scales of different proficiency levels are inclusive of writing ability (and strategies) (Pan et al., 2019). Test takers are reportedly found to be more critical of their writing performance and employ more appropriate writing strategies in the second- or third-taking of this online writing test (Pan et al., 2019). More importantly, in follow-up focus group interviews, test takers find that the descriptors are rather concrete in serving as a point of reference to make progress in English writing. Teachers also find these self-assessment descriptors useful because feedback to test takers includes not only writing quality but also use of writing strategies, thus enriching the feedback of writing test. It can be perceived that in similar contexts of writing assessment, the CSE-W descriptors can also be referred to in conducting self assessment.

7.3.2 Applying to Peer Assessment

Apart from applying the CSE-W descriptors to self assessment, they also have a role to play in peer assessment. This can be particularly true in the case of young learners. While self assessment is practical for English learners at the tertiary level, it may be less practical for young learners (Carless, 2005). This is because young learners tend to under- or over-estimate themselves (Matsuno, 2009; Patri, 2002) and may have just a haphazard knowledge of what the descriptors mean (Butler & Lee, 2006). In addition, when assessed, young learners may not be certain about their own writing performance. For instance, when asked to assess their own written output by directly referring to the self-assessment grid of the CSE-W, young learners need to be trained to reach the intended understanding of each statement. But the results may still be that teachers and students perceive the self-assessment differently (Butler & Lee, 2010). As such, it is suggested that the CSE-W descriptors, after adaptation and localization, can also be applied to peer assessment, where face-to-face interviews are conducted in assessing each other's writing ability. In the case, both parties of peer assessment can reach a comparatively stable consensus of what is assessed. Below is an introduction of an on-going study, where localization and revision of descriptors are emphasized when applied to peer assessment.

The study aimed at investigating the effectiveness of junior high school students' peer assessment of writing in classroom assessment. As suggested by the CSE, this group's ability to write English generally fits CSE-W3. Therefore, in order to generate the interview questions to be used in peer assessment, the researchers initially screened the CSE-W3 descriptors. However, as low- and high-achievers both exist in this age group, the descriptors of CSE-W2 and CSE-W4 were also looked into. In particular, the higher level (CSE-W4) might be more important as average students need to be informed of what they manage to do with help, or zone of proximal development (Vygotsky, 1978) in a dynamic assessment term. The following steps were tried in the study.

Step 1: Localize the descriptors. The initial step was selecting related descriptors. The relatedness in this case was (1) whether the CSE-W descriptors are pertaining to the teaching content and context, and (2) whether they are within the students' cognitive limits. Table 7.3 lists some descriptors from the CSE-W self-assessment scale.

In Table 7.3, regarding CSE-W3, descriptors (1)–(4) are related to writing ability; whereas the other descriptors are about writing strategies. Aware that descriptor (4) *I can write itineraries/schedules for class activities* is not covered by the teaching content, the researchers removed it. Likewise, regarding CSE-W4, the researchers also dismissed descriptors (4) and (5) (shaded descriptors), as they were not up to the cognitive level of junior high school students. As such, the remaining descriptors were selected for the next step.

Step 2: Rewrite the statements and construct the interview questions for peer assessment. As peer assessment is conducted in the form of interview in this case, the selected statements should be changed into questions, with a corresponding change in personal pronouns. For example, the descriptor *I can write short stories based on prompts given by my teacher* was changed into *Can you write short stories based on prompts given by your teacher?* Wherever necessary, the researchers also provided close-ended responses for students to choose from, such as *can do it in whatever circumstances*, *can do it with the help of teachers or your classmates* and *cannot do it in any cases*. These options may provide a lead-in for students to conduct peer assess, which can be followed by more evidence (to be explained below).

Step 3: Peer assessment training. Based on the revised descriptors, the researchers played a rather important role in ensuring that all students had a correct and complete understanding of all the interview questions. In doing so, more elaborations should be made. For example, when explaining *Can you use conjunctions to connect sentences?*, the researchers cited some examples of *conjunctions* for a

Table 7.3 Selected descriptors from the CSE-W self-assessment scale

Levels	Descriptors
CSE-W4	(1) I can write my views on topics I am familiar with or interested in. (2) I can write a summary of what I have read. (3) I can write a brief report on a certain social practice. (4) I can write my resume. (5) I can write brief news reports for media such as university newspapers. (6) I can write an outline before I start writing. (7) I can use a topic sentence to emphasize the main idea of a paragraph. (8) I can check my writing and correct errors in word use and connection.
CSE-W3	(1) I can write short stories based on prompts given by my teacher. (2) I can write compositions on familiar topics. (3) I can write letters or email to tell my friends about my current situation. (4) I can write itineraries/schedules for class activities. (5) I can collect useful words and sentences before writing. (6) I can use conjunctions to connect sentences. (7) I can check and correct obvious grammar errors.

clearer understanding of what this question really meant. In a similar vein, in elaborating on *Can you write compositions on familiar topics?*, the researchers also provided more examples of familiar topics, such as campus life and community volunteering, so that students can respond to their peers' questions more to the point. It should be noted that this step was regarded crucial especially for young learners.

Step 4: Conduct peer assessment in the form of interviews. By pairing up students, the researchers asked them to conduct an interview with each other. It may seem like self assessment as one student simply asks the other questions. However, in the case of young learners, interaction (reading the questions aloud) improves engagement, where more consciousness is raised and more attention is paid to accurately understanding the questions. Thus, peer assessment in the form of interviews may enhance the validity of peer assessment. In the process of being interviewed, both students are encouraged to provide more evidence, such as their writing performance in everyday language learning, besides just a yes or no answer. Therefore, peer assessment of this form was also regarded advisable as peers may challenge the response(s) they get from each other, based on their familiarity with the interviewee(s). In case of disagreement, the researchers stepped in for more objective responses.

Step 5: Provide timely feedback in peer assessment. After conducting peer assessment, the researchers collected the results and provided the feedback of students' writing ability in a timely manner. In real practice, the researchers also provided an inventory of students' strengths and weaknesses in English writing, such as improper use of adjectives and lack of writing strategies.

Although collecting the evidence for conducting peer assessment deriving from the CSE-W descriptors is still ongoing and the overall effectiveness is to be further investigated, the feedback from teachers is largely positive. On the one hand, teachers found their students quite engaged in collecting concrete evidence for judging their peer's writing ability, such as referring to their essay writing in the workbook and/or teacher's previous comments on their essay assignment. On the other hand, teachers found some students would gradually request more detailed feedback from their peers and teachers.

7.4 Conclusion

How language proficiency scales can be applied in various assessment scenarios usually invites much concern. The CSE-W, ever since its inception, is not an exception. Teaching practitioners, in particular, are rather interested in knowing the procedures of how the scales can be used in their teaching and assessment. Therefore, starting from an introduction of the CSE-W rationale and construction, this chapter looks into the applications of the CSE-W in some assessment contexts in China.

The CSE-W descriptors have a role to play in enriching the feedback of writing tests by including more details from test takers' self assessment results. This application goes beyond traditional feedback report in that apart from the results of writing quality, there is also more detailed information regarding self-reported writing strategy use. In addition, customized from the CSE-W descriptors, peer assessment can be perceivably conducted in the form of interview for young learners. The general procedures can be summarized as descriptor localization, interview question construction, peer assessment training and implementation as well as feedback providence.

However, it has to be admitted that there are also challenges when the CSE-W is applied in different assessment contexts. First, in the initial implementation of the CSE-W in China, one challenge is whether users, particularly frontline teachers, can familiarize themselves with the rationale and structure of the CSE-W. Unlike many other assessment instruments and proficiency scales, the CSE-W descriptors are organized in terms of text types and writing strategies. Second, although the element of writing quality is implicitly or explicitly embedded into the CSE-W sub-scales, they are still felt to be less observable in the descriptors. As a result, it can be challenging for users to align test takers' writing performance with particular sub-scale descriptors. In fact, NEEA has launched a round of revision for the CSE, with a view to offsetting this weakness. Third, the CSE-W sub-scales are not advisably to be applied in assessment scenarios intact. Adaption, localization or customization in wording or structuring should be made wherever and whenever necessary, depending on different assessment scenarios. In certain cases, one CSE-W level can be developed into fine-grained sub-levels. Fourth, though the Chinese and English versions of the CSE-W co-exist, the Chinese version was publicized prior to its English version, the counterpart of which might incur subtle distortions in the translation process. Thus, it is more advisable that users who can read Chinese should refer to the CSE-W Chinese version.

Appendix: An Example of Self Rating Descriptors

Descriptors ☹ 😐 ☺

1. My writing was highly related to the given topic.
2. In my writing, punctuation was used correctly.
3. In my writing, tenses were used correctly.
4. In my writing, different word(s) of similar meanings were used interchangeably to achieve lexical variety.
5. In my writing, topic sentences were written for different paragraphs.
6. In my writing, figures of speech were used for expressiveness.

……

References

Alamartgot, D., & Fayol, M. (2009). Modelling the development of written composition. In R. Beard, D. Myhill, M. Nystrand, & J. Riley (Eds.), *The SAGE handbook of writing development* (pp. 23–47). Sage Publications.

Alderson, J. C., Figueras, N., Kuiper, H., & Nold, G. (2006). Analyzing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 3–30.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bazerman, C., et al. (2017). Taking a long view on writing development. *Research in the Teaching of English, 51*(3), 351–360.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal, 90*(4), 506–518.

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5–31.

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assessment in Education: Principles, Policy and Practice, 12*(1), 39–54.

Chanquoy, L. (2009). Revision Processes. In R. Beard, D. Myhill, M. Nystrand, & J. Riley (Eds.), *The SAGE handbook of writing development* (pp. 80–97). Sage Publications.

China's Ministry of Education, & Language Work Committee. (2018). *The China's Standards of English Language Ability*. Higher Education Press/Shanghai Foreign Language Education Press.

Council of Europe. (2001). *Common European Framework of Languages for Reference: Learning, teaching and assessment*. Cambridge University Press.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. ETS.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. ETS.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481–506.

Galbraith, D. (2009). Writing about what we know: Generating ideas in writing. In R. Beard, D. Myhill, M. Nystrand, & J. Riley (Eds.), *The SAGE handbook of writing development* (pp. 48–64). Sage Publications.

Glynn, S. M., Britton, B., Muth, D., & Dogan, N. (1982). Writing and revising persuasive documents: Cognitive demands. *Journal of Educational Psychology, 74*, 557–567.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective*. Longman.

Halliday, M. A. K. (1973). *Explorations in the functions of language*. Edward Arnold.

Halliday, M. A. K. (1976). The form of a functional grammar. In G. Kress (Ed.), *Halliday: System and function in language* (pp. 101–135). Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford University Press.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Edward Arnold.

Hatim, B., & Mason, I. (1990). *Discourse and the translator*. Longman.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 1–27). Lawrence Erlbaum Associates.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*, 369–388.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229–249.

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 355–365.

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *American Journal of Psychology, 103*, 327–342.

Liu, J., & Pan, M. (2019). English language teaching in China: Developing language proficiency frameworks. In A. Gao (Ed.), *Second handbook of English language teaching* (pp. 415–432). Springer.

Liu, J., & Wu, S. (2019). *An investigation into the China's Standards of English Language Ability*. Higher Education Press.

Matsuno, S. (2009). Self-, peer- and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100.

Norris, J. M. (2005). Book review: Common European framework of reference for languages: Learning, teaching, assessment. *Language Testing, 22*(3), 399–405.

North, B., & Panthier, J. (2016). Updating the CEFR descriptors: The context. *Cambridge English Research Notes, 63*, 16–24.

Pan, M. (2017). Towards exemplary writing activities for the China's Standards of English: A Systemic-Functional-Linguistics text typology perspective. *Foreign Language World, 2*, 17–24.

Pan, M. (2018). Investigating the writing scales of the China's Standards of English Language Proficiency: A perspective of writing ability development. *Foreign Language in China, 3*, 145–152.

Pan, M. (2019). The construction of the writing sub-scales of the China's Standards of English: From theories to practices. *Foreign Languages in China, 3*, 38–45.

Pan, M., & Zou, S. (2020). *An investigation into the China's Standards of English writing sub--scales*. Higher Education Press.

Pan, M., Song, J., & Deng, H. (2019). Developing and validating the self-assessment scales in an online diagnostic test of English writing. *Foreign Language Education in China, 6*, 33–41.

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment. *Language Testing, 19*(2), 109–132.

Schoonen, R., Gelderen, A. V., Glopper, K. D., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*(1), 165–202.

Vygotsky, L. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wei, R., & Su, J. (2012). The statistics of English in China. *English Today, 28*(3), 10–14.

Weir, J. (2005). Limitations of the Common European Framework of Reference for Languages (CEFR) for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Dr. Mingwei Pan is Professor of Applied Linguistics at Shanghai International Studies University. His main research interests include large-scale test development and validation. Dr. Pan is now Member-at-Large of the Asian Association for Language Assessment and also Board Member of the China's Association of Language Testing and Assessment.

Part II
Research and Practice of English Language Writing Assessment in China

Chapter 8
Research and Practice of English Language Writing Assessment in China: An Introduction to Part II

Liz Hamp-Lyons

Abstract Chapter 8 begins with an overview of research and practice in English language writing assessment around the world and highlights the contribution of this edited collection to a detailed picture of what is possible in Chinese L1 contexts, from 'home-grown' tests of writing proficiency in English to major international tests, to explorations of newer practices. A chapter-by-chapter summary is then provided for Chapters 9 to 14, which have a clear focus on English writing assessment for improving teaching and learning in the Chinese contexts. The topics covered in Part II include the conceptual framework for academic writing assessment, the development and use of scoring rubrics, validation of a writing proficiency scale of Business English proficiency, effective feedback implementation, focused written corrective feedback, and teacher training on using student writing portfolios.

Keywords Educational assessment · Assessment for learning · Scoring rubrics · Feedback · Portfolios

In our Preface we discussed the significance and long history of assessing knowledge through writing. In the Introduction to Part I Yan Jin discussed the introduction of written exams in and of English in China from 1987. This was also the point from which the use of international tests of English such as ELTS/IELTS and TOEFL became more common, in line with the growth in the academic exchange market and the popularity in some countries for young people to go to an English-speaking country to do graduate work. In the UK and the US, assessing knowledge through written examinations became increasingly prevalent from the early 1900s, and by the late 1970s, written exams had become ubiquitous as the enrolments in colleges and universities grew rapidly. The 1970s was also a period of rapid expansion in the number of students from outside the UK and US wanting to study disciplines key to their own country's economic development at prestigious western universities began to grow. Increasingly significant funding from USAID and the British Government as well some from Canada and Australia was made available for outstanding

L. Hamp-Lyons (✉)
University of Bedfordshire, Luton, Bedfordshire, UK

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_8

overseas students, and by the end of the 1970s the presence of such students was becoming a prominent element in universities, and a sociocultural as well as economic asset to the host countries (Hyland & Hamp-Lyons, 2002; Jordan, 2002).

Some of our authors have themselves followed that path, and these varied backgrounds reflect in microcosm how research and practice in English language writing assessment is now taking place around the world, as more and more countries see the ability to write competently in English as a valuable skill in international business, international politics, and the exchange of academic knowledge. The international reach of research in writing assessment is reflected in the growth of the international journal *Assessing Writing* from its beginning in 1994, which arose from a series of small conferences on this subject sponsored by the University of New York and accessible to almost no-one outside the United States of America. From 1994 to 2001 very few papers from beyond the US were published (or submitted). In 2002, under a new publisher, Elsevier, the journal began to be edited by Liz Hamp-Lyons, who had worked at US universities from 1986 to 1996, but is British with experiences in Europe, Asia and Australia. The journal began to solicit articles internationally under Hamp-Lyons, and since 2018 under its new Editors Martin East and David Slomp. The July 2020 issue had articles from authors working and studying in Hong Kong SAR, Australia, mainland China, the US, Lebanon, Italy, the UK, Germany and the Republic of Korea.

Educational systems as well as national and local cultures vary significantly, as Lam's chapter mentions. Educational assessment journals often refer to differences in education systems, such as class size, and in learning styles, as being impediments to 'progress' in educational practice. This is true to some extent: but we hope that the chapters in this edited collection will contribute to a more detailed picture of what is possible in Chinese L1 contexts, from 'home-grown' tests of writing proficiency in English to major international tests, to explorations of newer practices. The use of scoring rubrics that define how writing is to be perceived and valued is becoming widely accepted, and is being used increasingly in what we might call 'assessment for learning' classrooms as well as in large-scale writing tests. The practice of giving feedback rather than only marking student writing is now prevalent, and teachers are becoming increasingly skilled at providing useable feedback to their learners without giving up a personal life. It is increasingly understood that certain kinds of assessment can be more beneficial to learners than others as well as more rewarding for teachers than others. Furthermore, the use of fully researched and validated scoring rubrics, and the rapidly increasing literature on how feedback supports learning as well as formal assessment, have supported the use of student writing portfolios as a record and a measure of development in writing by making these key affordances available to students as they prepare to create their portfolios. Our collection of studies does not include any attention to self- and peer-assessment or to collaboration: but we know that in mainland China the practice of learners establishing their own study groups is common, much more common than in, for example, Hong Kong. We are also aware that our collection does not have a chapter on the increasingly common and significant area of automated scoring of writing, nor on the perhaps more controversial use of automated writing feedback. Indeed, there is still a lot to do if the field is to reach an agreed understanding of what "best practice" in writing assessment means, but there is also plenty to be celebrated.

8.1 Outline of Chapters 9 to 14

In Chap. 9 Cecilia Guanfang Zhao focuses on academic writing and critiques the long-standing practice of assessing writing with a single, context-less task used to elicit impromptu writing within a strict time limit: what Hamp-Lyons & Kroll (1997: 18) have called a "snapshot approach" to writing. Having overviewed the existing theoretical models of writing ability, she argues that the abilities underlying successful academic writing go beyond linguistic correctness to demand real cognitive engagement, and a contextually based social interaction between the writer and her. Zhao then looks at the writing components of four large-scale tests in China, and finds weaknesses in them all when judged by the models of writing ability now accepted. She proposes a conceptual framework for writing assessment that would reduce construct under-representation by finding ways to evaluate writing process and to bring in some role for social interaction. These are ambitious goals, and Zhao illustrates her thinking with a useful Figure; however, there is more to be explored in relation to the building into writing tasks a role for engagement with multiple texts and other voices.

In Chap. 10, Li Liu and Guodong Jia also report on a study of a rating scale developed in and for China, with the important difference that this one was designed for use within their own institution. Liu and Jia attempt to apply elements of the current, argument-based, approach to validation. Messick (1989) emphasized the need for "empirical evidence and theoretical rationales [to] support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." (p. 13) The study provides substantial useful information for the further development of this local scoring instrument and method: but as the authors acknowledge, the study focused only on the Evaluation and Explanation components of an argument-based approach, collecting two types of empirical evidence and putting forward several rationales in arguing for the validity of the instrument developed for the Renmin University and did not collect backing evidence for the extrapolation, utilization and consequences inferences in the argument-based model. Their discussion of raters' comments on the 'holistic' scale echoes arguments made by Hamp-Lyons (e.g., 2016) about the choice between holistic and multi-trait/analytic scoring instruments and methods.

In Chap. 11 Li Wang correctly identifies the paucity of validation studies of writing assessment instruments aimed at specific purposes teaching and learning. She reports a mixed methods study designed to validate a writing proficiency scale of Business English proficiency, which, as she describes, was based on the CEFR and CSE frameworks. The early parts of her chapter make clear the large amount of work that went into the development of this new scale. Her own small-scale validation study which uses ten ‘experts’ understandings of and views on the descriptors in the scale is thorough and interesting. The selection of 5 ‘language’ experts and five ‘domain’ experts made possible interesting and potentially very valuable insights, and is reminiscent of some much earlier ESP projects using a similar approach which were, regrettably, limited by both lack of research funding and heavy workloads from pursuing valuable data to practical conclusions. Drawing closely on her quantitative and qualitative analyses, she is able to identify descriptors and levels needing changes or further investigation. The Chapter can be of value within China and well beyond.

Chapter 12, by Jing Yang, addresses an area of growing interest for researchers in writing assessment and for classroom writing teachers, i.e., feedback. Yang begins her chapter by talking about the needs of writing teachers, and this focus on teachers runs through the chapter. Chinese college and university teachers have been aware of the potentials of feedback for 25 years or more, but Yang reminds us of the difficulties of effective feedback implementation where there are large classes and an abiding emphasis on test success. Her two project participant teachers were experienced and taught fairly small classes. Yang’s study used methods now familiar in writing feedback research, but her description of her analysis is unusually detailed and illuminating, as is her use of quotes from teachers and students. These two teachers professed, and put into practice, a strong belief in the value of feedback and clearly had developed skills in teaching students its value and guiding them towards self-reliance in giving feedback to others and to themselves. The emphasis on creating links between teaching, tasks and rubrics helps build a coherent approach for students to make the most of the feedback they get.

Chapter 13 is by Icy Lee, Na Luo and Pauline Mak, and looks at a very different aspect of feedback. There has been, and remains, a range of views about what kind of written feedback to give and when to give it. Lee, Luo and Mak argue that too much written corrective feedback can be counter-productive for teachers as well as for students, providing plenty of evidence, while also acknowledging that some studies have shown that it can be effective. Lee et al. go on to argue for focused rather than written corrective feedback. However, because in their experience most existing studies on focused WCF lack ecological validity they designed a study that used a diagnostic assessment approach to providing focused written corrective feedback. Their choice of errors to focus on was based on previous studies of error patterns, and the data came from four secondary classes in Hong Kong, making this chapter unusual in this book. The findings of the error pattern data are interesting, but what readers may find most helpful is the detail about the methodology and the argument made for using diagnostic assessment in the classroom before finalizing teaching plans, rather than relying on ad hoc observation for target error selection which, Lee et al. argue, otherwise teachers will miss a valuable chance to connect WCF with pre-writing instruction on language features.

In Chap. 14 Ricky Lam continues his ground-breaking work on portfolio writing assessment with second language learners. In this study Lam explores whether providing teachers with assessment training necessarily leads to real competence and confidence in assessment – that is, to assessment literacy. Reporting on a small-scale study with three members of an assessment training course involves careful data collection designed to provide him with rich feedback on participants' responses to the training. Before the course began, a questionnaire showed that these three teachers' prior knowledge of and attitudes to assessment was influenced by their personal teaching experiences. After the training they were positive about the training itself, but through post-training interviews and written assignments Lam was able to probe more deeply and identify individual, institutional, and cultural issues that influenced each of them more or less positively in terms of putting portfolio assessment into action. He characterizes each teacher: Rebecca is an *inquisitive practitioner* of portfolio assessment; Joan believes in the values of high stakes testing for the Chinese culture, and is a *disciple*; Taylor became confident and engaged, actively planning teaching reforms: Lam calls her a *game changer*.

References

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing*. https://doi.org/10.1016/j.asw.2016.06.006

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community, and assessment* (TOEFL monograph series report no. 5). Educational Testing Service.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for Academic Purposes, 1*(1), 1–12.

Jordan, R. R. (2002). The growth of EAP in Britain. *Journal of English for Academic Purposes, 1*(1), 69–78.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Liz Hamp-Lyons began her career as a teacher of English for Academic Purposes, before researching the writing test component of ELTS (the precursor to the IELTS) for her PhD at the University of Edinburgh. Much of her PhD was published, with other chapters, in *Assessing Second Language Writing in Academic Contexts* (1991). She worked at the Universities of Michigan and Colorado before becoming Chair Professor and Department Head at the Hong Kong Polytechnic University. For many years she edited the journal *Assessing* Writing (Elsevier), and founded the *Journal of English for Academic Purposes* in 2002. She has consulted on large scale as well as more local writing assessments.

Chapter 9
Theory-Based Approach to Academic Writing Assessment in Higher Education: A Conceptual Framework for Assessment Design and Development

Cecilia Guanfang Zhao

Abstract An examination of the current writing assessment practices indicates that unlike measurement theory, "writing theory has had a minimal influence on writing assessment" (Behizadeh and Engelhard, Assess Writ 16(3):189–211, 2011: 189). Despite the widely accepted theoretical conception of writing as a cognitive process and social practice situated in a particular socio-cultural context, the most often employed writing assessment task is still prompt-based impromptu essay writing, especially in large-scale English as a Foreign Language (EFL) assessment contexts. However, assessment specialists have long called into question the usefulness of impromptu essay writing in response to a single prompt. As a response to the above observation of the lack of theoretical support and generalizability of results in our current writing assessment practices, this chapter seeks to propose and outline an alternative writing assessment design informed by and reflecting more faithfully theoretical conceptions of writing and language ability. The chapter starts with a brief review of the existing writing theories and a survey of the current practices of writing assessment on large-scale high-stakes EFL tests, with a particular focus on those within the Chinese context. By juxtaposing theoretical conceptions of the construct and the actual operationalization of this construct on these EFL tests, the chapter highlights several salient issues and argues for an alternative approach to writing assessment, especially for academic purposes in higher education. A conceptual framework for such an assessment is presented to illustrate how writing theories can be used to inform and guide test design and development. The chapter ends with a discussion of the value and practical implications of such assessment design in various educational or assessment settings, together with potential challenges for the developers and users of this alternative assessment approach.

Keywords Conceptual framework · Writing assessment · Assessment design · Academic writing

C. G. Zhao (✉)
University of Macau, Taipa, Macau, China
e-mail: czhao@um.edu.mo

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_9

An examination of the current writing assessment practices indicates that unlike measurement theory, “writing theory has had a minimal influence on writing assessment” (Behizadeh & Engelhard, 2011: 189). Despite the widely accepted theoretical conception of writing as a cognitive process and social practice situated in a particular socio-cultural context, the most often employed writing assessment task, other than the earlier discrete-point test items, is still the prompt-based impromptu essay writing, especially in large-scale English as a Foreign Language (EFL) assessment contexts. However, assessment specialists have long called into question the usefulness of impromptu essay writing in response to a single prompt (Cho, 2003; Crusan, 2014). As a response to the above observation of the lack of theoretical support and generalizability of results in our current writing assessment practices, this chapter seeks to propose and outline an alternative writing assessment design informed by and reflecting more faithfully theoretical conceptions of writing and language ability. In the rest of the chapter, I will first offer a brief review of the existing writing theories developed over the past 50 years or so, and then survey the current practices of writing assessment on large-scale high-stakes EFL tests, with a particular focus on those within the Chinese context. I will then synthesize the problems in the current EFL writing assessment practices by comparing and contrasting our theoretical understanding of the construct and the actual operationalization of this construct on these EFL tests, and argue for an alternative approach to writing assessment, especially for academic purposes in higher education. A conceptual framework for such an assessment will then be presented to illustrate how writing theories can be used to inform and guide test design and development. The chapter will end with a discussion of the value and practical implications of such assessment design in various educational or assessment settings, together with potential challenges for the developers and users of this alternative assessment approach.

9.1 Theoretical Models of Writing/Language Ability

Existing theoretical models of writing or language ability over the past 50 years have changed from a more static and text-focused view to an increasingly more dynamic and contextualized conception of this construct. For example, writing was once conceptualized largely as mechanical and linguistic accuracy (Hatfield, 1935) and a set of linear processes (Britton et al., 1975; Rohman, 1965), before the influential cognitive process model (Flower & Hayes, 1981; Hayes, 1996; Hayes & Flower, 1980) presented it as a series of non-linear hierarchical mental processes and cognitive activities. Such activities as goal setting, generating, organizing, translating, reviewing, and monitoring, according to Flower and Hayes, could happen at any stage of the composing process, interacting with other factors such as the task environment and individual writers’ motivation, affect, and long-term and working memory capacity. This primary focus on cognition and individual writers was soon criticized for ignoring the sociocultural context in which any act of writing is situated, and for its ineffectiveness in preparing students, especially L2 students, for academic writing tasks they would encounter in actual educational settings (Horowitz, 1986; Hyland, 2003; Johns, 1991; Spack, 1988; Swales, 1990). Following this sociocultural turn, researchers and practitioners turned next to genre-based conceptions of writing from various perspectives, including the systemic functional linguistics, English for specific purposes, and rhetorical genre studies approaches (cf. Bawarshi & Reiff, 2010; Hyon, 1996; Johns, 2008). A general consensus is that "genres are both social and cognitive" (Johns, 2008: 239); therefore, the analyses of "context, complex writing processes, and intertexuality" are all critical (Johns, 2011: 64).

In addition to these theoretical models of writing, more general conceptualizations of language ability also abound (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale & Swain, 1980). The most influential, particularly for use in test design and development, is probably Bachman and Palmer's (1996) model of communicative language ability, where language knowledge and strategic competence together define the construct of language ability. The language knowledge dimension is further broken down to organizational and pragmatic knowledge, with the former covering grammatical and textual knowledge, while the latter functional and sociolinguistic knowledge. The strategic competence dimension, on the other hand, covers a series of metacognitive strategies of goal setting, planning, and appraising during actual language use. Bachman and Palmer (2010) believe that an individual's language knowledge, strategic competence, topical knowledge and affective schemata, and the external factors of language use task and situation, together comprise a conceptual framework for language use.

Beyond general descriptions of writing or language ability, existing literature also offers various conceptions of academic literacy/literacies in particular, although covering the similar perspectives of writing as cognitive processes (academic literacy, Scardamalia & Bereiter, 1991) and social practices (academic literacies, Lee & Street, 1998, 2006). While it is beyond the scope of this chapter to fully unpack the whole field, interested readers may find Bloome et al.'s (2018) review informative. Here, I will briefly review one such model proposed by Snow and Uccelli (2009) when theorizing the challenges of academic language use for native and non-native English speakers alike. In this nested, pragmatics-based model of academic language, Snow and Uccelli (2009) argue that the ultimate purpose of academic literacy practice is to achieve "the two ubiquitous features of communicative tasks—representation of self and of one's message—under particularly challenging conditions" (p. 122). According to their model, at the fundamental level is one's ability to "organize discourse," using discourse markers and reference terms to signal metatextual relationships and conform to conventions of a particular academic (often also technical) discourse community. This level of academic language ability is nested in a higher-order ability to "represent the message," which involves the proper use of "approved academic genres," appropriate level of detail and information for the intended audience, and the representation of "abstract, theoretical constructs, complicated interrelationships ... and other challenging cognitive schemas," while explicitly acknowledging "sources of information/evidence"

(p. 123). This ability to represent the message is further nested in a yet higher-order ability to “represent the self and the audience,” which entails effective academic voice and identity construction and establishment of a co-membership with an intangible, non-interactive, expert academic audience, through explicit display and extension of one’s knowledge and acknowledgement of “the epistemological status of one’s claims” (p. 123).

Despite the many models of writing/language ability, a general consensus among different schools of thoughts is clear: writing, especially academic writing, is more than producing a linguistically correct text; it is also a cognitive process, and a social interaction between and representation of the author and her audience in a particular communicative situation within in a particular historical and sociocultural context. Based on such an understanding, I now turn to the examination of operationalizations of the construct of (academic) writing ability, as represented by the writing tasks included on various large-scale high-stakes EFL tests in the Chinese context in particular. Juxtaposing the theoretical and the operational definitions of the construct helps reveal the extent to which current testing practices actually align with our theoretical knowledge about writing.

9.2 Operationalization of English Writing Ability on Chinese EFL Tests

Four nation-wide large-scale high-stakes EFL tests are in use for educational evaluation and selection purposes at various stages of postsecondary education in China. These four tests, namely National Matriculation English Test (NMET), College English Test (CET-4 & 6), Test of English for English Majors (TEM-4 & 8), and the national Graduate School Entrance English Examination (GSEEE), affect the entire student population in China. Given the high-stakes nature of these tests, the way they assess English language proficiency, and writing proficiency in particular, will certainly have a huge impact on how EFL writing is conceptualized, taught and learned. A quick survey and analysis of the writing components on these tests, in terms of their design, tasks, and scoring rubrics, would present an operational definition of English academic writing ability within the Chinese EFL setting.

9.2.1 National Matriculation English Test (NMET)

According to the official *Guide for NMET* (National Education Examinations Authority, 2019), the writing part of the test intends to measure students’ ability to (1) convey information in a clear and coherent manner, and (2) effectively use the language knowledge they have acquired. Only one writing task is presented on NMET, specifying the basic rhetorical situation and asking students to write a short text of approximately 100 words to convey specific information provided to them in the prompt in their native language. Prevalent genre types include emails, letters, memos, and announcements, although periodically picture descriptions and expositions can also be found. Four dimensions are included in the holistic scoring rubric, including the coverage of key points listed in the prompt, the diversity and accuracy of lexico-grammatical features, coherence and cohesion, and mechanics. As shown, the rubric ignores the effectiveness and appropriateness of communication despite that the task is framed as a rhetorically situated "authentic" task. Additionally, with the key information and ideas listed in bullet points and provided to the test takers in Chinese, the writing task is reduced to a translation task in essence (Dong et al., 2011), testing students' ability to use lexico-grammatical features and control basic mechanics. Cai (2002) further points out that such a task design pre-determines not only the content of the writing but also the organizational structure, as most test takers would follow the order of those listed bullet points in their writing. Based on such observations, therefore, this type of writing is also known, among some Chinese scholars, as a "quasi-writing" activity (e.g., Chen, 2017; Lu, 2010).

9.2.2 College English Test Band 4 (CET-4) and Band 6 (CET-6)

As outlined in the official *Guide for CET* (National College English Testing Committee, 2016), the writing part of CET-4 is designed to measure students' ability to describe and narrate personal experiences, feelings, emotions and events, to describe and explain simple tables, graphs or other graphics, to offer personal opinions on familiar topics, and to handle practical writing. CET-6 builds on CET-4 and requires students to express their opinions on common topics, describe, explain and discuss information presented in tables, graphs, and other graphics. The major dimensions explicitly stated in the *Guide*, defining the construct of writing ability, remain the same across the two levels of the test and include the presentation of ideas, text structure and organization, language use, and use of writing strategies. Unlike NMET, therefore, CET writing tests value author stance and opinions, in addition to organization and language use. Interestingly, they also highlight the proper use of writing strategies that would facilitate the conveyance of ideas and content, although no further explanation is given in either the rubrics or the *Guide* as to what this dimension means and how it would be evaluated.

Both tests require test takers to complete their responses within 30 min and write in response to a single prompt, which oftentimes calls for an expository or argumentative genre, with the length requirement being slightly different (120–180 words for CET-4, and 150–200 words for CET-6). As is the norm in large-scale testing practices, a holistic rubric is adopted for scoring written responses. The rubric, however, only covers the first three dimensions outlined in the official *Guide* for the test, leaving out the assessment of writing strategy use. Moreover, the descriptors on the rubric are often oversimplified and generic. As an example, the rubric defines the highest level of writing performance as one that is “on topic, with clear ideas, coherent organization, and correct language use” (National College English Testing Committee, 2016: 10). Probably because the rubric is generically constructed, it is applied to the scoring of both CET-4 and CET-6 writing samples. However, the test developers added a note in the *Guide*, stating that although the rubric is shared, CET-4 and CET-6 writing tests are “set at different difficulty levels and with different assessment requirements,” so that “the anchor papers of CET-4 and CET-6 that received the same-level ratings are in fact very much different in quality” (National College English Testing Committee, 2016: 10). With the use of the same scoring rubric and descriptors, it is hard to conceptualize and understand how a level 5 essay on CET-4 should be “very much different in quality” from a level 5 essay on CET-6. The single, prompt-based, often also decontextualized, writing tasks on CET tests also raise questions of its authenticity and interactiveness, which in turn threatens validity (e.g., Gu & Yang, 2009; Cai, 2002).

9.2.3 Test for English Majors Band 4 (TEM-4) and Band 8 (TEM-8)

According to Jin and Fan’s (2011) test review, TEM is an achievement test that intends to assess whether undergraduates majoring in English have achieved the required English proficiency at the end of their 4th semester and 8th semester during their undergraduate studies, hence TEM-4 and TEM-8, respectively. According to the official *Guide for TEM-4* (Pan, 2016), the writing section is designed to measure students’ basic competence in “written expression” through a performance task that requires students to write in response to a given prompt, graphic, or short reading excerpt. Students are expected to write approximately 200 words within 45 min in such genres as exposition, argumentation or narration. Written responses are evaluated in terms of content relevancy and adequacy, organization and coherence, and language accuracy and appropriateness. Similarly, the writing section on TEM-8 also measures writing ability through a performance task, although the official *Guide for TEM-8* states explicitly that TEM-8 only adopts an integrated reading and writing task (Deng, 2017). Students need to process *two* short reading excerpts and write approximately 300 words within 45 min on TEM-8; other than that, all the conditions and evaluative criteria stay the same as those for TEM-4. It should be noted that this integrated reading-to-write task type was only recently introduced onto TEM tests in 2016, after 25 years of impromptu opinion writing test.

9.2.4 Graduate School Entrance English Examination (GSEEE)

According to the official *Guide for GSEEE* (National Education Examinations Authority, 2018), the writing section comprises two tasks, one requiring students to complete a short 100-word practical writing in the genre of either letter, memo, abstract, or report (Task A), while the other requires students to write a conventional narrative, descriptive, expository, or argumentative essay of 160–200 words, based on a given prompt, picture, graph, or outline (Task B). The evaluative criteria cover the following four dimensions, as explicitly stated in the *Guide*: (1) correct use of grammar, spelling and punctuation, and appropriate use of vocabulary; (2) adherence to genre conventions; (3) appropriate organization that brings out clarity and coherence; (4) appropriate register in relation to the specified purpose and audience of the writing, if given. Written responses are rated holistically on a scale of 0–5. Despite the listed evaluative criteria in the *Guide*, however, the actual rubric seems to prioritize task completion (i.e., coverage of required content and points), lexico-grammatical accuracy, cohesive device use, proper register and format, as well as length of response as key evaluative criteria. Overall, therefore, the evaluation of writing ability, or that of text quality, still seems to focus on lexico-grammatical accuracy, due to either the neglect or the vague description of other dimensions.

9.3 Problem Statement and the Need for Alternative Approaches to Academic Writing Assessment

A brief review of the writing sections on these large-scale high-stakes national EFL tests points to the fact that they all assess writing based on a written product. In contrast, the writing theories developed in the past few decades have highlighted that "writing is text, is composing, and is social construction" (Cumming, 1998: 61), and that "effective writing integrates the product with the process within a specific context" (Hildyard, 1992: 1528). The EFL testing practices reviewed above, therefore, show a significant under-representation of the construct of writing ability.

Furthermore, an examination of the specific writing tasks reveals an overreliance on the use of decontextualized generic "essay" writing tasks. The endorsement of this impromptu opinion-based essay writing as more or less the only task type on these large-scale high-stakes EFL tests is particularly problematic. Such a task type fails to see writing as a social action and interaction situated in a particular rhetorical and sociocultural context, leading to not only construct underrepresentation but also a lack of authenticity in task design, which, from a testing perspective, could threaten the validity and usefulness of such an assessment approach (cf. Moore & Morton, 2005). Specifically, authentic academic writing tasks at the postsecondary level often involve in-depth and critical processing and use of sources, and evidence-based or data-driven argument construction and presentation, rather than a simple opinion statement or personal response to an everyday topic. Sadly, however, these types of topics accounted for over 64% of all the CET writing topics, based on Gu & Yang's (2009) study of CET writing tasks over a period of two decades (1989–2008).

In addition to such personal topics adopted by most of the EFL tests in China, the prompts would often list all the key information that test developers expect students to cover in their responses. Such a design reduces a writing task to either a translation task (in the case of NMET where such bullet points are listed in Chinese) or a task that does not involve much thinking or planning or organization of content. In fact, Gu and Yang's (2009) study showed that 97.5% of all the CET writing tasks would fall into this category that they termed "outline-provided" type of writing, which is also the most prevalent type on GSEEE. In such cases, writing is underrepresented as linguistic accuracy and rigid formality only (Chen, 2017), as reflected in the rubrics themselves.

To be fair, however, recent years have witnessed certain changes in the writing task design on some of these EFL tests. As mentioned earlier, the TEM test battery recently introduced the reading-to-write task. Students are now required to process reading materials, although still rather limited in length and complexity, before they are required to produce a written text. This is certainly better aligned with authentic academic writing tasks, at least for the English majors who are expected to complete their coursework and degree thesis in English (such considerations may also explain why out of the four large-scale nation-wide EFL tests, only TEM seems to have implemented such a new task design). While this reform represents a step forward in the EFL test developers' conception of writing ability, a scrutiny of the prompt itself and the scoring rubric for this new task type still reveals a surface-level application of source-based writing. For example, the new TEM writing tasks only ask students to first summarize the main points in the reading passage(s), and then express their opinions on a related topic. What we could infer from such prompts is that source materials are used only for some generic summary tasks, independent of the subsequent writing task. Although the prompt also includes a line saying that students "*can* support [themselves] with information from the excerpt(s)," the use of and interaction with source materials are not explicitly required, hence unlikely to be valued. Indeed, if we turn to the actual scoring rubric, it becomes clear that none of such aspects of writing ability as knowledge construction, social interaction, and representation of self and audience is being considered or assessed. As an example, the descriptors used to evaluate TEM-8 writing samples define the highest level of student responses as those that showcase "effective communication with accuracy," which is further defined as fully addressing the writing task (i.e., contain both summary and opinion) with "logical organizational structure, . . . clearly stated main ideas, and sufficient supporting details," and with "almost no errors of vocabulary, spelling, punctuation or syntax," while "[using] the language appropriately" (Deng, 2017: 30–33). Apparently, the adequate and critical use of source materials for academic interaction and communication is not included as part of the evaluative criteria. In fact, the descriptors are almost the same as those used in the rubrics for any conventional impromptu essay writing tasks. The only required use of source materials also stops at what Bereiter and Scardamalia (1987) would call the "knowledge telling" level, neglecting that writing, particularly academic writing in higher education, is often for the purpose of knowledge transformation and construction.

Based on the above analysis of the overall EFL writing assessment design, the specifics of the writing tasks and prompts, as well as the scoring rubrics, it is not difficult to note that indeed "writing theory has had a minimal influence on writing assessment" (Behizadeh & Engelhard, 2011: 189). The developments in our theoretical understandings of the construct of writing and language ability are inadequately reflected in writing assessment practices, particularly and especially in the EFL context. Furthermore, as Hamp-Lyons (2016c) pointed out, writing assessment for academic purposes in higher education (HE) in particular has significantly lagged behind our "knowledge in what the language(s) of higher education look and sound like and how they 'work' linguistically, socially, culturally and interculturally" (p. 17). It is obvious that alternative means and forms of academic writing assessment are needed to more faithfully reflect the authentic writing tasks people encounter in HE and to better capture both the breadth and depth of this construct of academic writing ability. Only with the use of such alternative assessments, especially on the large-scale high-stakes tests, will we be able to introduce a more positive washback and ultimately help EFL students to develop the much-needed writing competence to support and facilitate successful academic communications and knowledge making in HE. The next section will hence present an alternative approach to academic writing assessment design and illustrate how writing-theory-informed design of academic writing tasks may be used to better capture the breadth and depth of the construct of academic writing at the tertiary level.

9.4 A Theory-Based Approach to Academic Writing Assessment Design

In order to address the aforementioned issues of construct underrepresentation, lack of authenticity, as well as the minimal influence of writing theory on our current writing assessment practices, writing assessments should evaluate not only the written product (writing as text), but also the writing process (writing as cognitive activity) and the social construction and interaction as mediated by the text (writing as social act). Of course, some attempts have already been made in the field to cover the breadth of the construct. The earliest and most commonly referenced attempt is the development and use of analytic rubrics, or what Hamp-Lyons (2016a, b) would call multiple-trait rubrics, when scoring students' written products. By incorporating more dimensions and more detailed descriptors into the rubric, it is hoped that, in addition to the conventional trichotomy of content, organization, and language & mechanics, those often neglected components can also be evaluated, including for example, audience awareness, authorial voice, register and genre knowledge, pragmatic competence, communicative effect, citation and reference format, as well as paraphrasing, summarizing and synthesis skills for integrated academic writing tasks in particular (e.g., Banerjee et al., 2015; Chan et al., 2015; Knoch, 2009). While analytic or multiple-trait scoring certainly contributes to a more systematic evaluation of the various dimensions that together define writing ability, the use of such rubrics is, nevertheless, limited in that not all aspects of the writing ability may be explicitly manifested in the written product and readily translated into a dimension on a rubric. Most obvious of all is the aspect of cognitive process and metacognitive strategy use involved in the completion of a writing task. The end-product may not be able to provide enough evidence for raters to reliably evaluate students' competence in these areas. This probably also explains why the official CET *Guide* includes the use of writing strategies as one of the four key dimensions of writing ability, but leaves it completely unattended to in the actual scoring rubric.

Perhaps to address some of these unresolved issues, Beck et al. (2015) recently proposed that we should go "beyond the rubric" in our evaluation of students' writing, and to use "think-aloud as a diagnostic assessment tool" to help us gain insights into the composing process so as to identify students' "strengths and challenges as writers, beyond what is discernible from evaluating their writing alone" (p. 670). While think-alouds can tap into the implicit composing processes, hence adding that part of the construct back into our assessment of writing ability, the applicability of such an assessment approach is probably limited to classroom use only, due to practicality considerations. Hence, new means and forms of academic writing assessment are needed. One alternative, I believe, is to streamline the composing process to the extent possible, eliciting the cognitive activities and social interactions in particular in our task design. Based primarily on Flower and Hayes's (1981) cognitive process model of writing, Bachman and Palmer's (1996, 2010) conception of communicative language ability, and Snow and Uccelli's (2009) nested model of academic language ability, I will illustrate how these theories may guide us in our design and development of a cognitive-process-based academic writing assessment for use with students in higher education.

As Flower and Hayes's (1981) model highlights task environment as an important dimension in any act of writing, writing assessment should also seek to specify for the test takers the topic and communicative purpose of writing, the rhetorical context, as well as the intended audience. In terms of topic selection, large-scale language tests often spare no effort to make sure that test takers write on a familiar topic. This probably explains why most of the writing tasks on the Chinese EFL tests surveyed above are about some aspects of students' everyday life. While minimizing the influence of topical knowledge on language performance is desirable for the purpose of assessing test takers' language knowledge, it is not when the purpose of assessment is to measure one's writing ability, especially for academic purposes in higher education. After all, a major function of language use in higher education is precisely for learning. We use language to learn about and communicate new information and ideas, new knowledge and discoveries. Consequently, writing tasks would only be more authentic, and fair too, if students can write to learn about a relatively new topic. In fact, empirical studies indeed revealed that almost all the university writing tasks "involved a research component of some kind, requiring the use of either primary or secondary sources or a combination of the two," as opposed to the writing tasks on language tests that focus primarily on "[writing from] prior knowledge" (Moore & Morton, 2005: 52). Similarly, Deane and his colleagues also pointed out that "writing in a school context is almost always engaged with, and directed toward, texts that students read, whether to get information, consider multiple perspectives on an issue, or develop deeper understandings of subject matter" (Deane et al., 2008: 78). An integrated writing assessment design is therefore more representative of and congruent with actual writing practices in authentic educational settings. The key is to provide the right type of input that would offer and stimulate ideas, and at the same time be comprehensible to students at a particular level of language proficiency and cognitive maturity.

To be more authentic, such input could, and probably should, go beyond one or two short excerpt(s) to include multiple sources and materials. If technologically feasible, such input could be made accessible to students through hyperlinks that would lead to further processing of additional materials of different degrees of relevance. Students' ability to select relevant input for use in their writing may well be part of their academic writing ability, in that it would provide evidence into the information processing, critical thinking, and evaluation skills involved in actual academic writing. Meanwhile, the use of such materials and information also reflects another key dimension in Flower & Hayes's (1981) model, wherein the writer's long-term and short-term memory would interact and influence the composing process.

Once this task environment (i.e., rhetorical situation and topical knowledge) is specified in the prompt and input, the rest of the writing assessment can simulate the general process of composing and be organized into roughly three stages or sections, reflecting the three key components in Flower and Hayes's (1981) model: planning, translating, and reviewing (see Fig. 9.1). Of course, these cognitive processes are nonlinear and recursive; however, it does not mean that they cannot be represented somehow on a test using a combination of pre-writing items, a main writing task, and post-writing items, to tap into these different cognitive activities and metacognitive strategies employed and deployed in the composing process.

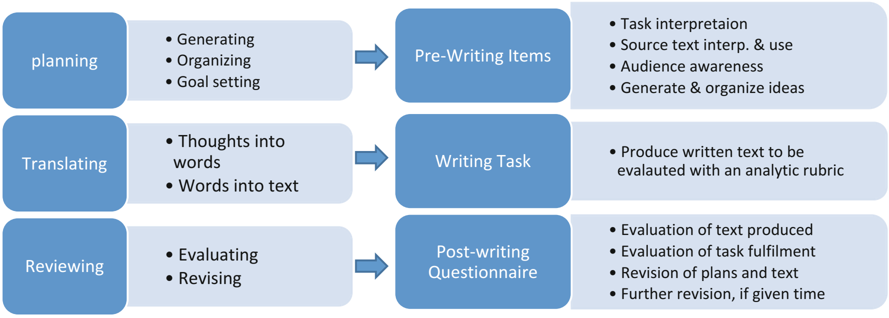

planning
• Generating
• Organizing
• Goal setting
Pre-Writing Items
• Task interpretaion
• Source text interp. & use
• Audience awareness
• Generate & organize ideas
Translating
• Thoughts into words
• Words into text
Writing Task
• Produce written text to be evalauted with an analytic rubric
Reviewing
• Evaluating
• Revising
Post-writing Questionnaire
• Evaluation of text produced
• Evaluation of task fulfilment
• Revision of plans and text
• Further revision, if given time

Fig. 9.1 Translating cognitive processes into test sections and items

As shown in Fig. 9.1, the planning stage in Flower and Hayes's (1981) model includes key components such as generating ideas, organizing ideas, and goal setting, which are also key elements in the strategic competence dimension of Bachman and Palmer's (1996, 2010) model of communicative language ability. In order to assess the ability of planning prior to the actual writing, pre-writing items may simulate the think-aloud process to elicit and examine the cognitive and metacognitive activities involved in the planning stage. Some items in the pre-writing section, for example, may ask students to articulate clearly their interpretations of the purpose, audience, and genre of the writing task in relation to the rhetorical situation given. Other items may ask students to list major ideas they think are relevant and important, and organize these ideas in an outline or bullet-point format. When the writing task involves and is based on students' processing of reading materials, pre-writing items could also measure students' understanding of the input and elicit their plans on *how* they intend to use such input for the purpose of the writing task.

Based on such planning, students can then translate these ideas and plans into an actual written text. Examining the written product eventually submitted in relation to their responses to the pre-writing items can reveal how original ideas and plans have been implemented, modified, or adjusted to varying degrees of success. This evidence may also be used to measure students' strategic competence, which is largely undealt with so far in existing writing assessment practices. Moreover, the written product could be scored using an analytic rubric designed to capture the multiple traits and dimensions of the construct of academic writing ability. In particular, the design of this analytic rubric should seek to restore the often-missing social dimension in L2 writing rubrics, highlighting the importance of the representation of the author and the audience in academic written interaction, as argued by Snow and Uccelli (2009) in their model of academic language use.

Conventionally, the writing section on most of the existing language tests would end here with the completion and submission of the final written product. Nevertheless, such a design does not faithfully reflect the recursive composing process. Successful writing almost always involves extensive revision, rewriting, and editing. Of course, writing assessment researchers and practitioners are not unaware of this mismatch. However, many believe it is simply impossible to address the recursive writing process in testing conditions. Hamp-Lyons and Kroll (1997), for example, pointed out that "other [non product-oriented] models that play a critical role in the field of composition studies may seem unhelpful, because they are not so much models of writing as a product as they are models of writing as a process," and noted specifically how the writing process model "is problematic for the design of academic writing assessment" (p. 7). Although no further explanation was given on why they believed the process model was "unhelpful" and "problematic," it was implied that tests and assessments can only be about products, despite the well-established process-oriented practices endorsed by writing teachers in various writing classrooms.

Unarguably, no test could fully emulate the authentic writing process due to practicality issues, particularly time constraints. However, this should not suggest that the assessment of writing ability could/should not go beyond that of the written product. The writing process, for example, could be captured, at least to a certain extent, by a pre-writing section that allows writers to demonstrate their planning and a post-writing section that prompts them to reflect on their writing processes as well as their plans for subsequent revision. In particular, this post-writing section could include items and tasks that ask students to (1) self-evaluate their writing and communicative success, and the overall task fulfillment, and (2) reflect on their own writing processes and strategies, including for example, how often, if at all, they evaluate their writing plans and products and revise their plans and texts while composing. Additional questions can be designed to probe into the subsequent revision plans by asking what types of revisions, if at all, they would focus on if they were given more time and resources. Such data, although self-reported, could still give us valuable information about the students' strategic competence, cognitive ability, and metacognitive strategy use that inform and influence their writing practices and performance.

9.5 Discussion and Conclusion

In response to the observation that current writing assessment designs and practices are inadequately informed by writing theories, an alternative design informed and guided by theoretical models of writing is proposed. As Cumming (1998) pointed out decades ago, "writing is text, is composing, and is social construction" (p. 61). Existing writing assessments, however, focus primarily on the assessment of written product, leaving out the composing process and the social construction. The design proposed here, therefore, expands on the current coverage of the construct by incorporating the cognitive processes (as informed by Flower and Hayes's cognitive process model) and strategic competence (following Bachman and Palmer's communicative language ability model) involved in composing and written interaction, and foregrounding the nature of writing as a social construction of meaning and relationships (as highlighted in Snow and Uccelli's nested model of academic language). Admittedly, in a large-scale testing context, not all aspects of the social functions of writing can be fully captured, particularly in terms of collaborative writing or using writing as a site for social and political actions (Cumming, 1998). In this chapter, therefore, the social aspect of the construct primarily focuses on the importance of situating meaning making in specific social cultural contexts and in relation to different communicative purposes and audiences. It also highlights the function of writing as a site for the author to build relations with readers, with prior texts, and gain voice and identity within a particular sociocultural context (cf. Bazerman, 2015; Beach et al., 2015; Snow & Uccelli, 2009).

Specifically, the pre-writing and post-writing items aim to make explicit the implicit cognitive and metacognitive activities involved in the composing process. This itemized design is also more practical than Beck et al.'s (2015) use of think-alouds, making it applicable to various testing and assessment contexts. In addition, a streamlined process-oriented design could also serve to raise students' awareness about the kind of thinking, planning, monitoring, and revising that are necessary for successful writing, making test taking a learning process in and of itself. Furthermore, highlighting the nature of writing as context-specific social construction of meaning and relationship in the design of the writing tasks and rubrics also help raise L2 writers' awareness about the "dialogic, [goal-oriented], and audience-directed quality of powerful writing, and … hone [their] understanding of how academic language choices are shaped by social contexts" (Beck et al., 2015: 680).

In addition to the positive impact on test takers and their test taking experience, such an alternative assessment design could also benefit other stakeholders, especially users of the assessment results and decision makers. Test takers' responses to the pre- and post-writing items would provide additional information about their writing performance and ability, adding discriminative power to the writing test as a whole. It is likely, for example, to have multiple, or sometimes even a large number of, test takers receiving the exact same score/rating on the essays they produce in response to a conventional writing prompt. It would be impossible to interpret, based on these essay scores alone, how one test taker may still differ from another. Data collected from pre- and post-writing items, however, could reveal varying levels of composing competence and strategy use, contributing to more nuanced and accurate interpretation of their writing ability. Such information could serve as the basis for important decision making by test users, including for example, placement decisions into various writing courses and curricula that target different instructional approaches and foci. Of course, when used by classroom teachers for diagnostic purposes, such information could greatly enhance pedagogical effectiveness and support differential treatment of individual writers' needs and challenges.

While such an alternative design creates opportunities for writing assessments to better represent the construct and bring about positive washback effect, it also poses a few challenges on the actual test development and administration. One such challenge is that it requires the test developers and item writers to have a solid understanding of relevant writing theories that could properly inform their task design and item writing. Without such theoretical knowledge, it is likely that the design of the items and tasks may misguide the test takers and distort the (meta)-cognitive processes, hence negatively influence students' writing performance. To address this issue, therefore, it is important that professional development and training be offered to test developers and item writers prior to the actual test development.

Another major challenge concerns the complexity of scoring. Such a contextualized, process-oriented design defies the use of any generic existing writing rubrics. Instead, it calls for the use of a combination of various scoring approaches and tailor-made rubrics to evaluate responses from different sections and items. In general, holistic primary-trait scoring may be used for specific pre- or post-writing items that tap into various context-specific interpretative or (meta)cognitive skills, whereas multiple-trait and analytic scoring can be used to evaluate the main written product composed in response to the task-specific writing prompt. The design of these scales or rubrics will also need to be context- and task-based, although they may still incorporate categories we often find on existing generic writing rubrics.

In addition to proper choices of scoring approaches, score reporting could be yet another challenge. Should we report scores based on sections, reflecting test takers' ability to control the writing process? Or should we report scores based on skill areas, such as test takers' audience awareness, which could be reflected in their interpretation of the task, their writing plans, the actual written product, as well as in their plans for subsequent revisions? The decision, of course, will have to be made based on the purpose and focus of the assessment, together with considerations of the different stakeholders' needs and intended uses of such score reports.

One more decision to be made and justified is whether or not to penalize students' less-than-optimal planning in the pre-writing section, knowing that initial plans are likely to change during the recursive writing process. Likewise, precise interpretations of any observed discrepancy between a pre-writing plan and an actual written product could be a real challenge, as it would be difficult to tell whether the discrepancy is a result of the writer's conscious modification of initial plans during the writing process, or a reflection of his/her inability to execute the plans in the actual act of composing. A potential solution to such a problem is to design the post-writing items in a way that would elicit students' explicit reflections on the choices they made prior to and during the writing. This would allow us to gather information similar to that obtained from a think-aloud session, despite the retrospective route.

All the aforementioned challenges, however, do not outweigh the value-added benefits derived from the use of such an alternative assessment design, especially in EFL contexts that have long had a skewed representation of the writing construct both on their tests and in various writing classrooms and programs. Hopefully, with a new mindset that goes beyond the conventional product-oriented testing practice, together with the technological affordances available to us in this new era, we are able to design new assessments that more accurately reflect our current understanding of the construct under examination, instead of prioritizing only the measurement or psychometric issues. Only in this way will we be able to materialize the next generation of writing assessment, one that reflects an understanding of writing assessment as "both humanistic and technological" and "a complex of processes in which multiple authors and readers are involved and revealed" (Hamp-Lyons, 2001: 117).

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Banerjee, J., Yan, X., Chapman, M., & Elliot, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19.

Bawarshi, A. S., & Reiff, M. J. (2010). *Genre: An introduction to history, theory, research, and pedagogy*. Parlor Press.

Bazerman, C. (2015). What do social cultural studies of writing tell us about learning to write? In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 11–22). Guilford Press.

Beach, R., Newell, G. E., & VanDerHeide, J. (2015). A sociocultural perspective on writing development: Toward an agenda for classroom research on students' use of social practices. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 11–22). Guilford Press.

Beck, S. W., Llosa, L., Black, K., & Trzeszkowski-Giese, A. (2015). Beyond the rubric: Think-alouds as a diagnostic assessment tool for high school writing Teachers. *Journal of Adolescent & Adult Literacy, 58*(8), 670–681.

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189–211.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Erlbaum.

Bloome, D., Carvalho, G. T., & Rue, S. (2018). Researching academic literacies. In A. Phakiti, P. D. Costa, P. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (1st ed., pp. 887–902). Palgrave Macmillan.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities (11–18)*. Macmillan.

Cai, J. (2002). The impact of CET-4 and CET-6 writing requirements and rubrics on Chinese students' writing. *Journal of PLA University of Foreign Languages, 25*(5), 49–53.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing, 26*, 20–37.

Chen, W. (2017). On NMET writing tasks and their future development. *China Examinations, 302*, 44–47.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8*(3), 165–191.

Crusan, D. (2014). Assessing writing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–15). Wiley Blackwell. https://doi.org/10.1002/9781118411360.wbcla067

Cumming, A. (1998). Theoretical perspectives on writing. *Annual review of applied linguistics, 18*, 61–78.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (ETS RR-08-55). Retrieved from https://doi.org/10.1002/j.2333-8504.2008.tb02141.x

Deng, J. (Ed.). (2017). *Guide to test for English Majors, Band Eight*. Shanghai Foreign Language Education Press.

Dong, M., Gao, X., & Yang, Z. (2011). A longitudinal study of writing tasks on the NMET national test papers (1989-2011). *Educational Measurement and Evaluation, 10*, 47–52.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Gu, X., & Yang, Z. (2009). A study of the CET writing test items over the past two decades. *Foreign Languages and Their Teaching, 243*, 21–26.

Hamp-Lyons, L. (2001). Fourth generation writing. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 117–129). Lawrence Erlbaum Associates, Inc.

Hamp-Lyons, L. (2016a). Farewell to holistic scoring? *Assessing Writing, 27*, A1–A2.

Hamp-Lyons, L. (2016b). Farewell to holistic scoring. Part Two: Why build a house with only one brick? *Assessing Writing, 29*, A1–A5.

Hamp-Lyons, L. (2016c). *Unanswered questions for assessing writing in the HE. What should be assessed and how?* Paper presented at the UKALTA Language Testing Forum (LTF 2016).

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community, and assessment* (TOEFL monograph series report no. 5). Educational Testing Service.

Hatfield, W. W. (1935). *An experience curriculum in English: A report of a commission of the National Council of Teachers of English*. D. Appleton-Century Company, Incorporated. Retrieved from https://archive.org/details/experiencecurric00nati (Original work published 1935).

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Lawrence Erlbaum Associates, Inc.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive process in writing: An interdisciplinary approach* (pp. 3–30). Lawrence Erlbaum Associates.

Hildyard, A. (1992). Written composition. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (pp. 1528–1540). Macmillan Publishing Company.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445–462.

Hyland, K. (2003). Genre-based pedagogies: A social response to process. *Journal of Second Language Writing, 12*(1), 17–29.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly, 30*(4), 693–722.

Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing, 28*(4), 589–596.

Johns, A. M. (1991). English for specific purposes: Its history and contributions. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 67–77). Heinle & Heinel.

Johns, A. M. (2008). Genre awareness for the novice academic student: An ongoing quest. *Language Teaching, 41*(2), 237–252.

Johns, A. M. (2011). The future of genre in L2 writing: Fundamental, but contested, instructional decisions. *Journal of Second Language Writing, 20*(1), 56–68.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Lee, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157–172.

Lee, M. R., & Street, B. V. (2006). The "academic literacies" model: Theory and applications. *Theory into Practice, 45*(4), 368–377.

Lu, Z. (2010). Possible path of reforming English tests for College Entrance Examination based on application ability. *Educational Measurement and Evaluation, 1*, 15–26.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*, 43–66.

National College English Testing Committee. (2016). *Guide to College English Test (2016 revised version)*. Retrieved from http://cet.neea.edu.cn/res/Home/1704/55b02330ac17274664f06d9d3db8249d.pdf

National Education Examinations Authority. (2018). *Guide to Unified National Graduate Entrance Examination-English I & II (for non-English major)*. Higher Education Press.

National Education Examinations Authority. (2019). *Guide to National Matriculation English Test*. Retrieved from http://www.neea.edu.cn/res/Home/1901/d15ec0514666ac280810099f9595b557.pdf

Pan, M. (Ed.). (2016). *Guide to test for English Majors, Band Four*. Shanghai Foreign Language Education Press.

Rohman, D. G. (1965). Pre-writing: The stage of discovery in the writing process. *College Composition and Communication, 16*(2), 106–112.

Scardamalia, M., & Bereiter, C. (1991). Literate expertise. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 172–194). Cambridge University Press.

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). Cambridge University Press.

Spack, R. (1988). Initiating ESL students into the academic discourse community: How far should we go? *TESOL Quarterly, 22*(1), 29–51.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Cecilia Guanfang Zhao obtained her Ph.D. in TESOL from New York University and is currently Associate Professor in the Department of English at University of Macau. Her research interests are in the areas of second language writing, language assessment, and writing assessment in particular.

Chapter 10
Validating a Rating Scale for a University-Based Writing Assessment: The RUC-TWPE Experience

Li Liu and Guodong Jia

Abstract The validity of rating scales of writing performance is essential for ensuring reliable test scores and valid score inferences. This study reports on the validation of a rating scale for a university-based writing assessment, *Test for Writing Proficiency in English*, in the Renmin University of China, by using the argument-based validation approach. Both quantitative and qualitative methods were employed to guide the collection and analysis of evidence informing the validation. Documents of developing the rating scale were reviewed, and Many-Facet Rasch Measurement and rater interviews were conducted to investigate the performance of the rating scale and raters. The findings provided preliminary evidence for the evaluation and explanation inferences of the validation framework of the rating scale. Challenges and problems involved in designing the rating scale and developing the school-based writing assessment are also discussed.

Keywords University-based writing assessment · Rating scale · Argument-based validation

10.1 Background

The rating scale under examination is the assessment tool for the writing test, Test for Writing Proficiency in English (RUC-TWPE), designed and implemented by the School of Foreign Languages, Renmin University of China. The purpose of the TWPE is to examine whether college students have attained the writing proficiency standards articulated in the Renmin University of China—Standards of Writing Proficiency in English (SWPE) (SWPE Project Group, 2016). In addition, the TWPE could also be used as an indicator of the quality of the teaching of college English writing and whether the students have reached the writing proficiency

L. Liu (✉) · G. Jia
Renmin University of China, Beijing, China
e-mail: liliu@ruc.edu.cn; gdjia@ruc.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_10

Table 10.1 Test format of TWPE

Task	Form	Weight	Task type	Word count	Time allocation
Task A	IBT	40%	Practical writing	120–150	20 min
Task B	IBT	60%	Argumentation	200–300	40 min

standards set in the national standard *College English Teaching Guidelines* (Ministry of Education, 2016) and the university language teaching curriculum.

With a particular emphasis placed on college English education, China's Ministry of Education initiated the development of the *College English Teaching Guidelines* and implemented it in 2016. The *Guidelines* specify a redefinition of English proficiency (including writing proficiency), a specification of fundamental objectives, as well as intermediate and development goals to accommodate the growing demand for international exchange. The SWPE is designed with reference to the three levels of writing proficiency: elementary, intermediate, and advanced, as defined in the *Guidelines* and to the university requirements on students' English writing skills.

The implementation of SWPE requires a scientific assessment system. For the elementary and intermediate level of standards, TWPE has been designed to test students' writing proficiency at the completion of the two compulsory writing courses. For the elective Academic English Writing course, formative assessment and course papers are used to evaluate students' academic writing proficiency. TWPE is a computer-based test, including "Writing Task A" and "Writing Task B" with a weight of 40% and 60%, respectively. The test format can be found in Table 10.1.

For Writing Task A, test-takers will complete one of the required types of writing on a given topic. The minimum number of words is 120, and the time allotted is 20 minutes. It comprises 40% of the total score. This task aims to test practical writing, that is, the writing students need in their daily lives, whether emailing their teachers or writing formal letters to the university. For Writing Task B, test takers may be asked to provide solutions to some questions, arguments or evidence for a point of view, contrast or comparison of different views, or comments or counter-arguments on a certain viewpoint. The minimum number of words is 180 and the time allotted is 40 minutes. It comprises 60% of the total score. This task aims to examine whether students can argue, provide clear evidence, exemplify, and draw a logical conclusion on a given argumentation.

10.2 Literature Review

There is no single definition of writing ability applicable for all situations because it could be approached from different perspectives for different contexts, cognitively, socially, and culturally (Hamp-Lyons & Kroll, 1997; Weigle, 2002). An agreed construct of writing practice will allow for research results that can achieve a greater degree of comparability, more opportunities for convergent research findings, and a set of common descriptors. Since writing ability is a type of performance (McNamara, 1996), many writing specialists have developed scoring rubrics identifying features that should be attended to. Though not fully developed, the scoring rubrics designed to evaluate L2 writing present an "implicit theory about the nature of writing. . .about the development of L2 writing skills" (Valdes et al., 1992: 334–335).

10.2.1 Scoring Rubrics as the Model of L2 Writing Ability

The criteria for scoring reflect different conceptualizations of writing proficiency. One of the best-known and most widely used scales in ESL was created by Jacobs et al. (1981). The writing criteria outlined evaluate the writing performance of international university students in the foreign language classroom. Scripts are rated on five aspects of writing: content, organization, vocabulary, language use, and mechanics with each possessing a different weight. Weir (1988) developed a slightly different approach for the Test in English for Educational Purposes (TEEP). The scheme consists of seven scales with the first four relating to communicative effectiveness, while the others relating to accuracy. The Michigan Writing Assessment Scoring Guide is another example. It is scored in three rating scales: ideas and arguments, rhetorical features and language control for grading an entry-level university writing examination (Hamp-Lyons, 1990).

More recently, Cumming and his colleagues (2000) proposed a detailed rubric for defining L2 writing ability. The core conceptualization of their proposed models is that writing is a communicative act and that communicative language ability must include contextualized language use (Bachman & Palmer, 1996; Grabe & Kaplan, 1996; Hamp-Lyons & Kroll, 1997). Cumming et al. (2000) pointed out two orientations for examining the writing. The first is the text-characteristics perspective which focuses on the characteristics of the written texts that people produce. The second is the reader-writer approach which considers the perceptions and judgments of readers of such texts. Following this line of thought, they provided a potential rubric of evaluative criteria useful for defining L2 writing ability.

There are similarities between this model for L2 writing ability and Bachman and Palmer's (1996) model for communicative language ability. In fact, where Bachman and Palmer added strategic competence to their model, some L2 writing researchers have suggested that a separable skill of writing proficiency might complement L2 ability during L2 writing performance (e.g., Cumming, 1989; Krapels, 1990). In short, a model of L2 writing ability must indicate how L2 writing ability is distinct from other types of L2 knowledge and how L2 proficiency and writing ability interact.

Grabe and Kaplan's (1996) model addresses this issue in that their model gives much greater consideration to the linguistic knowledge base and particularly to an account of communicative competence as applied to writing. According to them, writing competence comprises three competences: linguistic (grammatical), discourse, and sociolinguistic (Grabe & Kaplan, 1996: 217–222). Grabe and Kaplan's (1996) model is important for defining L2 writing, as it is founded on previous research in both writing and L2 studies. For a construct definition to be useful for testing purposes, however, especially in large-scale writing assessment, it must be usable by test developers.

10.2.2 Development and Validation of Rating Scales for Writing Assessment

The process of constructing assessment scales for performance testing is complex and multi-dimensional. Well-designed rating scales are essential and can help mitigate rater effects. As a result, a number of different approaches, both empirically and intuitively based, are employed by test developers (Bacha, 2001; Knoch, 2009). Rating scales constructed based on expert intuition or certain theoretical models have been under criticism for their potential problems. Current thinking argues for empirical validation of assessment scales (Council of Europe, 2001; Upshur & Turner, 1995). That is, assessment scales should involve a range of methods for establishing their validity. Therefore, there is a shift to data-driven approaches for scale construction and validation, which "place primary value on observations of language performance and attempt to describe performance in detail to generate descriptors" (Fulcher et al., 2011).

Since writing assessment is a complex and multifaceted activity that could not be represented by a simple numerical score, researchers have called for an in-depth investigation into this "black box" in their quantitative studies (Eckes, 2005; Weigle, 1998). Another line of qualitative inquiry has therefore been devoted to using learner performance data and/or rater perceptions as the basis for the construction and validation of assessment scales.

However, it is crucial not to lose sight of the value of expert judgment (such as researchers and experienced teachers) and the wealth of knowledge that experts can bring to the development process. It is then advisable to take advantage of the strengths of a range of intuitive, quantitative and qualitative approaches for rating scale development.

The argument-based approach provides a practical and systematic guideline for constructing validity arguments, linking validity evidence for the development and use of a test in a particular context (Chapelle et al., 2010a). Researchers can determine the claims and evidence depending on their testing contexts and test uses (Chapelle et al., 2010b). Drawing on Kane's conceptualization of inferences, warrants and assumptions, Knoch and Chapelle (2018) proposed a set of warrants and assumptions related to the rating process. They claimed that evidence from the rating process encompasses a wide range of inferences (evaluation, generalization, explanation, extrapolation, decision, and consequences) in the interpretative argument. Their framework provides a useful starting point for rating scale validation and we will draw on this framework to situate the current work.

10.3 Methods

10.3.1 Research Questions

The current study reports the validation activities conducted for the TWPE rating scale. The validation was based on an argument-based approach proposed by Knoch and Chapelle (2018). They proposed a series of warrants and assumptions in relation to a number of inferences with possible sources of backing. Following this line of thought, the evidence in relation to assumptions underlying the evaluation and explanation inferences was gathered (see Table 10.2). The present study, therefore, aims to address the following research questions:

RQ1: Is the scale based on a theoretical/pedagogical defensible model of English as a second language writing?
RQ2: Does the rating scale distinguish the writing proficiency of students at different levels?
RQ3: How do raters use the rating scale?

Table 10.2 The validation framework adopted in the study

Warrants	Assumptions	Sources for backing
Evaluation inference: Observations are evaluated using procedures that provide observed scores with intended characteristics.		
A. The scale properties are as intended by the developers	1. Scale steps are adequate to distinguish among the levels that appear in the test; 2. The scale is able to spread test-takers into different levels as needed for the test purpose	Many-facet Rasch analysis
B. Raters rate reliably at the task level	1. Raters are able to identify differences in performances across score levels 2. Raters can consistently apply the scale to test tasks; 3. Raters are comfortable when applying descriptors and confident in their decisions	Many-facet Rasch analysis; Rater interview
Explanation inference: Expected scores are attributed to a construct of language proficiency.		
A. The rating criteria are based on a clearly defined construct.	1. The rating scale is based on a defensible theoretical or pedagogical model of proficiency and/or development. 2. Rating scale criteria and descriptors cover. the construct (i.e., no construct-irrelevance or under-representation).	Review of test development documentation

10.3.2 The TWPE Rating Scale

The scale is a six-point holistic scale with explicit descriptors for each level (see Appendix I for sample descriptors at band 4–6). A holistic scale is chosen due to the large test population and limited time range for reporting the test results to students and the university academic affairs office. A satisfactory level of reliability is expected to be achieved after systematic rater training.

10.3.3 Writing Tasks

The test takers are 60 undergraduate students enrolled in college English courses. Each test taker completed two writing tasks designed in correspondence to the curriculum of the writing courses. A total of 120 written samples were collected from one live TWPE trial test administration. The writing courses aim to equip students with a general understanding of effective essay components and writing skills. Upon completing the courses, learners are expected to write clear, detailed descriptions, write narrative essays of real events and experiences, write logical argumentation and write for practical purposes, such as a letter, a report and a personal statement (SWPE Project Group, 2016).

The writing test is 60 long and consists of two writing tasks of 150 words and 250 words. In Task A, candidates are required to respond to a situation by writing a letter, for example, requesting information or explaining a situation in the university context. In Task B, candidates write an essay in response to a point of view, argument or problem. Specifically, in this trial test, in the first task, test takers wrote a letter to accommodation authorities of the university to apply for a change of dormitory. In the written argumentation task, candidates composed an argument relating to study abroad (see Appendix II for the two tasks). Since the test is computer-delivered, the prompts are expected to be concise and clear on one full screen, the format of writing tasks of international standardized tests (such as IELTS) was also referred to.

10.3.4 Teacher-Raters

Six raters were selected to participate in the rating session. They are all experienced college English teachers and teachers of writing courses. Each essay was assessed independently by four raters using the TWPE rating scale. Rater R1 and rater R2 were asked to make judgments on the first writing task while raters R5 and R6 were working on the second writing task. Rater R3 and Rater R4 had experiences in using

Table 10.3 Raters and rating in the study

Rater	Task A	Task B
R1	√	
R2	√	
R3	√	√
R4	√	√
R5		√
R6		√

rating scales in large-scale writing assessments and they marked all the scripts (see Table 10.3 for a summary of rating arrangement).

10.4 Data Analysis

For the first research question, the procedures of developing the rating scale were reviewed to see whether the rating scale captures the writing construct covered in the curriculum and set in the requirements of the SWPE.

For the second research question, specific information regarding the adequacy of the rating scale to distinguish writing performance was sought via Many-Facet Rasch Measurement (MFRM). It can enable researchers to examine individual raters' rating performance with greater detail and specificity. For one thing, MFRM is able to differentiate systematic and random sources of rater variability through calibrating raters in terms of the systematic differences in their overall severity (indicated by rater severity measures) and estimating their degree of inconsistency across the whole ratings (indicated by rater Infit Mean Square index). Furthermore, it can look into the apparent inconsistencies to reveal sub-patterns of raters' scoring behaviors through estimating significant bias interaction between raters and students, items or tasks. The MFRM analysis for the scores collected in the separate rating sessions was conducted using FACETS version 3.71.1 (Linacre, 2010), a Rasch measurement computer program. The program is "ideally suited for essay grading, portfolio assessment and other kinds of judged performance" (Linacre, 2004: 4). In the current study, since the raters rated the student scripts written on two tasks using one holistic rating scale, a three-facet model was therefore built for the MFRM analysis that included the facets of student, rater and task.

To address the third research question, semi-structured interviews were conducted with raters after they finished all the ratings. The raters were first asked about their general impression about the rating scale, then about the clarity and sufficiency of the descriptors in each band level of the rating scale. They were also asked about their difficulties of using the rating scale if there were any. Interview transcripts were summarized according to the main themes to identify different aspects of using the rating scale in the study.

10.5 Results and Discussion

10.5.1 The Theoretical/Pedagogical Relevance of the Rating Scale

The Standards of Writing Proficiency in English (SWPE) was put into use in 2016. The overall objective of the SWPE is to delineate the standards of English writing proficiency that undergraduates should attain and enhance students' writing skills as a preparation for their future study, work and international engagement. In addition, the SWPE aims to offer guidance for curriculum design, material development, classroom instruction and assessment of college English writing courses to ensure the realization of the overall objectives of English writing instruction. The rating scale was then developed for the writing assessment to examine students' level of writing proficiency at the completion of the two writing courses, English Writing I and English Writing II.

First, surveys in the form of questionnaires and interviews of students and teachers suggest that students have no access to systematic learning in English writing in secondary school and are therefore at the preliminary stage in their English writing. To lay a foundation of English writing for students, English Writing I was designed, in which students are required to learn commonly used types of writing (description, narration, practical writing, etc.), basic rhetoric, and mechanics. With the completion of the English Writing I course, students will continue their learning by taking English Writing II, which is designed to help students achieve a higher level of writing proficiency. Students are expected to have a complete mastery of expository and argumentative writing as a preparation for future international communication. Third, the SWPE has an optional Academic English Writing course designed to develop students' basic skills and strategies of academic English writing. All the courses are delivered in English. Students will take the writing assessment after completing the first two writing courses.

In the development of the rating scale, the consensus was that rating scales should be designed based on the purpose of the test and should be in alignment with the writing skills covered in the curriculum. Several theoretical models of language proficiency were drawn on, including Canale and Swain's Communicative Competence Model (1980) and Bachman and Palmer's (1996) conceptualization of language proficiency. Four dimensions of writing construct were considered in the development of the rating scale: linguistic, sociolinguistic, pragmatic, and strategic competence. The linguistic, sociolinguistic, and pragmatic competences are concerned with knowledge that students need to use correctly and appropriately in specific contexts to achieve successful written communication. The strategic competence assesses the evidence of using different writing strategies.

Besides, as language is a carrier of culture, written expression should no doubt be concerned about the cultural differences between languages and respect conventions in different cultures. Cultural awareness will facilitate the sense of intercultural exchange and improve competence of cross-cultural communication. Cultural awareness was therefore added to the rating scale. The writing proficiency in SWPE thus refers to the competence of written communication, including linguistic competence, sociolinguistic competence, pragmatic competence, strategic competence and cultural awareness (SWPE Research Group, 2016).

Similar considerations were given when designing the writing tasks in the assessment. The genre of the first writing task includes narration, description and practical writing, such as letters and emails. The second writing task examines argumentative writing skills. Test takers write an argument on a point of view, a question or a given topic. With different writing prompts, test takers might be asked to provide solutions to a question, arguments or evidence for a point of view, contrast or comparison of different views, or comments or counterarguments on a certain viewpoint. The topic is more abstract than task one and will include some intercultural communication elements. For example, the topic used in the trial test required students to weigh the pros and cons of studying abroad.

10.5.2 Performance of Raters and Rating Scale

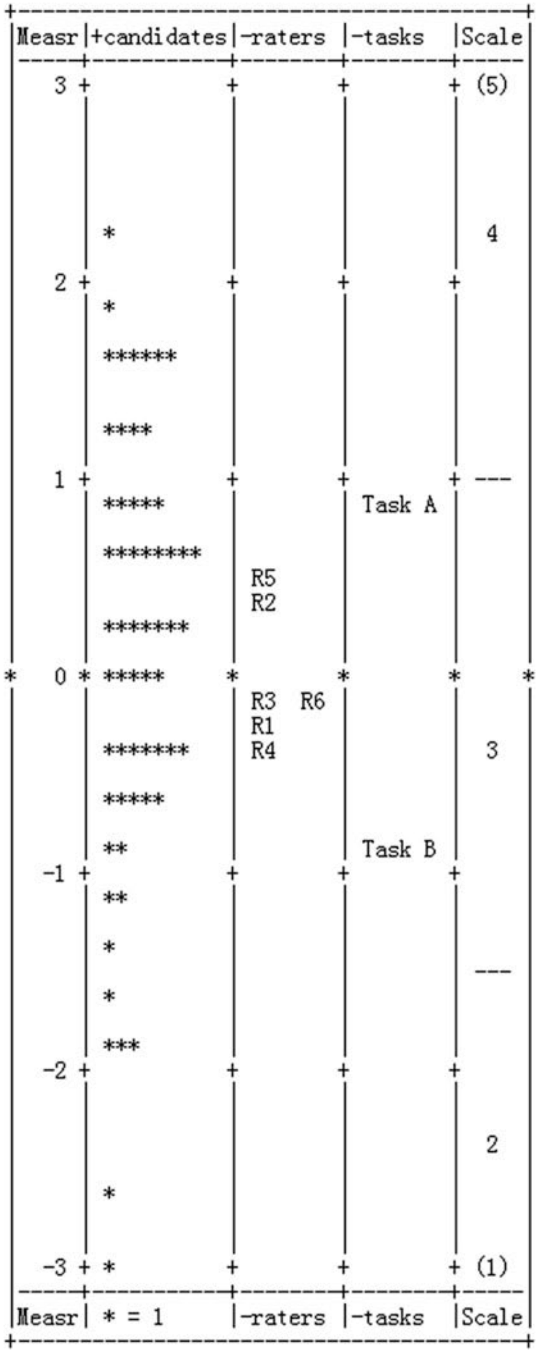
10.5.2.1 MFRM Overview

MFRM analysis can provide rich statistical outputs, including both global level statistics for each facet and individual level indices for each element in the facet. A brief summary of all the facets has been interpreted in this section. Figure 10.1 shows the facet map, displaying candidate ability, rater severity, task difficulty and the score used.

The leftmost column of the map represents the common scale in the unit of logit, against which all the measures in the following facets are calibrated. The second column displays the distribution of estimates for students' writing proficiency with those who are more competent listed at the top while the less competent at the bottom. The measures of student ability ranged from −2.95 to 2.25 (see Table 10.4). The fixed (all same) chi-square test was statistically significant ($\chi^{(2)} = 261.5$, $p = .00$). The value of candidate separation (strata) was 2.70, suggesting that the test was able to distinguish at least two distinct levels of writing proficiency among the candidates. High separation (>2) indicates that the test was able to differentiate between difficulty/ability groups of items/persons (Linacre, 2019). The person reliability was 0.76, indicating that the classification of the students is generally reliable.

The mean infit MnSq (0.97) is close to the expected value of 1.0. There are no fixed rules for setting the limits for the fit statistics (Aryadoust et al., 2020). In general, any individual Infit Mean Square value needs to be interpreted against the mean and standard deviation of the set of Infit Mean Square values for the facet concerned (Pollitt & Hutchinson, 1987). Under this consideration, the fit statistics can be

Fig. 10.1 Overall facet map



```
+-----------------------------------------+
|Measr|+candidates|-raters |-tasks  |Scale|
|-----+-----------+--------+--------+-----|
|   3 +           +        +        + (5) |
|     |           |        |        |     |
|     |           |        |        |     |
|     |           |        |        |     |
|     |           |        |        |     |
|     |           |        |        |     |
|     | *         |        |        |  4  |
|     |           |        |        |     |
|   2 +           +        +        +     |
|     | *         |        |        |     |
|     |           |        |        |     |
|     | ******    |        |        |     |
|     |           |        |        |     |
|     |           |        |        |     |
|     | ****      |        |        |     |
|     |           |        |        |     |
|   1 +           +        +        + --- |
|     | *****     |        | Task A |     |
|     |           |        |        |     |
|     | ********  |        |        |     |
|     |           | R5     |        |     |
|     |           | R2     |        |     |
|     | *******   |        |        |     |
|     |           |        |        |     |
*   0 * *****     *        *        *     *
|     |           | R3  R6 |        |     |
|     |           | R1     |        |     |
|     | *******   | R4     |        |  3  |
|     |           |        |        |     |
|     | *****     |        |        |     |
|     |           |        |        |     |
|     | **        |        | Task B |     |
|  -1 +           +        +        +     |
|     | **        |        |        |     |
|     |           |        |        |     |
|     | *         |        |        |     |
|     |           |        |        | --- |
|     | *         |        |        |     |
|     |           |        |        |     |
|     | ***       |        |        |     |
|  -2 +           +        +        +     |
|     |           |        |        |     |
|     |           |        |        |     |
|     |           |        |        |  2  |
|     |           |        |        |     |
|     | *         |        |        |     |
|     |           |        |        |     |
|     |           |        |        |     |
|  -3 + *         +        +        + (1) |
|-----+-----------+--------+--------+-----|
|Measr| * = 1     |-raters |-tasks  |Scale|
+-----------------------------------------+
```

Table 10.4 Summary of candidate measurement statistics

	Max.	Min.	M	SD
Measure	2.25	−2.95	0.08	1.13
Infit MnSq	3.85	0.16	0.97	0.79

Adj (True) S.D. .98 Separation 1.77 Strata 2.70 Reliability .76
Fixed (all same) chi-squared: 261.5 d.f.: 59 significance (probability): .00

determined within the range of two standard deviations around the mean. A value greater than the mean plus twice the standard deviation would indicate too much unpredictability, or misfit. There are thus three misfitting candidates (5%) in this case. One test taker received a much more generous score from one rater; two candidates had very different scores for Task A than for Task B.

The third column of the facet map provides information about the rater facet. It compares the raters in terms of the level of severity/leniency they exercised during rating. The severest rater was rater R5 while rater R4 was the most lenient in scoring. The fourth column presents tasks in terms of their relative difficulty and two tasks were not equivalent in difficulty.

The last column of the facet map is the scale used in the rating process. The horizontal lines among the points stand for the threshold where the likelihood of getting the next higher score exceeds that of getting the next lower score for a given student script under a given assessment context. The facet map also indicated that the student facet has a much wider span on the logit than the rater facet, suggesting that the most significant part in score variance lies in the student writing proficiency rather than the raters.

10.5.2.2 Raters and Rating

One of the most important considerations for rating scale validation is whether raters are able to make effective use of the scale when scoring the test takers' scripts. According to the report, the raters showed different levels of severity and these differences were statistically significant ($p = .00$). As can be seen from Table 10.5, the raters are ranked in order of severity with rater 4 as the most lenient and rater 5 as the most severe. We can see that the range of severity measures (0.83 logits) is far smaller than the range of ability measures (5.20 logits). O'Sullivan (2002) suggests that when the range of severity of raters is much smaller than the range of ability measures, we can interpret this as an indication that differences in rater severity do not have much practical impact on scores.

For the rater facet, the Infit statistics denote the degree of self-consistency when using the rating scale. Having an expected value close to zero demonstrated that these raters had a somehow consistent rating and used the rating scale in a consistent way. The acceptable range of Infit Mean Square was calculated to be 0.52–1.4. As can be seen in the table, the infit statistics of all raters fell within the acceptable range. That is, all the raters were internally consistent and their ratings were

Table 10.5 Rater measurement report

Rater	Measure	Model S.E.	Infit		Outfit	
			MnSq	ZStd	Mnsq	ZStd
5	0.49	0.21	0.81	−0.9	0.83	−0.8
2	0.39	0.19	0.57	−2.6	0.6	−2.3
6	−0.06	0.21	1.06	0.3	1.06	0.3
3	−0.18	0.14	0.96	−0.2	0.95	−0.3
1	−0.29	0.2	1.24	1.2	1.19	0.9
4	−0.34	0.14	1.14	1	1.16	1.1
M	0.00	0.18	0.96	−0.2	0.96	−0.2
SD	0.32	0.03	0.22	1.3	0.21	1.2

Adj (True) S.D. .27 Separation 1.45 Strata 2.27
Model, Fixed (all same) chi-squared: 18.6 d.f.: 5 significance (probability): .00
Inter-Rater agreement opportunities: 720 Exact agreements: 387 = 53.7% Expected: 319.5 = 44.4%

Table 10.6 Rating scale statistics

Category	Absolute frequency	Relative frequency	Average measure	Outfit Mnsq	Threshold	S.E.
1	25	5%	−2.31	0.9	–	–
2	61	13%	−1.2	1.1	−2.63	0.25
3	231	48%	−0.01	0.9	−1.93	0.15
4	142	30%	1.15	1	1.05	0.12
5	21	4%	1.93	1	3.51	0.24

predictable and reliable in estimating the students' writing proficiency. However, the ZSTD statistics indicate that rater 2 performed predictably (ZSTD = −2.6, −2.3). This indicates that rater 2 was a very cautious rater who might try to award scores he/she thought other raters would give.

10.5.2.3 The Rating Scale

The score category statistics help to investigate whether the rating scale functions as intended and whether raters can use the scale in an acceptable manner. The two important indicators of whether the rating scale is clear and distinguishable for the rater are Outfit Mean Square index (Outfit Mnsq) and the step calibration measures for score categories, as listed in Table 10.6. It is generally held that Outfit MnSq less than 2.0 and the well-orderliness exhibited in the step calibrations for all the score categories would suggest that there are no major overlaps or step-disorderings in the use of the rating scale. The outfit of the current rating data ranged from 0.9 to 1.1, suggesting that the model expected measures largely matched the average measures.

In terms of the average measure, as shown in Table 10.6, the rating scale functioned as expected in that average measures increase strictly monotonically with rating scale category (Eckes, 2015). Lastly, the distances between calibrations of the adjacent score categories (between the first and the second, the third and the fourth, the fourth and the fifth category) lie within the recommended range (1.4–5.0 logits) (Linacre, 2004). But the step calibrations advance from −2.63 to −1.93 logits with a distance of 0.7 between the second and third category. When the thresholds are too close, the categories involved are less distinctive than intended. In this case, it is suggested to consider redefining the categories to have wider substantive meaning or combining categories (Eckes, 2015).

At the same time, we can also find that the raters did not use the band score 6 for scoring either task. This might be due to the fact that the students at the highest level of language proficiency (the top 5% students based on the results of the placement test in the university) did not participate in the trial test since they were studying academic writing as one of the compulsory courses at that time. For the rest of the students, the academic writing course is an elective course in the curriculum. Another possible reason was that raters were more cautious and sometimes hesitant to award the highest band score.

For the two writing tasks, the measure logit is 0.90 for Task A and −0.90 for Task B, which suggests the two tasks were not equally challenging to the students. The variability of the task difficulty is also confirmed in the significant chi-squared values. Possible explanations for the variability could be students' familiarity with the topics or genres of the tasks. But no evidence was collected in this study to support any of the explanations. The role of task variation in writing test design is clearly an area needing further research. At the same time, a two-way bias analysis (Rater by Task) was also run and there was no significant bias interaction between raters and tasks.

The above results show that the rating instrument (including the rating scale and the rating procedures) can, in general, discriminate among student scripts with different scores and raters overall can carry out the scoring tasks with an acceptable level of quality.

10.5.3 Use of the Rating Scale by the Raters

As well documented in the literature, the rating process is much more than merely a cognitive procedure, conditioned by text features, the rating scale as well as the rater factors. It was noted that the use of the rating scale contains different problematic aspects for the raters in the current study. Since the rating scale was inadequate to the complexity they have observed in the student scripts, this will lead to a tension between the publicly accessible and visible scale descriptors and raters' inaccessible and intuitive impressions. The following aspects for teachers' difficulty were identified:

Use of Holistic Rating Scale

Since the current rating scale is a holistic one, where no weight has been given to different features, raters were found to arrive at the final score by balancing good aspects of the script and its richness of ideas in the current study. Therefore, the same score could have different interpretations for different students. It is also reported by the raters that using a holistic score somewhat increased their cognitive demand if the scale was not clearly explained and specified in the training session.

Discrepancies have long been observed in previous studies as regards the validity and reliability associated with holistic and analytic scales in terms of rater characteristics and the task (Barkaoui, 2010; Hamp-Lyons, 1991; Li & He, 2015; Ohta et al., 2018). For example, raters might refer to some features which are easier to operationalize or rely on their prior teaching and assessment experiences when they use the holistic scale. It is suggested by the raters that the current rating scale may include some bullet points to help the raters to focus on the required rating dimensions. At the same time, more consideration should be given to the beneficial consequence of using the scale for test takers and teachers.

Rater Training and Standardization

Raters mentioned their cognitive load since they were left with the task of operationalizing the rating scale and weighing different features in the rating process. Some raters suggested the use of range finders that are sample writings at each band level selected by the test developer. Though raters achieved a relatively high level of reliability using the rating scale, rater training programs and training materials need to be in place to facilitate their consistent understanding and use of the scale. Special attention should be given to avoid raters using criteria that are irrelevant to the rating scale.

Difficulty in Operationalizing Certain Feature in the Rating Scale

In the university writing standards, cultural awareness is one dimension of the construct we wanted to include. As language is a carrier of culture, written expression should be concerned about the cultural differences between languages and respect conventions in different cultures. Relevant descriptors in the rating scale were also developed, which raters found difficult to interpret and use in the test. For example, for the highest band level, there is a descriptor like "*be fully aware of the relationship between languages and the culture it exists in*". Some raters pointed out that the definition and operationalization of this construct have overlapped to some extent with sociolinguistic competence, which concerns the use of linguistic knowledge in a social dimension. The same concern was shown in terms of developing the writing task to examine this construct.

Conciseness of the Wording of the Scale

Since the writing standard and test shared the same level descriptors, it is evident that the description of the standard cannot be directly used in the description. It is expected that the rating scale should be clearer and more concise, which will save lots of time for raters who want to review the rating scale during rating. As noted, the more complicated and detailed the descriptors in a scale, the less likely it is to be used consistently (Orr, 2002).

Scoring Two Tasks Using One Scale
It was also suggested that the rating scale should be task-specific, since the two writing tasks addressed different writing elements. It is the task design and difficulty which led to the raters' judgment that the two tasks were not targeting at same aspects of writing, which is one limitation of the study.

10.6 Conclusion

The study reports the validation activities of the rating scale of the SWPE writing test by employing an argument-based approach. The data provides evidence that the rating scale is generally working as intended. There are also some considerations about the future revision and application of the rating scale.

As discussed, the holistic scale was expected to be the most effective way to use in the current context and may account for the overall picture of students' writing performance. The rating scale, however, was reported to be challenging to apply when a student's performance was strong in some aspects but weak in others. This could be one possible limitation of using the holistic scale. An analytic scale, however, can account for uneven performance among students and can also provide more detailed feedback to both teachers and students. The comparison of the usage of holistic and analytic scales could be a focus for future study. Text analysis of students' scripts may be desirable in the future to explore salient patterns of language as one of the evidences for explanation inference. Besides, in the current study, the scores awarded by machine were not used for examination. Further research is required to investigate the scoring validity of using both human rating and automated scoring.

It is clear that this study did not collect backing evidence for the extrapolation, utilization and consequences inferences in the argument-based model and further validation of the scale is thus needed in this regard. What's more important is how the test could help promote the teaching and learning of English writing in the university, which will contribute to usefulness for the intended use and interpretation of the test in the local context (Weir, 2005). We also plan to conduct further studies on the consequences of the use of scale, including the impact on writing instruction and learning in the university.

Acknowledgements Our sincere gratitude goes to Professor Liz Hamp-Lyons and Professor Yan Jin as well as the reviewers who contributed their time and expertise to this chapter regarding the improvement of quality, coherence, and content presentation.

Appendices

Appendix I: Level Descriptors of the RUC-SWPE (Band 4–6)

Level	Descriptors
6	Students at this level Can produce clear, smoothly flowing, complex reports, articles or essays with significant points. Can organize complicated ideas in such a logical way that helps the reader to find significant points. Can highlight major points and relevant supporting details with appropriateness. Can write with only very rare inaccuracies of structure or vocabulary. Can use a variety of rhetorical devices such as irony, oxymoron, etc. Can engage readers effectively exploiting various stylistic devices.
5	Students at this level Can construct clear, detailed, well-structured and developed writing on a variety of topics. Can underline the relevant salient issues and round off with an appropriate conclusion. Can use accurate and mainly appropriate complex language which is organizationally sound with only occasional inaccuracies of structure and vocabulary. Can use basic rhetorical devices such as simile, metaphor, etc. Can engage the readers by using stylistic devices such as sentence length, variety and appropriacy of vocabulary, idiom and humor though not always appropriately.
4	Students at this level Can write essays in support of or against a particular point of view. Can write formal letters to a standard conventionalized format. Can expand and support points of view at some length with subsidiary points, reasons and relevant examples. Can communicate clearly using extended stretches of discourse and some complex language despite some inaccuracies of vocabulary and structure. Can organize extended, generally coherent writing on topics related to his/her field of interest.

Appendix II: Writing Tasks for the Trial Test

WRITING TASK 1

You should spend about 20 min on this task.

You live in a room in college which you share with another student. However, there are many problems with this arrangement and you find it very difficult to work.

Write a letter to the accommodation officer at the college. In the letter,

- describe the situation
- explain your problems and why it is difficult to work
- say what kind of accommodation you would prefer

Write at least 120 words. Begin your letter as follows:

Dear Sir or Madam,

WRITING TASK 2

You should spend about 40 min on this task.

Write about the following topic:

Nowadays many students have the opportunity to study for part or all of their courses in foreign countries. While studying abroad brings many benefits to individual students, it also has a number of disadvantages.

Do you agree or disagree?

Write at least 180 words.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*. https://doi.org/10.1177/0265532220927487

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371–383.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54–74.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010a). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010b). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*(4), 443–469.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*, 81–141.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL monograph 18). Educational Testing Service.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197–221.

Eckes, T. (2015). *Introduction to many-facet rasch measurement analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt: Peter Lang.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective*. Longman.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69–87). Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Ablex.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000—Writing: Composition, community, and assessment* (TOEFL monograph series, MS 5). Educational Testing Service.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275–304.

Knoch, U., & Chapelle, C. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*, 477–499.

Krapels, A. R. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 37–57). Cambridge University Press.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*, 178–212.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*, 95–110.

Linacre, J. M. (2010). *Facets Rasch measurement computer program* (Version 3.71.1). Winsteps.com

Linacre, J. M. (2019). *A user's guide to WINSTEPSVR MINISTEP Rasch-model computer programs*. Program Manual 4.4.7. https://www.winsteps.com/winman/copyright.htm

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Ministry of Education. (2016). *College English teaching guidelines.* Unpublished manuscript.

O'Sullivan, B. (2002). Investigating variability in a test of second language writing ability. *Research Notes, 7*, 14–17.

Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing, 38*, 21–36.

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143–154.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessment: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72–92.

SWPE Project Group. (2016). *Renmin University of China-standards of writing proficiency in English*. Renmin University Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12.

Valdes, G., Haro, P., & Arriarza, M. P. E. (1992). The development of writing abilities in a foreign language: Contributions toward a general theory of L2 writing. *The Modern Language Journal, 76*, 333–352.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.

Weigle, S. C. (2002). Assessing writing. Cambridge: Cambridge University Press.

Weir, C. J. (1988). Construct validity. In A. Hughes, D. Ported, & C. Weir (Eds.), *ELTS Validation Project: Proceeding of a conference held to consider the ELTS Validation Project report*. The British Council and University of Cambridge Local Examination Syndicate.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Macmillan.

Dr. Li Liu is currently Associate Professor at Renmin University of China. She has been actively involved in language teaching and a number of research projects in the field of language testing and assessment and teacher training. She is also leading the university-based language assessments at the RUC.

Dr. Guodong Jia is Professor of Linguistics and Applied Linguistics at Renmin University of China. He has been involved in many research projects related to computer-assisted language teaching and learning as well as language assessment and has published widely in the field.

Chapter 11
Examining Validity Evidence of an ESP Proficiency Scale: The Case of a Business English Writing Scale

Li Wang

Abstract The chapter reports on a validation project for a scale of Business English (BE) writing proficiency. The scale was empirically developed to facilitate the teaching, learning and assessment of BE writing in the Chinese tertiary context. To further examine the validity of the scale, semi-structured interviews were conducted to seek experts' perceptions of the scale. Specifically, ten experts from the pedagogical domain and business domain were carefully selected, whose opinions were elicited on an individual basis concerning the quality and usefulness of the scale. The experts in general perceived the scale favorably, commenting that most of the descriptors in the scale were appropriately categorized, formulated in a lucid manner and ascribed to proper proficiency levels. The experts in particular endorsed the usefulness of the scale, elaborating on how it could be applied to their respective workplace contexts. In the meantime, areas desiring improvement were also identified, shedding important light on the formulation and refinement of English for Specific Purposes (ESP) proficiency descriptors. Grounded in the ESP domain in general and BE research in particular, the study fills an important gap in literature by validating a theoretically informed, data-driven and statistically calibrated BE writing proficiency scale for Chinese EFL learners. Although this study focuses only on the BE writing skill, its findings have significant implications for scale development and validation in other discipline- or occupation-specific domains that feature the interaction between language and content.

Keywords Business English writing · Language proficiency scales · English for specific purposes · Validation

L. Wang (✉)
Xi'an International Studies University, Xi'an, China
e-mail: wangli@xisu.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_11

11.1 Introduction

In recent decades, language testing communities are facing increasing pressure to provide explicit information on test-taker performance (Hudson, 2005). The resultant pursuit of score meaningfulness has led to widespread adoption of language proficiency scales, which has enabled discussions on issues of language learning, teaching and assessment in a more fruitful and transparent manner (North, 2000). There have been, however, few attempts to develop, on an empirical basis, proficiency scales of English for Specific Purposes (ESP). Even rarer are studies illuminating how such a scale might be validated, which presents an important gap in our understanding of the ubiquitous assessment tool. Such a neglect deserves immediate research attention given the fact that ESP programs and tests, especially those of Business English (BE), are expanding at an ever quickening pace, complementing significantly the mainstream teaching and assessment of English for General Purposes (EGP) (Swales, 2000).

Against this backdrop, a BE writing proficiency scale has been empirically developed at the tertiary level in China, where BE education is gaining great momentum, necessitating hence an instrument that can be used as a common point of reference to facilitate the discussion of central issues involved in BE teaching, learning and assessment (Wang & Fan, 2021). The purpose of this chapter is to report a validation study for the BE writing proficiency scale, for the purpose of shedding light on scale validation to serve the wider assessment community.

11.2 Background to the Study

Language scales, also called “band scales, profile bands, proficiency levels, yardsticks”, normally consist of “a series of ascending bands of proficiency” (Council of Europe, 2001: 40). Due to a general movement towards more transparency in education and assessment systems as well as moves towards greater international integration, language scales mushroomed over the past few decades. Cautioning against risks associated with indiscriminate use of language scales, Alderson (1991) distinguished language scales as user-oriented, assessor-oriented and constructor-oriented ones according to their intended purposes of use. Pollitt and Murray (1996) took Alderson’s line of thought one step further by pointing out that scales could also be diagnostically-oriented.

In the early history of scale development, a scale was designed to serve one major function. The American Foreign Service Institute (FSI) Scale, for instance, was solely created to aid the rating of learners’ speaking performance. However, as the use of language scales gradually extended from the field of language assessment to language teaching and learning in general, a few scales were developed to cater to all the four scale functions outlined above. Such scales, due to their comprehensiveness, were also called “frameworks”, “standards” or “language proficiency scales” and were designed to serve as common points of reference in a range of education and assessment contexts. A typical example of such scales is the Common European Framework of Reference (CEFR), which is extensively used in countries and regions around the globe as a reference for the learning, teaching and assessment of languages (Little, 2007). Similar to the CEFR, China's Standards of English (CSE) was also a comprehensive proficiency scale. Officially launched in 2018, the CSE serves as a set of standards guiding the learning, teaching and assessment of English in China (e.g., Jin et al., 2017; Liu, 2019).

As language proficiency scales are widely used throughout the years, an important gap that needs to be filled is that little is known about ESP proficiency scales, although the development of rating scales in the ESP domain is receiving more attention (e.g., Knoch, 2014; Pill, 2016). Very often, when descriptions for ESP proficiency are required, EGP scales, in most cases the CEFR, will be adopted to fulfill the purpose. Such a practice has been increasingly questioned today given the uniqueness of language use features and tasks in the ESP domain (Kim & Elder, 2009).

In light of the dearth of empirical studies in the field of ESP scale development, a proficiency scale of Business English writing has been developed in the context of English language education at the tertiary level in China, where BE teaching and learning are gaining momentum alongside China's socio-economic advancement. The development of the scale went through the following three phases.

Phase I: Establishing a Descriptive Scheme and Collecting Descriptors A descriptive scheme of the scale incorporating two broad categories, one theory-based, the other activity-based, was established to enhance the scale's theoretical rigor and practical utility (North, 2000). Specifically, the model of Genre Knowledge proposed by Tardy (2009) was employed as the theoretical framework, informing identification of the components of the descriptive scheme. By consulting sources such as widely-recognized models of language proficiency and language scales, the dimensions in the model of Genre Knowledge were operationalized into nine subcategories (i.e., *Vocabulary*, *Grammar*, *Orthographical control*, *Genre format*, *Cohesion & coherence*, *Strategic competence, Sociolinguistic competence Intercultural competence* and *Business knowledge*), representing different aspects of BE writing proficiency. In addition, subcategories of the activity-based category were determined mainly on the basis of (a) an analysis of BE writing textbooks widely used in China; and (b) a review of the task types included in the writing section of professionally developed BE tests. For practical reasons, the list was selective and included only the broad macro-genres (e.g., report) rather than the more specific genres under them (e.g., feasibility report, progress report). This process yielded 10 genre-based descriptive categories, including *Letter, E-mail, Report, Memo, Minutes, Note, Press release, Resume, Summary* and *Documentary materials.* As such, the descriptive scheme of the scale was made up of 19 descriptive categories, with 9 theory-based ones and 10 activity-based ones.

After the descriptive scheme was established, descriptors of BE writing proficiency were collected. Sources for descriptor collection included existing language scales, BE tests, BE curricular requirements and recruitment advertisements, etc. The descriptors were tentatively put under the categories in the descriptive scheme, which were then validated in Phase II of the study.

Phase II: Teacher Evaluation of the Appropriateness of the Descriptive Scheme and Descriptors in Two Workshops In the second phase of the study, two workshops were held, in which 15 teachers collectively examined the relevance of the descriptive scheme and descriptors to the Chinese tertiary context. These teachers had either linguistics or/and business-related academic backgrounds, in addition to rich experiences in teaching BE courses, especially BE writing courses. In the first workshop, after a brief training session familiarizing them with the purpose of the study and the important concepts such as language proficiency scales, the teachers discussed whether the descriptive categories included in the descriptive were relevant to BE teaching and learning in the Chinese tertiary context. Based on their comments, the initial descriptive scheme was revised by collapsing E-mail and Letter into one category (Letter & E-mail) due to their overlap in modern business communication. In addition, the category of *Documentary Materials* was deleted due to its elusive nature.

After the first workshop, the teachers were given 1 month to individually analyze and evaluate the quality of all the descriptors which were provisionally allocated under the categories of the descriptive scheme. The second workshop spanned three sessions. In the first two sessions, the teachers worked in small groups to compare their individual evaluation of the descriptors for the selection of the well-written ones. In the last session, all the teachers went through the descriptors selected on the basis of small group discussions in order to make sure that only the best descriptors would be used in the next phase of the study for the development of a questionnaire.

Phase III: Scaling the Descriptors and Setting Cut-Offs In the third phase of the study, a questionnaire was constructed, the bulk of which was made up of the descriptors selected from the teacher workshops during Phase II. Altogether 572 Chinese university students who had received formal instruction in BE writing were asked to rate the difficulty levels of the descriptors on a five-point Likert scale. On the scale, 1 indicated "not difficult at all" and 5 indicated "very difficult". The questionnaire data were then submitted to Rasch analysis using Winsteps version 3.70.0.3 (Linacre, 2012) to calibrate the difficulty levels of the descriptors and identify misfitting items. The descriptors were then assigned to three proficiency bands on the basis of their logit values, with each band demonstrating approximately equal intervals of logit difference. Besides, content analysis of the descriptors was also conducted to ensure there was an apparent gap of abilities across the three bands (North, 2000), marking the completion of the development of the scale.

The resultant scale is composed of both a horizontal dimension and a vertical dimension. The horizontal dimension is a descriptive scheme incorporating 16 descriptive categories (i.e., *Vocabulary*, *Grammar*, *Orthographical control*, *Genre format*, *Cohesion & coherence*, *Strategic competence*, *Sociolinguistic competence, Intercultural competence*, *Business knowledge*, *Letter & E-mail*, *Report*, *Note*, *Minutes*, *Memo*, *Resume* and *Summary*). The vertical dimension is a definition of BE writing proficiency at three consecutive levels—(a) Advanced (*Level C,* to indicate that the learner has achieved an advanced level of competence suitable for complex business writing tasks), (b) Intermediate (*Level B*, to indicate that the learner is effective enough to handle moderately complex writing tasks pertinent to his/her field of expertise), and (c) Lower Intermediate (*Level A*, to indicate that the learner can access the business world by handling daily, routine tasks in familiar occupational context). Integration of the horizontal dimension and vertical dimension is presented as 16 illustrative scales substantiated by 86 empirically calibrated descriptors (see three examples of the illustrative scales in Appendix A). It is envisaged that with comprehensive descriptions of BE writing proficiency, the BE proficiency scale can provide a point of reference for the elaboration of BE curricular guidelines, syllabi and test specifications, contributing hence to greater transparency and coherence of BE teaching, learning and assessment in the Chinese tertiary context.

11.3 Validation of the BE Writing Scale

To further enhance the validity of the scale, the current study was carried out to probe into experts' perceptions of the scale. Specifically, it addressed the following two questions:

1. What are the experts' perceptions of the quality of the scale in terms of descriptor categorization, descriptor clarity and descriptor level assignment?
2. What are the experts' perceptions of the usefulness of the scale in their respective workplace contexts?

11.3.1 Study Design

Ten experts, five teaching experts (denoted as TE1–5 in this study) and five domain specialists (denoted as DS 1–5 in this study), took part in the validation study. To enhance the participants' representativeness, purposive sampling technique was adopted by making a principled selection of the experts in light of their expertise in the pedagogical and occupational fields.

The five teaching experts had over 10 years' experience in teaching BE writing courses. All of them had worked as supervisors for new BE teachers and postgraduates majoring in BE in their respective institutions. In addition, all of them were involved in the development of the National Curriculum designed for Business majors in China. The five domain experts worked in the China-based branches of multinational companies. With over 10 years' working experience, the domain experts were selected via recommendation as experienced and competent business English writers in their respective companies. At the time of this study, all participating domain experts had acquired certificates in CET6 (approximately equivalent to B2 or C1 on the CEFR), BEC Higher (equivalent to C1 on the CEFR) or IELTS Band 7 (equivalent to C1 on the CEFR). Their daily language of written communication was English as their clients or co-workers included native English speakers in addition to ESL (English as a Second Language) speakers from countries such as Singapore, India and Germany. Holding an MA degree in fields such as Accounting, Engineering, and Marketing, they specialized in areas ranging from product design, human resources to procurement.

According to Corbin and Strauss (2008), interviewing is a valuable avenue of research inquiry, opening a door for going into intricate details about unobservable traits such as thought processes. Semi-structured interviews were therefore conducted on an individual basis to explore these experts' perceptions of the scale. An interview guide (see Appendix B) with a set of open-ended questions was designed to facilitate the interviewing process. The focus of the guiding questions was placed on the experts' comments on (a) the quality of the scale in terms of descriptor categorization, descriptor level assignment and descriptor clarity and (b) the usefulness of the scale. According to Berg (2009), it is important to pilot the interview guide before the actual interview takes place. A specialist in language assessment was therefore invited to examine whether there were unclear, inappropriate, or poorly worded questions. Minor revisions were then made on the basis of the specialist's suggestions.

11.3.2 Data Collection and Analysis

During the interviews, the experts were firstly introduced to the purpose of the study. They were also shown sections of the CEFR to facilitate their understanding of the structure, content and functions of language proficiency scales. During the interviews, the participants were asked about their general impression of the scale, their perception of the quality of the scale and how the scale might be applied to their working contexts. They were in particular encouraged to reflect on their BE writing or teaching experiences in support of their comments. Chinese was used during the interviews and the participants were allowed to code switch. All the interviews were audio-taped with the participants' permission.

The interview data were verbatim transcribed and analyzed by means of analytic induction (Goetz & LeCompte, 1984) and constant comparison (Miles & Huberman, 1994). To gain a holistic view of the data, preliminary readings were firstly conducted. The transcripts were then analyzed in greater detail to identify the salient themes, which were further refined by grouping similar or interrelated themes together. A specialist in language assessment was also invited to go through a second round of categorization and coding from scratch as an independent coder. In case of discrepancy of coding, the data were revisited and discussions were carried out until agreement was reached.

11.4 Results and Discussion

11.4.1 Quality of the Scale

The experts' judgments on the quality of the scale, which fall into three categories, are presented in Sect. 11.4.1.1: (1) descriptor categorization, (2) descriptor clarity and (3) descriptor level assignment. In Sect. 11.4.1.2, the experts' comments on the usefulness of the scales are analyzed.

11.4.1.1 Descriptor Categorization

On the whole, the experts considered the descriptors included in the scale well categorized. Descriptor 17, which read "*Punctuation is reasonably accurate*." under the category of Orthographic Control, stood out as a problematic one.

One domain specialist, for instance, resorted to his intuitive feelings and made the following remark:

DS5: I just feel it a bit weird to put punctuation under the category of spelling.

The teaching experts, by contrast, contributed more perspectives as to why this descriptor was inappropriately categorized. Two such comments are shown below:

TE2: I deem it improper to put students' ability to use punctuation adroitly under the category of Orthographical Control. Orthographical Control, to my understanding, is mainly concerned with the spelling of words, which is the focus of the other descriptors in the category.

TE5: Most of the descriptors in the category are related to the correct forms of words, such as spelling and capitalization rules. This descriptor, however, has nothing to do with word forms.

TE3 offered a valuable suggestion concerning the modification of the category by referring to her experience in teaching English writing in America, as is demonstrated below:

TE3: You might want to change the category heading from Orthographical Control to Mechanics... As we can see from many writing textbooks published in America, aspects like spelling, punctuation and capitalization are usually discussed under the umbrella term Mechanics.

TE3's suggestion was taken up and this category, originally labelled as "Orthographical Control" was revised as "Mechanics" to render the category name more inclusive.

11.4.1.2 Descriptor Clarity

With regard to the clarity of the descriptors, the experts expressed concern over three descriptors: Descriptor 37, Descriptor 45 and Descriptor 54. Problems identified by these experts were mainly related to (a) the use of terms unfamiliar to people without a linguistics-related background, (b) unclear examples in the descriptors, and (c) the differences between BE written communication and oral communication. Illuminating comments highlighting problems identified in these descriptors are presented below.

Descriptor 37 (under the category of Sociolinguistic competence): *Appreciates fully the sociolinguistic implications of language used by speakers of other cultures and can react accordingly.*

The major issue raised about this descriptor was the term "sociolinguistic implications", which caused confusion on the part of all the domain specialists. They reported that this term hindered their comprehension of the descriptor. The following comment is reflective of this problem:

DS2: I don't know what "sociolinguistic implications" means. Structurally, I can tell that it is emphasized in this descriptor. So I have to guess its meaning from the other words in the descriptor.

All the teacher experts, on the contrary, had no difficulty understanding this descriptor, which may be attributed to their background in Applied Linguistics. Nevertheless, in light of the problems mentioned by the domain specialists, the term "sociolinguistic implications" was removed from the descriptor. Eventually, the descriptor was revised as "Appreciates fully the effects of social factors (e.g., power relations) on language use and can react accordingly" to make it more reader-friendly to a wider range of potential users.

Descriptor 45 (under the category of Intercultural Competence): *Is aware of the effect of non-linguistic elements on intercultural business written communication, including time, space, etc.*

Many teaching experts viewed Descriptor 45 negatively. TE1, for instance, brought up an important point that in business written communication, the role played by nonverbal factors was minimal as compared to business oral communication:

TE1 Do nonlinguistic elements mean nonverbal clues here?… I think they typically refer to gestures and facial expressions in face-to-face communication. However, writing by itself is a special form of communication devoid of any nonverbal clues.

TE3, from a different perspective, pointed out that confusion might be generated by the example included in the descriptor. She said:

TE3 The example in this descriptor highlights the important role played by time and space in business written communication. Yet I am not sure how time

> and space can affect written communication... Does the word "space" refer to the distance kept between two people or the space intentionally left for better structuring of the written text? This example is very confusing.

Domain specialists, by contrast, stressed the importance of nonverbal factors in business communication. However, scrutiny at their comments reveals that the impact exerted by time and space actually has little to do with business written communication, as is demonstrated by the following remarks:

DS4 Time affects business communication greatly. For instance, our French clients attach great importance to vacations and don't have the habit of working overtime. We need to adjust our schedule to their cultural preferences in order not to be thrown into a mess, especially when we have an urgent project to complete.

DS3: When we negotiate with others, space arrangement is a very sensitive issue. The two parties' relationship determines the space kept between them.

Taking into consideration the experts' opinions presented above, descriptor 45 was deleted from the scale because (a) its emphasis on nonverbal elements was more related to the oral form of BE communication instead of the written form, and (b) the example included in the descriptor might lead to confusion on the part of the readers.

Due to similar reasons, Descriptor 54 which reads "Demonstrates awareness of the sources from which information of prospective clients can be accumulated to establish new business relationships" was also removed from the illustrative scale "Business Knowledge".

11.4.1.3 Descriptor Level Assignment

The experts considered that most of the descriptors were ascribed to appropriate proficiency levels. As to the controversial ones, analysis of the interview data reveals two important issues worth mentioning: (a) The inclusion of examples may greatly affect users' perception of descriptor difficulty; (b) The teaching experts and domain specialists differ quite significantly with regard to their perception of the difficulty of some of the tasks included in the scale descriptors. Of the problematic descriptors, descriptor 35 and Descriptor 58 serve as two typical examples reflective of these issues and will be discussed below.

Descriptor 35 (under the category of Strategic Competence): *Can provide an appropriate and effective logical structure (e.g., heading, topic sentence) to facilitate reader comprehension of the business text.*

The experts' comments on Descriptor 35 demonstrated how the examples included in a descriptor might affect reader judgment on its difficulty level. The quotes below are illuminating:

TE5 Initially, I put it at Level C because I thought it was fairly difficult to provide an appropriate and effective logical structure to facilitate reader comprehension of a business text… but later when I saw the examples "heading and topic sentence", I changed it to Level B, because according to my teaching experiences, students usually have little difficulty mastering the writing of headings and topic sentences.

DS1 considered the descriptor easier than TE5 by putting it at Level A. The rationale provided by her is presented as follows:

DS1 The "heading" of an E-mail, to my understanding, is equivalent to the "subject" of an "E-mail". Whenever we write an E-mail, we think of a subject for it. The subject just highlights the most important message in the E-mail and I think it is something that every BE writer can do. It is very easy.

The observations made above demonstrated that the examples included in the descriptor turned out to be a confounding factor preventing the experts from reaching an agreement in terms of descriptor difficulty. The examples were thus deleted from the descriptor, which was later revised as "Can effectively structure the business text so as to facilitate reader comprehension".

Descriptor 58 (under the category of Letter & Email): *Can write a(n) letter/E-mail in an official capacity to an organization, concerning a straightforward business problem or transaction.*

The proficiency level of Descriptor 58, which was concerned with the genre of Letter & Email, was perceived differently by the two expert groups. Specifically, all the domain specialists considered the task described in the descriptor much easier than the teachers. DS4, for instance, put the descriptor at Level A by offering the following explanation:

DS3: On a daily basis, we, representing our company, write E-mails to our clients. This is a very simple routine task. So I place this descriptor at Level A.

The teaching experts, by contrast, tended to put the descriptor at higher levels. TE5, for instance, ascribed the descriptor to Level C and below is his rationale:

TE5 I think it is very difficult for students who have no working experience to write E-mails "in an official capacity"… The task may involve challenging sales skills as it emphasizes business transactions.

After elaborating on his choice, TE5 went a step further and made an important distinction between pre-experience BE learners and job-experienced BE learners, contending that "the task may be daunting for students, but for in-service employees, I think the task is a very basic requirement".

The divergence of opinions identified and the remarks of TE5 revealed that this descriptor was population-sensitive: for experienced workers, the ability to complete such a task might be considered elementary, but for pre-service learners such as college students, the task might be quite challenging. In light of this observation, the phrases (e.g., "*in an official capacity*", "*business transactions*") in the descriptor were removed and the descriptor was simplified as "Can write a(n) letter/E-mail concerning straightforward business issues" to render it more context-independent.

In a similar vein, three other problematic descriptors were modified or revised after exploration of the qualitative data.

11.4.2 Usefulness of the Scale

Generally speaking, both teaching experts and domain specialists approached the scale favorably and explored its potential usefulness in their respective workplace contexts. Interesting differences were also found with regard to how the scale might be used by the two expert groups.

11.4.2.1 Perceived Usefulness of the Scale in the Pedagogical Context

The teaching experts expressed positive attitudes towards the pedagogical value of the scale. TE1, for instance, commented that "the detailed description of BE writing proficiency can provide a valuable reference for BE learners to know about their current writing proficiency and their objectives of learning". When it comes to BE teaching, TE2 applauded the comprehensiveness of the scale and explained how it could help BE teachers expand their scope of instruction.

TE2: The comprehensiveness of the scale, particularly the inclusion of illustrative scales such as Strategic Competence and Intercultural Competence that dwell on aspects other than language per se, can greatly enhance the effectiveness of BE teaching in China... You know, I have read extensively and conducted research on the status quo of BE writing instruction in China and found that teachers focus too much on the formal aspects of language teaching, such as grammar, vocabulary and genre format, at the expense of the rest of the equation.

In a similar vein, TE5 highlighted the scale's potential usefulness for complementing what is not included in BE textbooks, as shown in the following statement:

TE5 Currently, business English writing textbooks abound in China and there exist great differences among them. Some merely focus on the writing of E-mails to carry out international trade; others touch upon a very limited range of genre tasks. The scale can remind teachers of the important genre types that students are likely to encounter in the business community. The teachers can then prepare their teaching materials in a more informed manner to complement what is missing in their textbooks.

While acknowledging the usefulness of the scale, the teachers also provided suggestions for the sake of its betterment. TE4, for instance, offered the following advice to render the scale more accessible to classroom practitioners:

TE4 I think the inclusion of examples is very important to help teachers gain a better understanding of what the descriptor is driving at. However, as I have noticed, not all descriptors in this scale provide examples. Maybe this is an area that needs to be addressed later.

11.4.2.2 Perceived Usefulness of the Scale in the Corporate Context

Domain specialists also confirmed that the scale could be useful in their workplace context. They resorted to different perspectives as to how the scale would be helpful in their companies. DS3, for instance, stressed the scale's value for in-service language training in her company.

DS3: This scale touches upon different aspects of business English writing…It can be used as reference material for us to provide training for staff who will work in overseas projects.

While heralding the comprehensiveness of the scale, some domain specialists emphasized that when it came to individuals, the scale should be used selectively since in-service BE writers might not be interested in all the skill areas included in the scale; rather, they would focus on the ones most relevant to their job demands. This view was reflected by the quotes below:

DS1: E-mail writing is the most important writing task that staff in our company need to cope with… By contrast, they may never be required to write a summary… Also, emphasizing resume writing ability is like encouraging them to leave our company.

DS2 Staff in our company work in different departments and their writing needs vary according to their job responsibilities. For instance, report-writing may be more relevant to project managers than secretaries, who instead write minutes more often… They can find out the tasks most relevant to their writing needs and selectively use the scale as a reference material.

Taking a different angle, DS4 and DS5 touched upon the linguistic features of BE writing, highlighting that in real-life BE communication, meaning took precedence over form. They argued that elements such as grammatical complexity and genre format should be less heavily weighed in the sale than the ability to master terminology commonly used in the business domain. Below are their comments:

DS4 In our daily communication, complex language is actually discouraged. Simple language can maximize communicative effectiveness...Format is not important either, as long as the message is communicated.

DS5 When we write English E-mails, mastery of terminology shared by both parties involved in the communication is very important... much less emphasis is put on how well or how complex the sentences are formulated... Our primary goal is to get our messages across and make sure no ambiguity or misunderstanding would arise.

Apart from the above-mentioned comments, the domain specialists also offered suggestions to further enhance the scale's validity. They pointed out that "a finer distinction of proficiency levels might be more informative" (DS1) and "space should be provided to add more job-related descriptors to the scale" (DS3). In addition, they also raised concern over the illustrative scale under the heading of Note. Specifically, they argued that note-writing was an artificial activity that barely had any relevance to workplace realities. Two explanations are presented below:

DS1 I seldom write notes. If I have a message to deliver but I can't find the receiver, I will send him an E-mail. If he doesn't respond, I will call him or text him.

DS5 Note is often used to communicate unimportant matters, but this descriptor [i.e., Can write notes in appropriate language to convey important information of immediate relevance to superiors or new clients, getting across the key points] emphasizes the ability to use note to convey important messages, which is unlikely to happen in reality.

In response to the criticisms, the illustrative scale Note was deleted along with its descriptors.

11.5 Implications and Conclusion

This chapter presents the findings emerging from the qualitative validation study of a business English writing proficiency scale, which describes BE writing proficiency across three consecutive levels. Drawing on a sample of ten experts made up of teaching experts and domain specialists, empirical evidence has been accumulated, supporting the validity of the scale originally developed on the basis of students' self-reported data. Findings of the study have significant implications for BE assessment as well as ESP scale development and validation.

In terms of BE assessment, the construct of BE writing in the validated scale was conceptualized from the perspective of genre and the model of Genre Knowledge (Tardy, 2009) was adopted as the theoretical framework of the scale. As mentioned in Sect. 11.2, the model was operationalized as nine sub-categories, featuring a combination of linguistic and nonlinguistic factors such as Vocabulary, Sociocultural competence and Business knowledge. The experts' perceptions of the scale indicated that attention to both linguistic and nonlinguistic elements was important for the operationalization of BE writing proficiency. In the field of language assessment, however, nonlinguistic factors such as subject knowledge are often treated as construct-irrelevant variables that should be strictly controlled for the measurement of language proficiency (Bachman, 1990). This study reveals that consideration of the impact exerted by nonlinguistic factors on language performance is crucial to the valid conceptualization of BE writing proficiency. This finding thus lends support to the argument calling for a change of the prevalent employment of language-focused criteria in assessing BE writing proficiency for the sake of establishing more relevant ones.

With regard to ESP scale development and validation, three important issues are highlighted by the findings. Firstly, the study found that the experts' interpretation of a descriptor might be considerably affected by the inclusion of examples, as any incongruence spotted between the descriptor and the example would severely impair the interpretation of a descriptor. In light of the fact that informative examples do facilitate readers' comprehension of descriptors, this finding highlights the significance of being more rigorous in the selection of examples in the stage of descriptor formulation when developing an ESP scale. The second issue is concerned with the application of ESP scales in different contexts, especially when varying groups of users are concerned. For instance, comprehensiveness of the scale tends to be welcomed in the pedagogical context, which is understandable as one of the goals of BE writing education in China is to prepare students adequately for future real-world writing challenges. Selective use of the scale, by contrast, is advocated in the corporate context which places high demands on a narrower range of job-related writing skills or tasks. The third issue is concerned with the differences identified between teachers and domain specialists. On the whole, teaching experts in the study stressed factors such as linguistic accuracy, complexity and genre format in creating BE written texts. However, the importance of these formal aspects was played down by the domain specialists who gave priority to the effectiveness of meaning exchanges. The divergent foci from the two groups therefore reiterate the necessity of juxtaposing the insights derived from the pedagogical and workplace contexts to inform ESP scale development and validation (e.g., Elder & McNamara, 2016; Knoch, 2014).

Although the study was carefully executed, its findings should be interpreted by considering two limitations. Firstly, only ten experts were involved in the study. A larger sample size might have contributed to a deeper understanding of the issue under investigation. Secondly, only two groups of potential stakeholders of the scale, namely teaching experts and domain specialists, were invited to participate in the study. The exploration of other data sources, such as BE learners at collegiate settings, native speakers of English specialized in BE writing, language testing experts or novice BE writers navigating to the culture of real-world business communication, is desired to further enhance the validity of the scale.

According to Weir (2005), developing and validating language proficiency scales is a dynamic process rather than a one-for-all endeavor. Validation of the BE writing scale therefore needs to continue along with the development of language testing theories and practices. Views and comments offered by a wider range of stakeholders shall be collected to inform later revision of the scale. The next logical step is to examine the extent to which the descriptions of BE writing proficiency in the scale accord with learner proficiency levels derived from well-constructed language tests. Another area worth exploring is to explore whether the scale developed in the Chinese context can be applied to other contexts.

Grounded in the ESP domain in general and BE research in particular, the study fills an important gap in literature by validating a theoretically informed, data-driven and statistically calibrated BE writing proficiency scale developed for Chinese EFL learners. Although this study focuses only on the BE writing skill, its findings have significant implications for scale development and validation in other discipline- or occupation-specific domains that feature the interaction between language and content.

Acknowledgements I would like to thank Professor Liz Hamp-Lyons and Professor Yan Jin for their valuable suggestions for the betterment of this chapter.

Appendices

Appendix A: Three Examples of the Illustrative Scales of the Scale of Business English Writing Proficiency

Vocabulary

Advanced	Has a good command of a very broad business lexical repertoire including idiomatic expressions and colloquialisms. Demonstrates satisfactory control over synonyms such as *merger*, *consolidation* and *acquisition*, especially their connotative levels of meaning in the business environment. Can use a wide range of business vocabulary fluently and flexibly to convey precise meanings.
Intermediate	Demonstrates familiarity with business specialist terms (e.g., *offer, bill*), whose meanings are very different when they are used in non-business context. Demonstrates an awareness of current business terminology. Demonstrates familiarity with common acronyms and abbreviations used in business documents such as *HR* and *SWOT*.
Lower Intermediate	Has a basic vocabulary repertoire of isolated words and phrases related to general business topics.

Socio-Cultural Competence

Advanced	Appreciates fully the effects of social factors on language use and can react accordingly.
Intermediate	Can adjust the level of difficulty and complexity of business text appropriately to the writing occasion. For instance, can drop all technical jargons when the reader is an outsider of the field. Shows a high level of appreciation of register. For instance, can adopt a conversational tone where the situation desires informal communication. Can express him or herself confidently, clearly and politely in a formal or informal business register, appropriate to the purposes and audience (s) concerned. Is aware of the salient politeness conventions and acts appropriately. For instance, can keep discriminatory or derogatory language (e.g., sexist salutation) out of one's business communications.
Lower Intermediate	Can establish basic business contact by using simple polite expressions.

Report

Advanced	Can write a formal report that is analytical in nature to present a case or give critical appreciation of proposals. For instance, can write a feasibility report which provides data, analyses and recommendations to win approval of a project. Can write a report that synthesizes a large amount of data and complex information, collected through methods such as questionnaires, interviews. For instance, can coherently present information and state a position on a previously researched topic.
Intermediate	Can write a report in response to requirements of information. For example, can write a series of progress reports on a regular basis to provide information on the progress of a project. Can write a business report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options.
Lower Intermediate	Can write very simple reports which pass on routine information.

Appendix B: Interview Guide

1. Please briefly introduce your daily BE writing/teaching experiences.
2. What is your general impression of the scale?
3. Do you think the descriptors are properly categorized?
4. Do you think the descriptors are assigned to appropriate levels?

5. Are there any vague or unclear descriptors?
6. Are there any other areas that require future revisions?
7. To what extent do you think the scale is useful in your workplace context?

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). Macmillan.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Berg, B. L. (2009). *Qualitative research methods for the Social Sciences*. Allyn & Bacon.

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.

Council of Europe. (2001). *Common European framework of reference for language learning, teaching and assessment*. Cambridge University Press.

Elder, C., & McNamara, K. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing, 33*(2), 153–174.

Goetz, J. P., & Lecompte, M. D. (1984). *Ethnography and qualitative design in educational research*. Academic.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*, 205–227.

Jin, Y., Wu, Z., Alderson, C., & Song, W. (2017). Developing the China Standards of English: Challenges at macro-political and micro-political levels. *Language Testing in Asia, 7*(1), 1–19.

Kim, H., & Elder, C. (2009). Understanding aviation English as a Lingua Franca: Perceptions of Korean Aviation Personnel. *Australian Review of Applied Linguistics, 32*(3), 1–17.

Knoch, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes, 33*, 77–86.

Linacre, J. M. (2012). *Winsteps tutorial.* Retrieved April 7, 2014, from http://www.winsteps.com/tutorilas.htm

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645–655.

Liu, J. (2019). China's Standards of English language ability. *Foreign Languages in China, 16*(3), 10–12.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods*. Sage.

North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang.

Pill, J. (2016). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing, 32*(2), 175–193.

Pollitt, A., & Murray, M. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Language testing 3: Performance, testing, cognition and assessment* (pp. 74–91). Cambridge University Press.

Swales, J. M. (2000). Languages for specific purposes. *Annual Review of Applied Linguistics, 20*, 59–76.

Tardy, C. M. (2009). *Building genre knowledge*. Parlor Press.

Wang, L., & Fan, J. (2021). *Working towards a proficiency scale of business English writing: A mixed-methods approach*. Springer.

Weir, C. (2005). *Language testing and validation*. Palgrave Macmillan.

Dr. Li Wang is Associate Professor of Applied Linguistics at the School of Economics and Finance, Xi'an International Studies University. Her research interests include English for Specific Purposes and Language Testing and Assessment. She has published in international and local journals such as Asian EFL Journal, Asian TEFL Journal, and Foreign Language World.

Chapter 12
Factors Impacting Upon Writing Teachers' Feedback Choices

Jing Yang

Abstract The study investigated two expert writing teachers' feedback and assessment practices and their emic (insider) perspectives on factors impacting on those practices. Specifically, the teachers shared their expertise knowledge and skill in dealing with formative assessment. Data were collected at a distinguished university in Northeast China over a 17-week semester through written texts, interviews, think-aloud and stimulated recalls. The study identified multiple factors influencing the teachers' feedback choices. The first notable factor was the two teachers' belief in the value of multiple drafts on a regular basis. The second factor was their belief in peer review. To apply peer review effectively among Chinese writing students, they provided systematic sustained support and supervision for peer reviewers. Another important factor guiding their feedback choices was the alignment between writing assessment rubrics and class instructional focal points. The two teachers treated feedback on paper not as an isolated act but as part of the teaching cycle. Teacher feedback and teacher assessment were not only to reflect but also to inform instruction.

Keywords Teacher feedback · Formative assessment · Expert teacher

12.1 Introduction

12.1.1 Motivation for the Study

One major motivation for the study arose from the observation that in the teacher feedback research, teachers' own voice was not often heard. An in-depth understanding of teachers' knowledge and beliefs underlying their actual practices can lead to a fuller and more valid conceptualization of teaching, rather than a superficial behavioral representation of teaching (Borg, 2006). The value of understanding the

J. Yang (✉)
Dalian University of Foreign Languages, Dalian, China
e-mail: yangjing@dlufl.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_12

mental side of teachers' work is that insightful information gained from research can be put to effective use in teacher education and development programs by encouraging teachers to develop their personal systems of knowledge, beliefs, and understandings drawn from their practical experiences of teaching (Freeman, 2002; Freeman & Richards, 1996; Richards, 2010). In her investigation into distinctive qualities of expert teachers, Tsui (2009: 429) found that one key quality of expert teachers is that they are capable of *theorizing practical knowledge* (i.e., "making explicit the tacit knowledge that is gained from experience") and *practicalizing theoretical knowledge* (i.e., "making personal interpretations of formal knowledge, through teachers' own practice in their specific contexts of work"). Important as they are, there is limited research on the actual processes of writing instruction and teacher feedback as perceived by writing teachers themselves (Ferris et al., 2011; Goldstein, 2005). Meanwhile, inferences are frequently made about teachers' intentions to employ particular feedback strategies without consulting the teachers themselves. This is problematic because no matter how well researchers may know the teachers, their assumptions may be incorrect (Ferris, 2014; Ferris et al., 2011; Goldstein, 2001, 2005).

12.1.2 Context of the Study

This study is also intended to address a practical concern of Chinese EFL writing teachers. Studies on teacher knowledge and beliefs about feedback, very limited in number, have been largely conducted in ESL contexts. Little relevant research has been conducted so far in EFL contexts such as mainland China. This consideration of language education contexts is important because findings of research conducted in ESL contexts may not be applicable to the EFL context of mainland China, given its unique English writing curricula, assessments, and pedagogical approaches.

English teachers in mainland China seem to face several obstacles when teaching writing in college English classes. Insofar as approaches to teaching English writing are concerned, although process-oriented writing has been imported and encouraged, pre-writing and multiple-drafting activities have appeared, and concepts of peer review and portfolio assessment are being tested out in classrooms, the traditional product-based pedagogy still dominates the majority of writing classes at universities in mainland China (Mei & Yuan, 2010; You, 2004b; Zhang, 2005). This pedagogy, typical of many Chinese universities (see Wang, 2010), is problematic because in product-oriented classrooms teacher feedback tends not to be taken seriously when revision is not required (Ferris, 2003; Hyland & Hyland, 2006). Moreover, college English teachers carry the heavy burdens incurred by the large college enrollment expansion. For most writing teachers, one of the burdens is to give feedback to writing submitted by a large class of students (Wei & Chen, 2003; Yang et al., 2006).

12.1.3 Factors on Feedback Choices

Factors contributing to a predominant product-oriented feedback and assessment approach include the local culture of education (see Hu, 2002; Yu et al., 2016, for more about Chinese culture of learning), test-oriented teaching (see Pan & Block, 2011, for a detailed illustration of the "put exams first" view), and big class size (You, 2004a, b). The rapid expansion of Chinese higher education at both undergraduate and postgraduate levels has resulted in rising class sizes that pose a number of challenges to English teachers (Jin & Cortazzi, 2006). This is confirmed in Du's (2012) interviews with the Heads of English Departments from three Chinese top-tier universities, who reported class sizes ranging between 40 and 90 students.

Two studies (i.e., Yang, 2010; Zhang, 2008) with their particular focus on the beliefs of English writing teachers in mainland China are worthy of special attention. In terms of the relationship between writing teachers' beliefs and practices, the two studies came up with different findings. Zhang (2008) explored in a qualitative case study the beliefs and practices of five university English writing teachers who taught non-English major students. The study found teachers believed that writing was a complicated cognitive process and that teachers should design and use communicative activities like group discussion and peer feedback. However, observation data demonstrated that these beliefs were not accordingly enacted in class or in feedback.

Complementing Zhang's study with writing teachers in a middle-ranking university, the second study by Yang (2010) focused primarily on beliefs and practices of three writing teachers in an elite university, who taught both English and non-English major students. The study found that all the teachers believed that writing was a thinking process and that both language use and the development of thinking skills were key goals of writing instruction. But, different from Zhang's (2008) findings about the inconsistent relationship between beliefs and practices, Yang's study found that the three teachers all emphasized a balance of writing products and processes in their beliefs and their practices.

First, the conflicting findings might have to do with the different learning and teaching experiences of the participant teachers in the two studies. The three teachers in Yang's study had many years of English teaching experience (i.e., 15, 23, and 43 years, respectively) and writing instruction (i.e., 5, 12, and 15 years, respectively). Two of them had completed postgraduate studies in the US and worked as teaching assistants for writing courses in the US universities. Their overseas learning and working experiences had a great impact on their perceptions of the value of process writing and prepared them well for the implementation of the approach. In contrast, the five teachers in Zhang's study had less teaching experience (i.e., 2, 7, 8, 10, and 23 years, respectively). Even though no details about their experience of teaching writing were provided, the researcher indicated that those teachers were inexperienced in teaching writing and unfamiliar with the nature of the process-based pedagogical approach. It has been suggested in the literature that more experienced teachers are likely to have more experientially-informed beliefs than less experienced teachers, and that deeply held principles or beliefs informed by teaching experiences might be applied more consistently to teaching practices than principles acquired from teacher education (as it is expected in the case of new teachers) (Basturkmen, 2012; Breen et al., 2001).

Second, the conflicting findings of the two studies might have to do with the significant institutional differences existing between the elite university in Yang's study and the middle-ranking university in Zhang's study. The former university's students, less influenced by the prospect of tests since the passing rate of the tests (i.e., College English Test and Test for English Majors, CET and TEM in short) had remained high for a considerable time, welcomed the development of their writing skills more than enhancing their test taking skills. In addition, the class size at the elite university was around 24. In contrast, students from the latter middle-ranking university were more CET oriented and studied English in a bigger class. Therefore, teachers in the elite university had relatively favorable conditions for adopting pedagogical activities such as multiple drafts, multiple revisions, and peer feedback.

It is recommended in the literature that feedback should be provided on multiple drafts (not only final graded drafts) and from multiple sources (not only teachers) (Ferris, 2014; Lee, 2017). Peer review, as a prominent feature of process-oriented writing instruction, has many potential benefits (Huisman et al., 2018; Hyland & Hyland, 2006; Tsui & Ng, 2000; Zhang & Mceneaney, 2019). In spite of certain advantages of peer review, it is "not readily embraced by teachers in L2 school contexts" (Lee, 2017: 98). In the context of mainland China, two reasons evident in Zhang's (2008) study may account for the limited use of peer feedback: one, teachers' unfamiliarity with the nature and value of peer review; two, teachers' perception of contextual constraints they have to deal with when it comes to the implementation of peer review. While the bulk of peer review studies investigated students' perceptions and attitudes (Chang, 2016), studies that look at writing teachers' perceptions and attitudes are still scarce.

In response to these issues, there are calls for more research that can take account of teachers' practitioner knowledge about formative feedback in specific teaching contexts. The present study, therefore, investigated two expert writing teachers who implemented their ideal feedback practices (e.g., multiple drafts, peer review, self-assessment) despite constraints that seem get in the way in other teachers' attempts to do so in the EFL context of mainland China. The study incorporated the teachers' own voice to address the "how" and "why" questions: How do the teachers give feedback? More importantly, why do they give feedback in the ways they do?

12.2 Methods

12.2.1 Participants and Teaching Context

The study is part of a large research project on feedback practices and beliefs of Chinese university EFL writing teachers. The large research project adopted a mixed-methods design: a qualitative multiple-case study of 10 teachers and a quantitative questionnaire survey ($N = 202$). The case study in the first phase investigated teachers at a distinguished university in Northeast China. Purposeful and snowball sampling was adopted. Teaching experience was the major consideration in recruiting participants. Among the 10 teachers, two teachers taught English writing for over 10 years and three teachers less than 2 years. Teachers who taught English majors and those taught non-English majors were both recruited. For the purposes of identifying the potential impact of personal learning/training and research experiences, special effort was also made to recruit teachers who had overseas learning/training experiences and those who did not, and teachers who had research interests in EFL writing instruction and those who did not.

The study reported in this paper focused on two female teachers Anna and Bella (pseudonyms) who taught English major students in the School of English Studies. Anna, the former writing course coordinator in the school, had 11 years of experience teaching English writing, and was also one of the staff that had led a writing pedagogy reform in the school since 2006. As a result of the reform lasting many years, the product approach to writing that had been dominant in the school was replaced by the process-genre approach (see Badger & White, 2000). At the time of the study, Anna was doing a research project on formative assessment in EFL writing instruction and had already published extensively in that area. Bella was the current course coordinator. She had taught English writing for 8 years. She obtained a PhD degree in Applied Linguistics, and her research area was teacher feedback. Anna and Bella were both considered by their colleagues not only as experienced teachers but also expert teachers in writing instruction. They were called 'backbone' writing teachers in the school. Their expertise was manifested in their mentoring of novice writing teachers, their knowledge of writing pedagogical approaches, their engagement in the writing pedagogy reform, and their publications in the field of writing instruction.

Generally speaking, their writing instruction was devoted to teaching English major sophomores how to write argumentative essays and notes – two types of writing tasks tested in TEM-4. There were 24 students in Anna's class and 35 students in Bella's. The average class size was 30 students in the school. Assessment was based on students' performance in weekly writing tasks (75% in total) and an end-of-semester project (25%). It was not mandated how teachers should go about responding to student writing. None of the course documents specified guidelines teachers should follow in marking student written texts, except that each semester teachers should select a minimum of three compositions by each student to comment on and grade. The scores given for the three compositions would be added to account for a great proportion (75%) of the final assessment of student writing performance. Apart from the minimal feedback workload required, there was no mention of feedback criteria in any course document. Peer feedback and multiple drafts were not compulsory. In other words, teachers had complete freedom and flexibility in terms of giving feedback to student writing.

12.2.2 Instruments

Multiple instruments were used to collect data from the participants. First, teacher interviews elicited data about participants' self-reported practices, rationales for the practices, and relevant personal experiences. The interviews conducted in the case study were semi-structured, and most questions were open-ended in nature. The interviews were conducted in Chinese. I translated all the interview data from Chinese to English. Second, think-aloud provided data about teachers' decision-making and thought processes while they were giving feedback. Third, stimulated recall sessions focused on how teachers explained and justified their specific feedback strategies. Last but not the least, marked student texts with teacher feedback elicited data about teachers' actual feedback practices. Collectively, these instruments were intended to create maximal opportunities for the teachers to speak for themselves. They were also aimed to achieve data triangulation, providing corroborating evidence from different sources to shed light on the research questions (Barnard & Burns, 2012; Miles et al., 2014).

In addition, student texts with peer feedback were also collected and two student interviews conducted. These additional data were collected because, in the midst of data collection, it was found that the two teachers, unlike other teachers in the case study from the same school, frequently used peer review. This discovery led quickly to a modification to the research design. It is worth noting that all the written peer feedback was not generated for the study's purposes but was naturalistic data.

12.2.3 Data Collection and Analysis

Table 12.1 provides details of data collection procedures. I met the teachers three times spaced out over a 17-week semester (i.e., meetings with Anna in weeks 2, 8 and 13 and Bella in weeks 3, 8 and 15). Considering the teachers' busy work schedules and preferences for online communication, they were further contacted mainly via the university's Office Application system or the networking app WeChat. Whenever it was necessary for them to add on, confirm, or clarify the interview data, they would be contacted.

Table 12.1 Information on data collection

	Instrument	Anna	Bella
Teacher meeting 1	Interview	46 mins	53 mins
Teacher meeting 2	Think-aloud	14 mins	11 mins
	Stimulated recall	25 mins	24 mins
Teacher meeting 3	Interview	40 mins	37 mins
Student meeting	Interview	22 mins	13 mins
Teacher feedback		14 texts	13 texts
Peer feedback		42 texts	13 texts

As for the collection of student texts, students were invited to provide their assignments that had already been marked up by their teachers for any task. The specific procedures were as follows: Immediately after a teacher agreed to participate, I went on to contact one student in the teacher's class (either the class monitor or the subject representative) and sought his/her help to promote the study among the students. The student was then asked to help collect his/her classmates' texts that had been written by Week 6 and Week 14. This method of seeking students' help for text collection was suggested by a participant teacher. She suggested that it would be more feasible to ask the student representative rather than the busy teachers to collect student texts and ask students to sign a consent form if they agreed to participate.

Coding of the data was conducted using NVivo 10. Altogether, the interview data were coded in three cycles. The first cycle was to establish a list of *open codes* (Saldana, 2000). The coding unit was set as a single sentence, but extended to a whole paragraph for the majority of the texts. Each unit in the text was assigned one or multiple code names, using either words in vivo (i.e., words or short phrases taken from the participants' own language), a descriptive label, or a concept in the literature (Saldana, 2000). The second cycle was to generate *pattern codes*. The main purpose was to chunk and sort data into categories. Three overarching categories were (a) self-reported feedback practices, (b) rationales (knowledge, belief, view), and (c) relevant teaching experience. The sub-categories under each overarching category were not pre-designed but mostly emerged from the interview data. The third cycle was also to generate *pattern codes*. However, different from the second cycle, it aimed to establish pattern codes that could reflect "relationships among people" (Saldana, 2000: 88). Specifically, cross-case comparison was made to identify similar and different self-reported feedback behaviors and beliefs among the 10 teachers in the large study. I composed a summative narrative (one or two pages long) for each teacher in order to achieve a comprehensive understanding of their distinctive feedback practices. I also drew up a cognitive map for each teacher to visually display their networks of beliefs and knowledge.

The teachers' actual written feedback on the student written texts was coded using NVivo 10, too. The coding started with the identification of feedback points. *Counting feedback points* is the most widely adopted method in the textual analysis of teacher written feedback (e.g., Lee, 2011; Montgomery & Baker, 2007). A feedback point refers to any mark, correction, or comment made by teachers that constitutes a meaningful unit. Each feedback point was then categorized in terms of feedback focus, error correction strategy, and feedback type, with reference to existing schemes in the literature (e.g., Hyland & Hyland, 2006; Lee, 2008).

To enhance the trustworthiness of the data analysis, the results of the preliminary analyses were sent back to the participants for clarification and confirmation that the results matched their interpretations.

12.3 Findings

The study identified multiple factors influencing the teachers' feedback choices. The first noticeable factor was the two teachers' belief in the value of peer review of multiple drafts on a regular basis. Peer review was frequently and extensively used by the two teachers and was a priority focus of their feedback. Another factor was their belief in the alignment between writing assessment rubrics and class instructional focal points. Teacher feedback and teacher assessment was not only intended to reflect but also to inform instruction. Last but not least, they believed that successful peer feedback relied on systematic sustained support and supervision on the teachers' part.

12.3.1 Frequent Use of Multiple Drafting and Peer Review Liberates Teacher Feedback from Primary Focus on Linguistic Errors

The cross-case comparison in the large case study found that Anna and Bella contrasted sharply with other teachers in the same school and the teachers in other schools of the university. Those teachers did not use peer feedback at all or used peer feedback only once or twice in the semester. Anna and Bella, however, used peer feedback frequently and extensively throughout the semester. Anna organized peer feedback on a regular bi-weekly basis. Within these 2 weeks, her students were encouraged to produce as many drafts as possible based on feedback from peers. Peer feedback was conducted within groups of three or four who were usually roommates. Each student read and commented for the other two or three members of the group. Among the 14 student texts with teacher feedback collected from Anna's students, six texts were third drafts, two texts fourth drafts, and two texts fifth drafts.

Anna confessed that, even though she told her students that every draft mattered in the final grade of one writing task, she could not afford time to examine the corrections and revisions in detail in every draft. But she did take the number of drafts and the "first-final-draft-difference" into consideration. That is, the more efforts a student put into revision, the higher grade would be awarded.

> Individuals make their best efforts; peers also do their best. The fifth draft, the tenth one, I don't set the limits. I ask them to submit all the drafts and bind them in order, with the first draft put at the bottom and the final one on the top, all labeled in number. (Anna, first interview)

The student interviews confirmed Anna's reported practice of using peer feedback. They also gave evidence for Anna's encouragement for additional revisions.

> Sometimes when the final drafts are handed back to us, I will go on to revise. Topic sentences, concluding sentences, and coherence issues in between, we are asked to give another check at all these important points. (Anna's student, interview)

The peer feedback of Anna's students featured a focus on content and organization. For example, a classification essay on Different Types of Shoppers went through five drafts in total. Here are the comments from three peer reviewers. Originally, the peer comments were in English mixed with Chinese. I translated them into English.

Reviewer 1: The first paragraph does not exhibit the topic explicitly. For a classification essay, you had better come up with a theme running through the whole essay.

Reviewer 2: The classification does not follow one consistent standard. My suggestion: 1st type of shoppers buy just for need; 2nd type buy whatever they like; 3rd type do not buy anything.

Reviewer 3: Classification is good; three types are just alright, no more no less. The second type is inadequately discussed. Please elaborate. Sentence structures are somewhat simple.

It is very obvious that the peer editors had pointed out many issues in relation to content and organization: lack of clarity and controlling ideas in the topic sentence, inconsistent classification standards, and inadequacy of one supporting detail, etc.

Similarly, Bella told her students that she would check all the drafts and wanted to see "the original and the raw stuff – the true process." Slightly different from Anna's requirement of group peer review (i.e., three students reviewing one text, each student reviewing three times), Bella required her students to review in pairs. Another difference was that Bella required them to write down self summaries of their revision work. She checked on those self summaries and included them as the priority of her feedback focus. More accurately, her feedback started off with reading these self summaries and peer editors' comments in reference to student texts. Her think-aloud data provided evidence supporting what she said. Bella started off her think-aloud like this, "This is news report. . . . This is the second draft. I read the peer comment first."

The two teachers perceived peer review as effective but did not take it as a panacea. They acknowledged that they were aware that there would still be errors undetected in spite of peer correction, there would even be wrong corrections at times, and there would be good peer comments that were not appreciated. Despite their awareness of the issues surrounding peer feedback, both teachers strongly believed that peer feedback would be effective and helpful as long as adequate training, guidance and supervision were provided, another interesting finding to which Sect. 12.3.3 will attend in detail.

12.3.2 Teacher Feedback Not Only Reflects But Also Informs Instruction

The instructional objectives of Intermediate English Writing were to introduce the writing of a three-paragraph essay (opening paragraph, body paragraph and concluding paragraph) and to provide students with vocabulary, structures and techniques they would need in order to write four types of essays (exemplification, classification, cause and effect, and comparison and contrast). Moreover, topic sentences, concluding sentences, and transitional devices for the four types of essays were among the important instructional points focused on throughout the semester.

The two expert teachers gave their primary attention in feedback to issues that corresponded to their focal instructional points in class. The writing task that was assigned afterwards must relate to the points, so as to check whether students had grasped the teaching content and were able to apply what they learned in class to the actual writing.

The feedback foci were not absolutely fixed but varied in accordance with the lesson foci. The teachers shared that instructional focal points could be a particular genre feature or a type of student problem teachers wanted to address urgently. For example, apart from the regular lesson foci such as topic sentences, concluding sentences, and transitional devices, Anna and Bella both reported that they would set aside about two teaching weeks for intensive sessions targeting at sentence-level errors only, a focused issue they perceived necessary to address urgently. As they provided short-term intensive training to tackle students' grammar errors, accordingly they gave special attention to language errors in the following one or two tasks after the training sessions.

The alignment between instructional foci and feedback foci was mainly achieved via the use of checklists. The two teachers reported that checklists were frequently emphasized and used in their class and also referred to in their feedback. For example, Bella used checklists to let the students know in advance what they should attend to in a news story. If students did well for the point(s), she would give a pass; but if students do not fulfill the requirement concerning the focal point(s), it would still be a definite fail, despite other appropriate aspects of the writing. Teacher assessment was based on their selective feedback on focal issues.

The alignment between class instruction and teacher feedback is a two-way process. The two teachers did not see feedback on paper as an isolated act but as part of the teaching cycle guided by the *preceding* instruction and preparing for the *subsequent* instruction. The excerpt below provides an example.

> This student didn't know how to use the transitional phrase 'on one hand, on the other hand'. This is not a problem in her essay only. She might think this phrase can be used for listing out two things. She might not know that the phrase should be used for two different aspects of the same thing. I will explain it again in class. (Anna, second interview)

Figure 12.1 illustrates the teachers' perception of the relationship between teacher feedback and class instruction. The two arrows on the top indicate the teachers' perception of feedback as being "guided" by, and a natural extension of, class

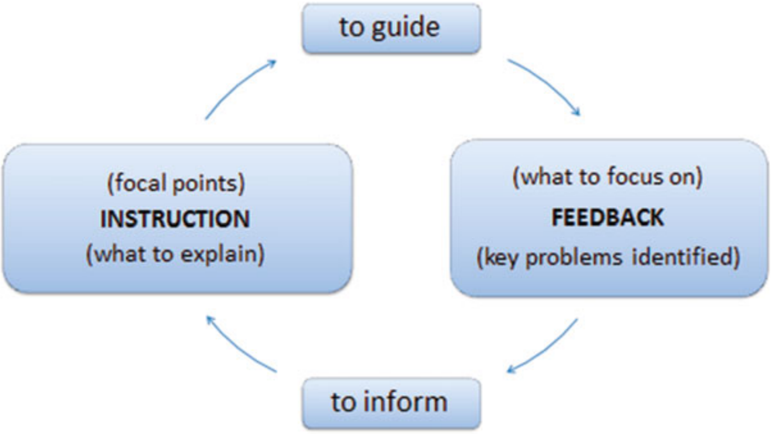


Fig. 12.1 Relationship between class instruction and teacher feedback

instruction. The two arrows at the bottom indicate that while the important points these teachers emphasized in class had found their ways into their feedback process, the problems they identified during the feedback "informed" them of what issues they should address in class.

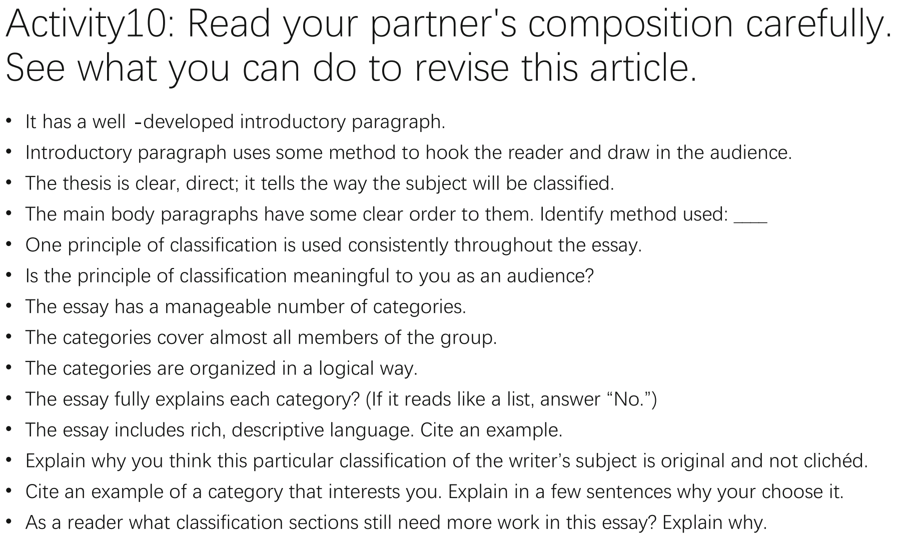
12.3.3 Systematic Sustained Support and Supervision Is Important

12.3.3.1 Systematic Peer Feedback Is Aided But Not Restricted by Checklists

Anna reported that she trained her students to do peer feedback by referring to checklists. The checklists, in her words, served as "a blue print." She emphasized that checklists can help reduce the subjectivity in feedback and the students need them as guidance. Otherwise, they would have no idea how to respond to peers' work appropriately, as illustrated in the following excerpt:

> I observe that what is characteristic of my students in peer review is they rush to look for grammatical and mechanical mistakes, circle them, and consider the job done. I remind my students that marking out mistakes and errors should be the very last thing to do in peer review. (Anna, first interview)

Anna suggested to her students that peer editors should mainly choose and cover global issues. She asked them to refer to the rubrics in the checklists, which were intended to give them reliable and appropriate guidance on what issues to look at. Figure 12.2 is one checklist Anna used for peer review of classification essays.



Activity10: Read your partner's composition carefully. See what you can do to revise this article.

- It has a well -developed introductory paragraph.
- Introductory paragraph uses some method to hook the reader and draw in the audience.
- The thesis is clear, direct; it tells the way the subject will be classified.
- The main body paragraphs have some clear order to them. Identify method used: ____
- One principle of classification is used consistently throughout the essay.
- Is the principle of classification meaningful to you as an audience?
- The essay has a manageable number of categories.
- The categories cover almost all members of the group.
- The categories are organized in a logical way.
- The essay fully explains each category? (If it reads like a list, answer "No.")
- The essay includes rich, descriptive language. Cite an example.
- Explain why you think this particular classification of the writer's subject is original and not clichéd.
- Cite an example of a category that interests you. Explain in a few sentences why your choose it.
- As a reader what classification sections still need more work in this essay? Explain why.

Fig. 12.2 A PPT slide for a classification essay checklist

The two teachers explained that peer review should be based on, but not restricted by, the rubrics specified in the checklists. Originally, Anna had asked her students to put ticks/crosses on the lists. For a couple of tasks she had requested the students to print out checklists and attach them to the submitted compositions. Anna explained that the writing teachers in the university she visited in the US mostly ticked or crossed items on checklists; they seldom gave corrections or comments on students' compositions. She had intended to keep to the same practice but gave it up because she observed that, more often than not, ticks/crosses on the list could not effectively reflect and monitor the students' efforts in peer feedback. She recalled an experience to illustrate her point:

> I once provided a checklist about paragraph writing to students. One student ticked high grades for many items for a paragraph written by his peer, but I read through the paragraph and found instantly it didn't deserve the high grades. Based on my experience, I ruled out the possibility that the student editor was unable to find out the problems with that paragraph.

That is why she decided to modify the use of checklists, asking her students to put down specific textual comments because she believed it would push them to be more committed. They should not only give in-text corrections but also put down their comments at the bottom of the composition, either in English or Chinese. The comments should not cover every issue but focus on major issues regarding content, viewpoints and structures. The summative comments should fall into three categories: strengths, weaknesses, and suggestions. She explained that in this way written feedback became more focused, constructive and flexible; comments could cater to students' individual problems.

12.3.3.2 Sustained Peer Review Support Is Achieved Through Regular Dialogues and Discussions in Class

Anna preferred to give support for peer feedback in class. Anna found that it was necessary for her to model for the students the review process and to follow up by reviewing one or two tasks together with them. More importantly, it was necessary to sustain discussions about peer performance in class whenever necessary. Even though checklists served as the guidance, class discussion activities must follow up so as to check whether students applied the rubrics to the actual peer feedback accurately and objectively.

Anna requested her students to give oral presentations of peer feedback in class. There were three specific procedures. First, peer feedback was done within groups. Second, each group decided on one composition for presentation in the second session. Anna gave three selection criteria. They could choose (1) the one they agreed to be the best written, (2) the one with the biggest improvement when compared to the first draft, or (3) the one that all members in the group had no idea how to help to improve, that is, with problems that they felt incapable of dealing with. Third, in the second session each group presented the selected essay together with peer comments. Afterwards, Anna and the student audience would give further assistance and guidance, working together with the groups to comment on the selected essays.

Anna observed that when it came to topic sentences, peer editors were likely to fail to provide effective evaluation. It happened often that her students thought a composition had an effective topic sentence, whereas actually she later found out that the topic sentence was either just a statement of a fact (but not the writer's opinion) or a statement that contains a topic (without controlling ideas that the following sentences can support or prove). The same was true of the supporting details. Under these circumstances, she would follow up to point out in class what was wrong with the peer comments. The student interview confirmed Anna's self-reported practice.

> What to focus on in peer review? At the beginning we didn't know, only to mark out small grammar mistakes. Later, Miss Anna taught us to focus on global issues like organization and selection of supporting details. She gave a lot of emphasis on these two aspects. One time, girls in my dormitory chose one essay we all thought were well-written, the one we could not perfect any more. It turned out, however, that when we put it on the ppt slide in class, Miss Anna detected a problem with supporting details, which we didn't notice previously. There are situations like this: we all believed an essay was good enough, but when Miss Anna pointed out the issues, we were kind of taken aback, "indeed, there were problems." We are not having as sharp eyes (as the teacher). (Anna's student, interview)

Apart from the group presentation of peer comments in the second session, Anna believed the conferencing in the subsequent session was also necessary in that it could allow her to check the students' uptake of those suggestions derived from the class discussion. Here is an example.

> One time, a student, already in the middle of the semester, didn't recognize run-on sentences in his writing. . . . It just happened that his essay was selected by his peer editing group as the representative sample for presentation in the second session. I of course pointed out that he had written run-on sentences. However, he didn't correct them in his revision. When the final draft was collected for me to grade, I was surprised that the problem was still there. Then I pointed out this problem in the fourth session to the whole class again. That student kind of protested, "Miss Anna, I have been writing sentences like that since long ago. How come they are wrong?" I then realized that he was not aware what a run-on sentence was. I wrote his problem sentence on the blackboard and the whole class analyzed and discussed it again. The other students confirmed to him that it was indeed a mistake. (Anna, first interview)

On reflection, Anna realized that "a mistake, if repeated, becomes a false truth. This is a good lesson for both that student and the rest of the class." She also realized that because of their low language proficiency, follow-up class discussions were much needed to assist and evaluate the peer feedback performance and monitor the uptake of peer/teacher comments.

12.3.3.3 Supervision of Peer Feedback Is Read and Monitored Regularly on Written Texts

While Anna worked together with her students on peer feedback in class, Bella preferred to give further written comments on peer commentary and required her students to write reflective self-editing summaries. She reported that after one or two sessions teaching her students about how to do peer review, more importantly, she needed to keep "pushing" or "monitoring" the students to do peer feedback by various means.

Bella would read peer feedback and make comments next to it, like "He (peer editor) gave good comments," "very to the point," and "good suggestions." She also required her students to submit their second drafts with reflective summaries of revisions and to indicate in the summaries where they took up peer comments.

> The student editors, I know, usually would check if the teacher responded to their comments, and if the teacher approved of their comments. I feel only when I attended to the peer performance this time, could they carefully do peer review next time, because they knew not only the student writer would read (their comments), but the teacher too. . . . Peer feedback requires student commitment. Only when they feel like doing, willingly and carefully, could it be a practice that improves their skills. (Bella, first interview)

Additionally, Bella used grades to incentivize peer and self editing.

> I told students that the peer review and self reflection were part of evaluation. I said that merely with the intention to push them. If student A edited for B, I asked A to put down his/her name. Just wanted to push them but didn't actually grade peer comments. If I had really counted it as part of the final assessment, it would have been too complicated and troublesome. (Bella, second interview)

12.4 Discussion

The study identified three pedagogical factors behind the teachers' feedback choices: multiple drafts, peer feedback and feedback foci alignment with the learning goals of instruction. In what follows, the feedback practices of the two teachers in the present study are discussed and compared with practices recommended by feedback literature.

Previous peer feedback studies of Chinese students mostly investigated its effectiveness in comparison with teacher feedback in an experimental design and students' perceptions of its effectiveness (e.g., Hu & Lam, 2010; Hu & Ren, 2012). Previous studies have also reported the difficulty of using peer review (e.g., Hu, 2005; Hu & Lam, 2010; Yu et al., 2016) and multiple drafts (e.g., Wang, 2010) among Chinese students. There was a wide difference in teachers' beliefs about peer review. These varied beliefs centered on three questions: whether students that grow up with Chinese learning cultures are capable of doing peer review, what contributions it can make to student writing, and whether it can be implemented in their specific teaching contexts. Anna and Bella in the current study strongly believed that students were capable and that peer review had many benefits. The two teachers put emphasis on the process of peer review, seeking the pathway to high-quality peer feedback that could lead to better revised texts. In other words, they shifted the focus from *whether* peer review was effective to *how* peer review could be effective in their own classrooms. The present study has contributed to the research base of peer feedback by looking at how peer feedback was perceived by the teachers who actually used it in their natural teaching contexts.

The effective implementation of peer feedback by the two teachers in the present study can be explained as having three aspects: systematic training, sustained support, and sustained supervision. Firstly, systematic training was reflected in that the teachers used task-specific checklists to train their students before each peer feedback activity. The primary advantage of using checklists as the training tools was that the checklists informed their students of what focal issues they should selectively and primarily target in peer review, because checklists could explicitly guide the students on the *content* of peer review (Zhao, 2014). Students' understanding of evaluation points through checklists before embarking on a peer feedback activity could help them know how to give appropriate and substantive feedback (Baker, 2016). Another advantage was that the checklists assist peer reviewers with appropriate *language* they could use, because linguistic strategy was an essential part of an effective peer review training session (Hu, 2005; Hu & Ren, 2012; Sanchez-naranjo, 2019). Secondly, sustained support was reflected in that the two teachers maintained regular communication with their students to hear concerns and difficulties they encountered in the process of peer feedback. The communication for sustained support in the present study was either in the form of in-class oral group presentations or in the form of written self-reflective reports. As soon as Anna found out reviewer-reviewer or reviewer-writer conflicts in group presentations, she would provide timely interventions and solutions. Lee (2017) advocates that teachers “let students share their experience and concerns”, and “provide opportunities for students to incorporate self-feedback/assessment”. These were two teacher-supported strategies that the teachers in the present study had well adopted. Thirdly, sustained supervision is reflected in that the two expert teachers’ belief that it was not realistic to expect their students to be committed to peer review, unless sustained follow-up teacher feedback on peer review was provided, be it oral praise for excellent peer performance or written comments asking for additional revisions. Through class discussions, Anna was able to evaluate the peer feedback performance and monitor the uptake of peer/teacher comments. It is recommended in the literature that in order to actualize the optimal benefits of peer review, a very important consideration is that students get feedback on how successful they have been in giving feedback – they need evaluative feedback on their actions (Hu, 2005; Zhao, 2014). The advantage of their sustained supervision of peer feedback was that by including accountability and evaluation mechanisms, students could take the activity seriously (Ferris, 2014).

Though it was not within the scope of the present study to investigate whether revisions undertaken as a result of the peer review had enhanced the quality of writing, both teachers and their students had acknowledged its several benefits. The benefits were also evident from the student texts. Peer comments on their students’ texts were specific and constructive. Overall, final drafts were of better quality. For one classification essay-writing task, there were almost no content and organization issues in the fifth (and also final) draft for Anna to comment on, since peer reviewers and the writer had effectively addressed those global issues in the previous drafts. The findings of the present study are in line with previous studies that teachers’ supportive intervention strategies involving discussion and interaction with their peers had a positive impact on students’ attitude on peer review and in turn their writing performance (Hu, 2005; Sanchez-naranjo, 2019).

Another remarkable finding was the two expert teachers’ experimentation, observation, modification and reflection on what worked best in their specific contexts of work. For example, Anna modified rating scales in checklists to open-ended questions and required her students to give formative textual comments in peer review. This finding lends support to Hu’s (2005) conclusion that, in order to actualize the optimal benefits of peer review, teachers should not just understand effective training for successful peer review from published research (i.e., to think globally) but also reflect on their own less successful activities and work out effective ones in their specific teaching context (i.e., to take adequate local action). Through monitoring systematically the success of new activities/actions, the two teachers developed their practical experiential knowledge about a set of strategies that worked effectively in their own teaching context. The finding also confirms that one of the distinctive qualities of expert teachers is their capability of theorizing practical knowledge (Tsui, 2009).

Finally, the present study found that the alignment between class instruction and teacher feedback helped teachers to integrate feedback into part of the teaching class and helped students understand the rationales for teacher and peer feedback. Some writing teachers are worried that if they do not give comprehensive feedback to students, their students will consider them irresponsible teachers (Lee, 2011). Students may hold unrealistic beliefs about a teacher's responsibility and other aspects of teacher feedback, usually based on their previous experiences, experiences that may not necessarily be beneficial for the development of writing. There is much teachers can do to alter student expectations of and views of teacher feedback. One way is to engage students in the discussion of feedback criteria for different writing tasks and explain them clearly to the students. Anna and Bella provided specific checklists for their students to refer to when they wrote essays and when they did peer reviews. What the two expert teachers did is that they explained to their students explicitly what their feedback criteria were. Otherwise, their students may not have been able to interpret their feedback or act on it as they had intended.

12.5 Conclusion

The findings of this study on the two expert teachers have pedagogical relevance for front-line writing teachers. Against the assumption that peer review in groups on multiple drafts is not feasible (see Yu et al., 2016, for the cultural issues and other constraints), the study found that students were actually very capable of doing peer review. Teachers should be prepared to understand that peer review is not easy in the beginning. Their students may not feel like doing peer review and may start off simply correcting a few errors, or even make wrong corrections and inappropriate comments. These two expert teachers also shared that these problems were normal when they started to trial peer review. They, however, came to learn from their own experiences that successful peer feedback relied on equipping students with peer review strategies and providing them with systematic sustained supervision and support.

Unlike experimental studies of feedback on limited types of errors conducted in controlled environments that "lack ecological validity" (Storch, 2010: 43), this study reflected real classroom conditions where the teachers provided feedback on valid and authentic writing tasks over a 17-week semester. Acknowledgment should be made here about the practical constraints on the implementation of the recommended practices that include class sizes, exam pressures, shortage of time, etc. The two expert teachers in the case study university also faced these constraints. Thus, by sharing their practices, this study hopes to offer something of interest and use to writing teachers whose teaching contexts resemble those of this study.

Acknowledgements I thank my supervisor Professor Hu Guangwei. I also thank Nanyang Technological University for awarding me a graduate research scholarship.

References

Badger, R., & White, G. (2000). A process genre approach to teaching writing. *ELT Journal, 54*, 153–160.

Baker, K. M. (2016). Peer review as a strategy for improving students' writing process. *Active Learning in Higher Education, 17*, 179–192.

Barnard, R., & Burns, A. (Eds.). (2012). *Researching language teacher cognition and practice: International case studies* (p. 2012). Multilingual Matters.

Basturkmen, H. (2012). Review of research into the correspondence between language teachers' stated beliefs and practices. *System, 40*, 282–295.

Borg, S. (2006). *Teacher cognition and language education: Research and practice*. Continuum.

Breen, M. P., Hird, B., Milton, M., Oliver, R., & Thwaite, A. (2001). Making sense of language teaching: Teachers' principles and classroom practices. *Applied Linguistics, 22*, 470–501.

Chang, C. Y.-H. (2016). Two decades of research in L2 peer review. *Journal of Writing Research, 8*, 81–117.

Du, H. (2012). College English teaching in China: Responses to the new teaching goal. *TESOL in Context, 3*, 1–13. Retrieved from http://www.tesol.org.au/files/files/278_hui_du.pdf

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Lawrence Erlbaum.

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing, 19*, 6–23.

Ferris, D. R., Brown, J., Liu, H. S., & Stine, M. E. A. (2011). Responding to L2 students in college writing classes: Teacher perspectives. *TESOL Quarterly, 45*, 207–234.

Freeman, D. (2002). The hidden side of the work: Teacher knowledge and learning. *Language Teaching, 35*, 1–13.

Freeman, D., & Richards, J. C. (1996). *Teacher learning in language teaching*. Cambridge University Press.

Goldstein, L. M. (2001). For Kyla: What does the research say about responding to ESL writers. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 73–90). Lawrence Erlbaum.

Goldstein, L. M. (2005). *Teacher written commentary in second language writing classrooms*. University of Michigan Press.

Hu, G. W. (2002). Potential cultural resistance to pedagogical imports: The case of communicative language teaching in China. *Language, Culture and Curriculum, 15*, 93–105.

Hu, G. W. (2005). Using peer review with Chinese ESL student writers. *Language Teaching Research, 9*, 321–342.

Hu, G. W., & Lam, S. (2010). Issues of cultural appropriateness and pedagogical efficacy: Exploring peer review in a second language writing class. *Instructional Science, 38*, 371–394.

Hu, G. W., & Ren, H. W. (2012). The impact of experience and beliefs on Chinese EFL student writers feedback preferences. In R. Tang (Ed.), *Academic writing in a second or foreign language: Issues and challenges facing ESL/EFL academic writers in higher education contexts* (pp. 67–87). Continuum.

Huisman, B., Saab, N., Den Broek, P. V., & Van Driel, J. H. (2018). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*, 863–880.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*, 83–101.

Jin, L., & Cortazzi, M. (2006). Changing practices in Chinese cultures of learning. *Language, Culture and Curriculum, 19*, 5–20.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*, 69–85.

Lee, I. (2011). Working smarter, not working harder: Revisiting teacher feedback in the L2 writing classroom. *The Canadian Modern Language Review, 67*, 377–399.

Lee, I. (2017). *Classroom assessment and feedback in L2 school contexts*. Springer.

Mei, T., & Yuan, Q. (2010). A case study of peer feedback in a Chinese EFL writing classroom. *Chinese Journal of Applied Linguistics, 33*(4), 87–98.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: An expanded sourcebook* (3rd ed.). Sage.

Montgomery, J. L., & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing, 16*, 82–99.

Pan, L., & Block, D. (2011). English as a "global language" in China: An investigation into learners' and teachers' language beliefs. *System, 39*, 391–402.

Richards, J. C. (2010). Competence and performance in language teaching. *RELC Journal, 41*, 101–122.

Saldana, J. (2000). *The coding manual for qualitative researchers*. SAGE.

Sanchez-naranjo, J. (2019). Peer review and training: Pathways to quality and value in second language writing. *Foreign Language Annals, 52*(3), 612–643.

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies, 10*(2), 29–46.

Tsui, A. (2009). Distinctive qualities of expert teachers. *Teachers and Teaching: Theory and Practice, 15*, 421–439.

Tsui, A., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*, 147–170.

Wang, P. (2010). Dealing with English majors' written errors in Chinese universities. *Journal of Language Teaching and Research, 1*(3), 194–205.

Wei, Y., & Chen, Y. (2003). Supporting Chinese learners of English to implement self assessment in L2 writing. *Proceedings of the Independent Learning Conference 2003*. Retrieved from http://www.independentlearning.org/uploads/100836/ila03_wei_and_chen.pdf

Yang, L. X. (2010). Gaoxiao yingyu zhuanye jiaoshi xuezuo jiaoxue xinnian yu jiaoxue shijian: Jingyan jiaoshi ge'an yanjiu [University English writing teachers' beliefs and practices: A case study of experienced teachers]. *Waiyu Jiaoxue Lilun yu Shijian* [Foreign Language Learning Theory and Practice]*, 2*, 59–68.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*, 179–200.

You, X. (2004a). "The choice made from no choice": English writing instruction in a Chinese University. *Journal of Second Language Writing, 13*, 97–110.

You, X. (2004b). New directions in EFL writing: A report from China. *Journal of Second Language Writing, 13*, 253–256.

Yu, S., Lee, I., & Mak, P. (2016). Revisiting Chinese cultural issues in peer feedback in EFL writing: Insights from a multiple case study. *The Asia-Pacific Education Researcher, 25*, 295–304.

Zhang, D. (2005). Teaching writing in English as a foreign language: Mainland China. In M. S. K. Shum & D. L. Zhang (Eds.), *Teaching writing in Chinese speaking areas* (pp. 29–45). Springer.

Zhang, P. (2008). Yingyu jiaoshi xiezuo jiaoxue xinnian yu jiaoxue shijian de shizheng yanjiu [An empirical study on English teachers' beliefs about and practices in writing instruction]. *Yuwen Xuekan* [Journal of Language and Literacy Studies]*, 11*, 157–160.

Zhang, X., & Mceneaney, J. E. (2019). What is the influence of peer feedback and author response on Chinese university students' English writing performance? *Reading Research Quarterly, 55*(1), 123–146.

Zhao, H. (2014). Investigating teacher-supported peer assessment for EFL writing. *ELT Journal, 68*, 155–168.

Dr. Jing Yang is currently a lecturer at Dalian University of Foreign Languages. She has been teaching English Academic Writing (Paragraphs to Essays) and Integrated College English for over 10 years. Her main research interests are in the areas of English writing instruction and teacher feedback.

Chapter 13
Diagnostic Assessment of Written Accuracy: New Directions for Written Corrective Feedback in Secondary Writing Classrooms

Icy Lee, Na Luo, and Pauline Mak

Abstract There has been a proliferation of research on focused written corrective feedback in recent years. The bulk of this research, however, has adopted experimental/quasi-experimental designs, targeted at a limited number of error categories, hence neglecting teachers' authentic needs in the classroom. The present study seeks to investigate how diagnostic writing assessment can be used to enhance focused WCF practice in authentic L2 writing classrooms. This chapter, part of a larger study, is intended to shed methodological light on how a diagnostic writing assessment conducted in class by teachers can be applied to inform systematic focused WCF practice in authentic L2 writing classrooms. The findings show that there is a great variation of error density across the student texts, with learners of higher proficiency level having lower error density and writing longer texts. Regarding error frequency and gravity, students tend to make frequent errors in verb tense, word choice and punctuation, and gravity of errors falls on sentence structure, expression and word choice. We propose a general set of pedagogical procedures for teachers interested in carrying out diagnostic assessment of students' writing, and call for the need of diagnostic writing assessment to ensure a systematic focused WCF practice.

Keywords Corrective feedback · Written accuracy · Diagnostic assessment

I. Lee (✉)
The Chinese University of Hong Kong, Hong Kong SAR, China
e-mail: icylee@cuhk.edu.hk

N. Luo
Wuhan University of Science and Technology, Hong Kong SAR, China

P. Mak
The Education University of Hong Kong, Hong Kong SAR, China
e-mail: pwwmak@eduhk.hk

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_13

13.1 Introduction

Writing is often considered a very difficult language skill for L2 students (Wang et al., 2016). Apart from the challenge arising from composing, developing, and organizing ideas in writing, one key problem many L2 students face is how to translate ideas into accurate language (Harris & Silva, 1993; Hinkel, 2002). To help students improve writing, teachers need to provide feedback on their texts in addition to classroom instruction. This is because feedback is pivotal to student learning, perhaps comparable to direct instruction (Hattie & Timperley, 2007). As Hyland (2013: 180) aptly puts it:

> Feedback offers the writer an outsider's view of a text and so provides a sense of audience and what that audience values in writing, contributing to his or her acquisition of disciplinary subject matter and patterns of argument and evidence.

Quality feedback to student writing should be timely, individualized, focused, and attend to different dimensions of writing in a balanced way (Ferris, 2014; Lee, 2019). However, in a large number of L2 contexts, particularly EFL settings, writing is primarily viewed as a vehicle for language reinforcement (Hyland & Anan, 2006; Lee, 2008). Such a predominant focus on language form is reflected in a conventional feedback approach where teachers respond to written errors comprehensively, which is referred to as "comprehensive written corrective feedback" (comprehensive WCF). As a pedagogical practice, comprehensive WCF is fraught with problems, such as posing cognitive overload to students (Bitchener, 2008), confusing and demotivating them (Hyland & Hyland, 2006).

To combat the preponderant influence of comprehensive WCF, opponents advocate a focused approach – that is, giving feedback to target error categories in student writing. Such an approach is referred to as "focused written corrective feedback" (focused WCF). While research on focused WCF has proliferated in recent years, it has mostly consisted of experimental/quasi-experimental studies, targeting a very small number of error categories (e.g., articles) that are chosen by researchers for research purposes rather than by teachers in response to authentic needs of the classroom. There is a need for more classroom-based focused WCF research that is guided by the systematic selection of target error categories based on a clear understanding of students' pervasive error patterns and written accuracy performance.

Against this backdrop, we embarked on a research project on focused WCF in Hong Kong secondary writing classrooms, aiming to explore whether and how this approach can benefit students in terms of language accuracy and other dimensions of writing. This chapter, part of the larger study, is intended to shed methodological light on how diagnostic assessment, in the form of a writing test, can be applied to inform systematic focused WCF practice in authentic L2 writing classrooms. By focusing on the secondary writing context, it also addresses a context gap in feedback and writing assessment research that has primarily been conducted in college contexts.

13.2 Literature Review

In this section, we first review relevant literature on focused WCF (alongside comprehensive WCF), which provides the impetus for our study. We then situate error selection pertaining to focused WCF within classroom writing assessment, with a specific focus on diagnostic assessment, emphasizing the importance of classroom assessment administered by teachers in real classroom contexts. Finally, we examine the key parameters of diagnostic assessment of written accuracy, which informs the error analysis of the study.

13.2.1 Focused WCF

It has been widely acknowledged in the literature that comprehensive WCF can be counterproductive for both teachers and students (Ferris, 2011; Lee, 2019; Truscott, 1996). For teachers, comprehensive WCF is time-consuming and distracts them from providing feedback on other important issues, such as content, organization and genre (Lee, 2019). It can negatively influence teachers' identity development, relegating them to "error hunters" (Hairston, 1986: 122) and "marking machines" (Lee, 2010: 148). Additionally, the large amount of time required for comprehensive WCF easily burns them out, causing them to respond to student writing hastily, putting them at the risk of producing illegible WCF or even inaccurate error corrections (Lee, 2008). For students, comprehensive WCF leads to formidable problems as well because the marked papers are often flooded with red ink (Lee, 2008), which not only saps students' writing motivation (Hyland & Hyland, 2006; Lee et al., 2018) but also causes "information overload" (Bitchener, 2008: 109). Overall, there is a lack of conclusive research evidence about the effectiveness of comprehensive WCF (e.g., Kepner, 1991; Truscott, 1996).

Although some recent studies found that comprehensive WCF could improve written accuracy (e.g., Bonilla López et al., 2017; Hartshorn & Evans, 2015; van Beuningen et al., 2008, 2012), most of them did not compare comprehensive WCF with focused WCF and were conducted in experimental/quasi-experimental conditions with limited relevance for authentic classrooms. While Rahimi (2021), by comparing the efficacy of the two approaches, shows that comprehensive WCF can be more successful than focused WCF in improving students' overall written accuracy, he acknowledges that it may be more appropriate for writing classrooms that are geared more toward learning language than writing.

A viable alternative to comprehensive WCF is focused WCF – that is, teachers responding to students' writing selectively (Ferris, 2011; Lee, 2019; Lee et al., 2015). Focused WCF is likely to be more helpful than comprehensive WCF as it finds much stronger support from SLA. To process WCF, students go through several cognitive stages (Gass, 1997): (1) consciously attending to the WCF; (2) noticing the difference between their output and WCF; (3) understanding the

WCF, analyzing it with reference to stored knowledge; (4) hypothesizing and testing new output; and (5) producing final output. With focused WCF, since a smaller number of error types is targeted, learners are more likely to attend to WCF consciously and more prone to noticing and understanding the feedback (Ellis et al., 2008).

Therefore, it has been argued that responding to recurrent patterns of errors in a focused manner, especially rule-governed items (e.g., verb tense and form, articles, subject–verb agreement) is more beneficial than responding to all errors comprehensively (Ferris, 2011; Lee, 2019). To test the aforementioned hypothesis, researchers have investigated feedback given on one or two linguistic domains (Bitchener, 2008; Bitchener & Knoch, 2008, 2009, 2010; Sheen, 2007; Shintani et al., 2014), yielding evidence for the effectiveness of focused WCF in improving the grammatical accuracy in the short and long run. Most studies which investigated the relative effectiveness of focused WCF and unfocused WCF (two of them defined unfocused WCF as feedback on a range of error categories rather than all error categories) also suggested that the former may be more effective than the latter in enhancing the accuracy of the target language features (Ellis et al., 2008; Rahimi, 2021; Sheen et al., 2009).

Overall, research has suggested that focused WCF is more manageable than comprehensive WCF for both teachers and students. For teachers, because they spend less time marking errors, they free up energy for other meaningful aspects of writing (e.g., responding to content and organization, and preparing materials to teach writing). For students, when their papers are no longer inundated with red ink, they are more likely to develop confidence and motivation in writing, which may in turn affect their feedback uptake (Lee et al., 2018; Lightbown & Spada, 2006; Mahfoodh, 2017; Storch & Wigglesworth, 2010).

However, most existing studies on focused WCF lack ecological validity as they have predominantly adopted the (quasi-)experimental design in controlled and laboratory-like conditions (Storch, 2010). For instance, researchers focused on a very narrow set of grammatical features, such as the referential indefinite and definite articles (e.g., Bitchener & Knoch, 2008, 2010; Sheen, 2007), the past tense *–ed* (Frear & Chiu, 2015), and the indefinite article together with the hypothetical conditional (Shintani et al., 2014). It remains questionable whether teachers grappling with various constraints in real classroom contexts can afford to focus on such a narrow set of language features (Xu, 2009). While Ferris et al. (2013) targeted a wider range of language domains based on students' needs, the heavy involvement of the researchers in the feedback process called into question the pedagogical application of its findings. Although Rahimi (2021) compared focused WCF and comprehensive WCF in natural classroom settings, he did not attend to the full complexity of the teaching context. To show the effectiveness of focused WCF in ecologically and pedagogically valid contexts, naturalistic studies which attend to the contextual dynamics are urgently called for.

13.2.2 Diagnostic Assessment for Focused WCF

In authentic classroom contexts, teachers who intend to implement focused WCF are immediately faced with the question of which errors should be selected as target error types. We postulate that perhaps it can be best answered by the administration of a diagnostic writing test at the outset of the writing class. Ferris (2011) questioned the feasibility of diagnostic assessment for focused WCF as it could be time-consuming for teachers. Alternatively, she suggested teachers rely on *ad hoc* observation in choosing which errors to mark during WCF. Different from Ferris, we believe pre-selected error types based on diagnostic assessment should be combined with teachers' *ad hoc* observation if focused WCF is to be implemented systematically to maximize student learning. Through diagnostic assessment, students' strengths and weaknesses in written accuracy are identified, which can inform not only the selection of error types for focused WCF but also pre-writing grammar instruction. In this way, a close connection can be fostered between assessment and teaching, which is definitely conducive to student learning (Carless et al., 2011).

Unfortunately, research on diagnostic assessment for writing is few and far between. Of the scant research on diagnostic assessment for writing, much of it has adopted analytic rubrics to assess ESL writing performance comprehensively (e.g., Erling & Richardson, 2010; Knoch, 2009, 2011; Kim, 2011; Llosa et al., 2011) without a specific focus on students' language accuracy. The only study focusing on diagnostic assessment on language accuracy of ESL writing is Xie (2019), who developed and validated a list of error types to gauge the language performance of university students in Hong Kong. While this error list can help teachers select target errors for focused WCF systematically, its application for the secondary context is questionable. A case in point is that the error list contains a total of 33 error categories, which is cumbersome and unwieldy for both secondary teachers and learners.

13.2.3 Major Parameters in Diagnostic Assessment of Written Accuracy

To diagnose written accuracy, a writing test relevant to the students' current level can be used to elicit errors students make (James, 2013). While existing literature on diagnostic assessment for writing hardly pinpoints how diagnostic writing tests differ from other kinds of writing tests (e.g., proficiency or placement tests) (Alderson, 2005), researchers suggest that the difference between the two does not lie in the writing prompt but the way the test paper is marked (Knoch, 2009; McNamara et al., 2002). Thus, to gauge learners' written accuracy, teachers can readily use writing prompts designed for proficiency or placement test purposes but mark errors in the test papers comprehensively.

In diagnostic assessment of written accuracy, at least three parameters have to be identified: error density, error frequency and error gravity. Error density refers to the number of different errors per unit of text (e.g., per 100 words) and helps teachers understand students' overall language accuracy (see James, 2013). A high error density rate can seriously impede writers from transmitting their intended meaning (ibid.), rendering the text less intelligible even if the errors are minor like spelling (Gunterman, 1978; Zola, 1984). If students display a high error density rate in their writing, teachers need to spend more time on language accuracy issues. A precondition for determining error density is to decide text length, that is, the total number of words in a given text. In addition to serving the role of identifying error density, text length itself may help teachers predict students' writing proficiency. As noted by Lee, Mak and Burns (2015), more proficient writers tend to write longer texts.

Error frequency refers to the total incidence of the same error category in a text, whereas error gravity is defined as the relative seriousness of different error categories (James, 2013; Xie, 2019). For teachers implementing focused WCF, the basic principle for target error selection is to focus on the most frequent errors and/or the most serious ones (Ferris et al., 2013; James, 2013; Xie, 2019). In other words, both error frequency and error gravity are pertinent to error selection for focused WCF. While deciding error frequency seems quite straightforward as it involves counting the total number of errors in each error category, determining error gravity can be thorny. Although errors that impede communication are usually considered more serious and thus bear greater gravity than others (e.g., Burt & Kiparsky, 1976; Vann et al., 1984), it seems impractical to establish a hierarchy of error categories in terms of gravity because its judgment is influenced by multiple variables such as students' L1 background and education (McCretton & Rider, 1993; Rifkin & Roberts, 1995). To date, the only consensus is limited to the following:

- Global errors violating rules of the overall sentence structure are more serious than local errors which only affect a single sentence constituent (i.e., a word or a group of words that functions as a single unit, such as subject and predicate, within the sentence) (Burt & Kiparsky, 1976; Tomiyana, 1980; Xie, 2019);
- Errors of semantic deviance (e.g., lexical errors) are often more serious than grammatical errors (Engber, 1995; Khalil, 1985; Rahimi, 2021; Santos, 1988; Xie, 2019).

In this study, the diagnostic assessment of written accuracy is informed by the above parameters – i.e., error density, error frequency, and error gravity.

13.3 Participants and Contexts

The participants in the present study were four Secondary 3 (henceforth S3) classes from two secondary schools in Hong Kong, two from a band 1 school (School A) and the other two from a band 2 school (School B). Secondary schools in Hong Kong are divided into three bands based on students' academic abilities. Band

Table 13.1 Basic information of the 4 participating classes

	School A		School B	
School band	1		2	
Participating classes	Class A1	Class A2	Class B1	Class B2
Students taking the test	30	32	31	28
English proficiency	Mixed ability	Mixed ability	Most proficient	Less proficient

1 schools have students of the highest academic abilities while band 3 schools take in students of the lowest academic abilities. The four classes were taught by four different teachers who volunteered to participate in our research project on focused WCF.

At School A, the S3 students were randomly assigned to their classes, and hence students in the two participating classes (henceforth Class A1 and A2) had similar overall English proficiency. In contrast, at School B, the students were placed into their classes based on English test results in the previous academic year, with the most proficient in Class B1 and the students in the lower half in Class B2 and another class. In total, 121 S3 students took part in the study. The basic information of the four participating classes is summarized in Table 13.1 above.

13.4 The Procedures

13.4.1 Developing a Categorization of Error Types

To begin with, a set of error categories was developed for the diagnostic assessment of written accuracy. The first author conducted a pilot study, collecting 90 student texts from S3 students of another school to generate error codes that suit Hong Kong secondary school students. The errors were coded by the first author and validated with the help of the second and third authors, generating a set of 16 error categories that guided the data analysis of the present study (see Appendix A). During the error coding process, we took into consideration the realities of real Hong Kong classrooms. In cases where more than one way is possible to code an error, we opted for a code that we believe both teachers and students can easily understand and apply. Our error coding scheme is based on a simple principle: if any error can be corrected by fixing a punctuation mark or a word, even though the error may be alternatively considered as a sentence structure error, we code it as a punctuation or word choice error, as shown in the examples below (each error underlined and marked with an asterisk):

There *had many dirty things on the beach. *(error of word choice)*

Smoking is harmful to health*, it will make you tired easily. *(error of punctuation)*

Careful distinction is made between sentence structure errors and multiple errors in the same sentence, as shown in the following examples:

Luckily, they *have a man came. (a sentence structure error)
Afther about 10 minutes. The police arrived and *tell all the *person on the beach to leave. (multiple errors in one sentence)

In addition, we add the category of "expression" errors to designate those which cannot be fixed by changing one word but fall into a single sentence constituent, as in the following examples:

We not only enjoyed *the sea water, but saw a real shark.
. . .when we went to travel *the air-plane, I saw. . .
*The industrial killed many sharks. . .

An expression error differs from a word choice error in that the latter involves only one word, as in the examples below:

. . .the lifeguard complimented Peter *that his bravery.
*One they arrived at the beach, Jacky's parents went swimming. . .

13.4.2 The Diagnostic Writing Test

A picture writing task (see Appendix B) was used for the diagnostic test. The prompt was adapted from an S3 English test paper of the Hong Kong Territory-wide System Assessment, a proficiency test which the Hong Kong Education Bureau annually administers to S3 students. Based on the proposition that diagnostic writing test differs from writing tests of proficiency and placement tests not in the prompt but in the way the test paper is marked (Knoch, 2009), we decided to adopt the test paper of an authoritatively validated test rather than design one by ourselves.

At the beginning of the larger study, the writing test was conducted with the intention to provide baseline data about students' written accuracy performance and error patterns. In the writing task, there were four pictures, with the first two about a family travelling by plane and arriving at an airport, the third and fourth about the family having fun on a beach and spotting a big sea animal. The students were asked to write a story based on the pictures within 40 min. The participating teachers administered the test to their respective classes in an English lesson at the beginning of the 2018–2019 academic year under examination conditions in which the students were not allowed to use dictionaries and seek support from others.

13.4.3 Analyzing the Test Papers

After the 121 student texts from the picture writing test were collected, we took a series of steps to analyze them, as described below. First, the second author counted the number of words in each paper, yielding information on text length. Next, the three authors got together to code the errors of one randomly chosen paper from each class to make sure that they had similar interpretations of the error codes. As we worked through the error analysis, we also came up with clearer definitions of error density and error gravity for our study. While we were aware that error density is traditionally defined as the number of different errors per unit of text (James, 2013), we found it too time-consuming to decide which errors were the same, given a corpus of 121 student papers adding up to 34,149 words. To make the task manageable, we only excluded repetitive spelling errors. For example, one student misspelt the word "Thailand" as "Tailand" six times, and the error was only counted once. Thus, we operationally defined "error density" as the total number of errors per 100 words, excluding repetitive spelling errors. We also defined grave errors as those meeting the following two criteria simultaneously:

- The error either affects more than one sentence constituent (global) or involves semantic deviance (often lexical);
- The error is moderate or high in frequency.

After that, the second and third authors chose three other papers from each class randomly (10%) and marked them independently. The inter-reliability rates between them were 95.8% for error identification and 91.3% for error correction. The second author then coded errors in the rest of the papers and counted the number of words in each paper to calculate error density. Finally, she entered the error analysis results into SPSS (Version 25) for statistical analysis. Descriptive statistics applied including calculation of the total, percentage and standard deviation. Independent t-tests were run to compare whether there was significant difference between the two classes within the same school.

13.5 Results and Discussion

Below, we report the findings of the study in three subsections: error density, error frequency and error gravity.

13.5.1 Error Density

The students' writing varied considerably in error density, ranging from having only 1.1 errors to as many as 40 errors per hundred words. As shown in Table 13.2, the mean error density of School A, the band 1 school, was 10.283 while the mean error density of School B is 22.112. Although there was some difference in the mean error density of the two classes at School A, independent *t*-test showed that this difference remained statistically insignificant ($p > 0.05$). This result aligned well with the fact that both classes at School A included students of mixed abilities and were largely similar to each other in terms of overall English proficiency. In contrast, at School B, Class B2, the weaker class, had a much higher mean error density (mean = 28.448)

Table 13.2 Error density of each class and school

		N	Maximum	Minimum	Mean	Std. deviation
School A	Class A1	30	17.9	3.8	9.940	3.4371
	Class A2	32	22.8	1.1	10.604	4.1653
	School	62	22.8	1.1	10.283	3.8138
School B	Class B1	31	30.9	6.1	16.390	4.9908
	Class B2	28	40	16.3	28.448	6.4556
	School	59	40	6.1	22.112	8.3161
Overall		121	40	1.1	16.051	8.7221

than Class B1 (mean = 16.390), the most English proficient class at the school. Independent *t*-test showed that this difference was statistically significant ($p < 0.05$).

The great variation of error density across the student texts means that they differed greatly in texture. Below, we show three excerpts from the students' texts with high, moderate and low error density.

> They ran and followed the shark. After running for 10 minutes, they reached the other end of the beach and the shark disappeared from view. "It's gone!" Julie said. "I don't want my adventure to end so fast." Tom was disappointed. Then, he realized that there was something like a door knob on the rocks beside him. (low error density excerpt, by a student of School A)

> Then, we went to *beach. My *parent *sit on the beach and *look after *me and my sister. My sister and I ran *to the beach. We went *to swimming and *play * the sand. After *Then, we were very *tried so we decided to eat *fishball and ice-cream. When we * eating the food*. Suddenly, we saw something in * sea. Therefore, we went *saw. (moderate error density excerpt, by a student of School B)

> Jack and *he family *today *go to *the Japan. They *go *to by *the plane. They *are so *exciting. When they * in *the Japan*. They *buy some special *thing such as *the special *sweet and *the *toy. They *are so happy. A few *day *, they *go to the beach. The beach * so beautiful. There *have *whit small sand and cold clean water. (high error density excerpt, by a student of School B)

The students with lower error density rate tended to be those who wrote longer texts. Pearson correlation tests showed that there was a negative relationship between error density and text length with a correlation coefficient of 0.684. Although the writing task and the time allowed were identical, the number of words students wrote in different classes varied considerably, ranging from as many as 659 words to as few as 44. We show the text length of the four classes in Table 13.3 below.

13.5.2 Frequency of Error Types

13.5.3 Overall Frequency

The frequency of different error types of the two schools is shown in Table 13.4 below.

Table 13.3 Text length of each class and school

		N	Maximum	Minimum	Mean	Std. deviation
School A	Class A1	30	604	190	335.20	94.671
	Class A2	32	659	193	359.41	107.387
	School	62	659	190	347.69	101.342
School B	Class B1	31	394	159	263.61	63.789
	Class B2	28	237	44	157.86	53.878
	School	59	394	44	213.42	79.328
Overall		121	659	44	282.22	113.155

Table 13.4 Descriptive statistics of different error types among the students

	N	Minimum	Maximum	Mean	Std. deviation
Vt (verb tense)	121	0	25	6.36	4.917
Wc (word choice)	121	0	14	5.77	2.851
P (punctuation)	121	0	24	4.31	3.786
Prep (preposition)	121	0	9	3.06	2.079
Sp (spelling)	121	0	16	2.97	2.735
Art (article)	121	0	15	2.92	2.290
Vf (verb form)	121	0	14	2.79	2.595
SS (sentence structure)	121	0	13	2.64	2.187
Exp (expression)	121	0	10	2.37	2.009
Pron (pronoun)	121	0	10	1.39	1.508
Ne (noun ending)	121	0	8	1.28	1.507
Wf (word form)	121	0	4	.77	.873
C (connectives)	121	0	8	.75	1.254
Wo (word order)	121	0	3	.61	.789
Ag (agreement)	121	0	4	.54	.885
M (miscellaneous)	121	0	3	.09	.387

Based on Table 13.4, in terms of frequency the 16 error types can be roughly divided into three categories (in descending order):

1. Highly frequent errors: verb tense, word choice and punctuation
2. Moderately frequent errors: preposition, spelling, article, verb form, sentence structure, expression, pronoun and noun ending
3. Infrequent errors: word form, connectives, word order, agreement and miscellaneous

The error frequency patterns of the two schools are shown in Fig. 13.1 below.

According to Fig. 13.1, the students of School A and School B followed similar patterns in their frequency of different error types despite occasional differences. At both schools, the three most frequent errors were tense, word choice and punctuation while they differed in the frequency of the first two. Similarly, the moderate and least frequent errors remained the same regardless of occasional differences in the ranking of some error types. Since reducing error density is a key objective for improving language accuracy, we elaborate on the three most frequent error types below.

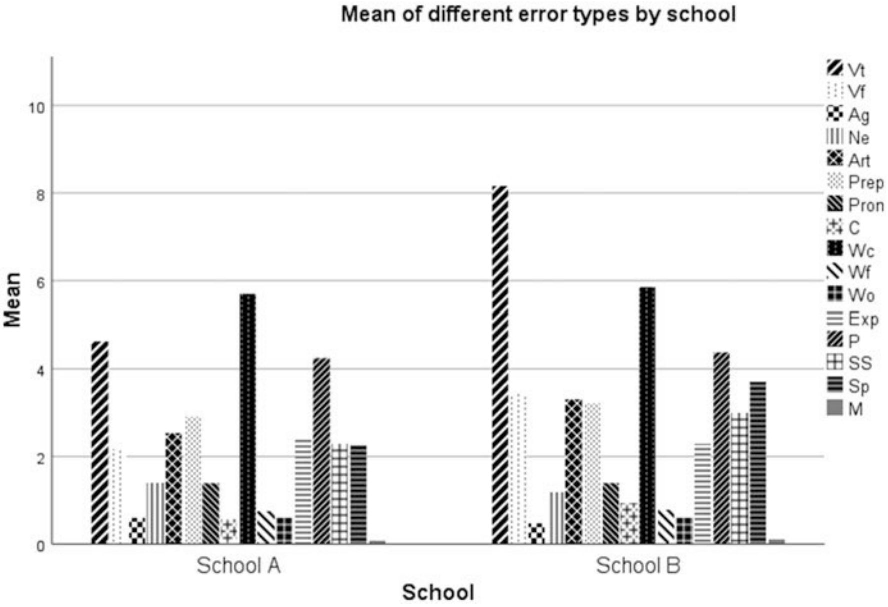


Fig. 13.1 Frequency of different error types at the two schools

13.5.3.1 The Three Most Frequent Error Types

In general, the students erred most frequently in verb tense, word choice and punctuation. While the two schools were similar in the mean value of word choice and punctuation errors, the students at School B made considerably more verb tense errors (8.17/student) than their counterparts at School A (4.63/students) despite that they wrote much shorter texts (see Fig. 13.1 and Table 13.5).

As the writing task was story writing based on pictures, the students were supposed to use the past tense in most sentences. Their high mean value of tense errors shows that many students, particularly those at School B, had difficulty using past tense. As “tense” is rule-governed, it may be relatively easy for teachers to rectify this problem if selected as a target language feature. However, two caveats are in order. First, tense errors are highly genre-sensitive. If the task is to write a leaflet, for example, the past tense verbs are unlikely to feature. In that case, verb tense errors may be less frequent. Second, the students erred mainly in the past tense and were not necessarily confused with the entire tense system.

The second most frequent error type among the students was word choice (5.77/student, see Table 13.4), similarly distributed across both schools (see Table 13.5). The high frequency for this error type may result from three factors. Firstly, ESL/EFL writers often suffer from a limited repertoire of lexical items and nuanced understanding of them (Hinkel, 2002). Secondly, since the students in this study were just S3 students learning to write English as a second/foreign language, their lexical repertoire was far from fully developed. In addition, the way we defined word

Table 13.5 Descriptive statistics of the three most frequent error types

		N	Minimum	Maximum	Mean	Std. deviation
Vt	School A	62	0	15	4.63	3.716
	School B	59	0	25	8.17	5.382
Wc	School A	62	1	14	5.69	2.826
	School B	59	1	14	5.85	2.900
P	School A	62	0	24	4.24	4.218
	School B	59	0	14	4.37	3.306

choice errors may affect the frequency. For us, word errors are any problems (wrong, unnecessary or missing) of a single word not falling into otherwise specified types (e.g., prepositions, connectives). In this way, errors which would have been classified traditionally as other errors were counted as word choice errors, as shown in the examples below:

There *have (√ is) a shark! (not sentence structure error)
I *fill (√ feel) very cool. (not spelling error)

However, this additive effect may be somewhat reduced by the fact that lexical errors affecting two or more words were subsumed under the category of expression errors.

The third most frequent error type was punctuation, with a mean of 4.31/student (see Table 13.4). Although punctuation errors are usually less severe as they seldom disrupt communication, their high frequency and highly rule-governed nature may render focused work on them particularly effective in reducing error density.

13.5.4 *Error Warranting Attention Due to Gravity*

Another important objective in helping L2 students improve written accuracy is to focus on the most serious errors that impair communication more than others. Based on the criteria in Sect. 13.4.3 as well as results of the study, we find that three grave error types which teachers should pay special attention to: sentence structure, expression and word choice (see Table 13.4 for frequency). We only elaborate the first two below as errors of word choice have already been discussed above as a high frequency error type.

13.5.4.1 Sentence Structure Errors

Sentence structure errors are inherently global in this study as we defined them as deviations affecting two or more sentence constituents, unfixable by changing a single word, punctuation or sentence constituent. They could seriously affect texture, as shown in the following excerpt from one student's text:

> Last year, I *going to pictures with my father, mother and sister. First, we *by the airport to go to French. We were so happy *but we were first by the airport.

Sentence structure errors are only partially rule-governed. While there are rules governing sentence structure (e.g., SV, SVO), L2 students can produce errors of sentence structures beyond imagination (James, 2013). Given their severity and moderate frequency, they merit attention from teachers and students participating in this study.

13.5.4.2 Expression Errors

Another serious error type is errors of expression. As they affect more than one word in a single sentence constituent (see examples in Sect. 13.4.1), expression errors are in nature non-rule-governed lexical problems which cause semantic deviance. We envisage that expression errors are not rare in student writing and suggest teachers consider targeting this error type for focused WCF and/or in post-writing grammar instruction.

13.6 Implications and Conclusion

In this chapter, we report the error patterns of S3 students in two secondary schools in Hong Kong, identified through a pre-study diagnostic writing test of a research project on focused WCF. First, the results of the study confirm a commonsensical view that more proficient students have lower error density and write longer texts and vice versa. Secondly, based on the criteria of error frequency and gravity, we conclude that the teachers in this study needed to give priority to errors of verb tense, word choice, punctuation, sentence structure and expression out of the 16 error categories. However, since error patterns are influenced by multiple factors such as L1 background, L2 proficiency, education level and language aptitude (Chan, 2010), we are fully aware that the error patterns identified in this study may not be transferable to other contexts. Teachers who want to practice focused WCF in other contexts should administer their own diagnostic writing task, probably following procedures similar to those described in this study. Despite the limitations, our results provide some interesting insights into WCF, which we elaborate below.

13.6.1 The Necessity for Diagnostic Assessment for Systematic Focused WCF

As mentioned earlier, if focused WCF is to be practiced systematically, diagnostic writing assessment is necessary rather than optional. Focused WCF has been criticized for being unsystematic due to the idiosyncratically selected target errors in the previous (quasi-)experimental studies (Hartshorn & Evans, 2012). We believe that diagnostic assessment of written accuracy at the beginning of a writing class will preempt such criticisms against the approach.

Because of the time-consuming nature of diagnostic assessment, Ferris (2011) recommends teachers rely on *ad hoc* observation in choosing which errors to mark during WCF. For us, this recommendation is not suitable for secondary teachers in Hong Kong for two main reasons. Firstly, teachers' *ad hoc* observation may run counter to students' needs (Kurzer, 2018), while a diagnostic writing test will bring a much more objective picture. For example, we had never anticipated that errors of punctuation would need to be prioritized in WCF before we analyzed the data presented in this chapter. Unexpectedly, punctuation appeared to be a consistent problem for all the four participating S3 classes. Accordingly, the teachers in our larger study were alerted to treat errors of punctuation in both WCF and classroom instruction. Secondly, as mentioned in the literature review, relying on *ad hoc* observation for target error selection means that teachers will miss a chance to connect WCF with pre-writing instruction on language features, a practice which finds support from not only the feedback literature (e.g., Carless et al., 2011) but also from skill acquisition theory (DeKeyser, 2007) (see 6.5 for suggested pedagogical procedure). Based on skill acquisition theory, errors will decrease after a combination of explicit rule-based instruction, exposure to abundant examples and frequent application (DeKeyser, 2007; Lyster & Saito, 2013). Feedback on target language features will be more effective if teachers provide explicit instruction on the target language features before writing (Lee, 2004). Without diagnostic assessment, they may pick on language features idiosyncratically for pre-writing instruction, thus failing to foster a strong alignment between assessment and instruction. Admittedly, analyzing errors gathered from diagnostic assessment can be time-consuming. However, if it is only conducted once a year or a semester and if the writing is relatively short, the workload seems manageable, particularly if assistance is available. Teachers can adopt some time-saving strategies, such as marking and coding the errors themselves but asking students or teaching assistants (if available) to count the errors and put the results in a pre-designed error analysis sheet.

Despite the importance of diagnostic writing assessment, we do not mean that the results of error analysis should rigidly dictate the teachers' selection of target errors in giving WCF. One caveat is that error patterns are likely to be genre-specific. Nevertheless, a pre-course diagnostic writing test is able to provide teachers with useful baseline information about students' written accuracy performance, which can inform both instruction and WCF.

13.6.2 The Need to Find Effective Ways to Address "Untreatable" Errors

Ferris (2011) labels rule-governed errors as "treatable" and non-rule-governed ones as "untreatable", acknowledging that some linger in between. Our results support her division of errors in terms of whether they are rule-governed. For instance, in the errors to be given priority in WCF for the participating teachers in this study, there are highly rule-governed (treatable) ones of verb tense and punctuation, partially rule-governed ones of sentence structure and non-rule-governed (untreatable) ones of word choice and expression.

This situation points to the need for writing teachers to deal with both treatable and untreatable errors in focused WCF, though research has mainly focused on treatable errors. Since untreatable or less treatable errors (e.g., word choice, expression and sentence structure) disrupt communication more seriously than treatable errors (Ferris, 2010), they are often more difficult for students to self-correct (Ferris & Roberts, 2001). Accordingly, they may merit more attention in both WCF and post-writing grammar reinforcement. Recent research has found that dynamic WCF (Hartshorn & Evans, 2012; Kurzer, 2018), an approach in which teachers give several rounds of comprehensive WCF on students' short texts on a given topic (written within 10 min) until students edit them to error-free pieces, can help L2 students reduce untreatable errors through revision. As a useful alternative to grammar exercise and/or instruction for improving written accuracy, dynamic WCF may be more suitable for ESL students in university writing classes, as in Hartshorn and Evans (2012) and Kurzer (2018). For younger L2 students in the secondary context, research on how to deal with untreatable errors through focused WCF is urgently needed.

13.6.3 Some Thoughts About Error Categories

Different researchers have developed different sets of error categories, and the number of error categories can range from three (James, 2013) to over 30 (Chan, 2010; Xie, 2019). Even seasoned writing researchers grapple with what error categories to use when marking students' papers, notably Ferris and her co-researchers, who had to switch between different sets of error categories in performing error analyses (Ferris & Hedgcock, 2005; Ferris & Roberts, 2001; Ferris et al., 2013).

In this study, we developed a set of 16 error categories to suit our research purpose and the teaching context in Hong Kong. While we found the list manageable in analyzing errors for research, we were aware that some categories may be too broad for teaching purposes. Take the category of verb tense as an example. While there are eight tenses in English, students in this study primarily erred in the past tense. If teachers concluded from the findings that students had difficulty with the entire tense system, it would be an overgeneralization. To accurately describe and target the students' errors, teachers may need to utilize subcategories, such as past tense within verb tense errors in this study. In terms of research, however, the more error categories, the more cumbersome and time-consuming is the error analysis. Based on our research, we conclude that it may be hard to have a set of error categories that suits the purpose of research and teaching simultaneously. As such, a separate set of error categories may have to be developed for pedagogical purposes.

13.6.4 Pedagogical Procedure for Diagnostic Assessment of Written Accuracy

We would like to end the chapter by proposing a general set of pedagogical procedures for teachers interested in diagnostic assessment of students' written accuracy:

1. Administer a pre-course impromptu, timed writing test in class, choosing a text type that is of great relevance to student needs.
2. Mark and code all errors of students' pre-course writing tests (i.e., comprehensive WCF) based on or adapted from an existing source of error codes (e.g., Ferris, 2011) appropriate for the specific students/context.
3. Perform error ratio analysis based on the WCF provided on students' pre-course writing tests (see Ferris, 2011; Lee, 2017). Provide each student with the result of error analysis that lists total errors for each error category (i.e., diagnostic in nature, showing strengths and weaknesses in written accuracy), so that students can find out their own error patterns.
4. Teachers can compile the error analysis results of all students and work out the error patterns for the entire class to guide grammar instruction. For example, if articles, prepositions, and run-on sentences are found to be the most frequent error types, teachers can prioritize these items in their teaching.
5. Individual students can look at the error analysis result based on the pre-course writing tests, heed their frequent error types, and monitor their own written accuracy development accordingly.
6. To track students' progress in written accuracy, a post-course timed, impromptu writing test can be administered to students under the same conditions of the pre-course writing test, using the same text type and a writing prompt of a similar level of difficulty. The same error analysis procedure can be adopted, with results (when juxtaposed against those of the pre-course writing tests) showing students' improvement with regard to specific error categories, and areas for further improvement.
7. One caveat is that the text type involved may influence the nature of errors made by students.

In conclusion, administering diagnostic assessment of students' written accuracy performance is a useful means to find out their strengths and weaknesses in their control of language in writing. The results can inform not only teachers' error selection in focused WCF but also their pre-writing instruction. They can also provide useful diagnostic information for students to help them monitor and evaluate their own written accuracy development. Although the implications for students' own learning are beyond the scope of this paper, diagnostic assessment carries potential benefits for both teachers and students and thus has a promising role to play in classroom writing assessment.

Acknowledgements The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 14606317).

Appendices

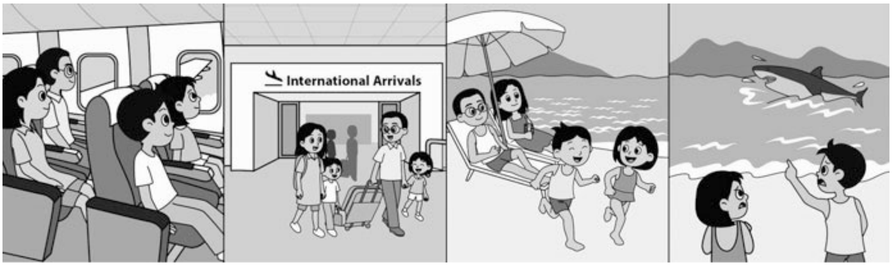
Appendix A: The Error Codes

Code	Error type	Brief definition
Vt	Verb tense	Errors in verb tense
Vf	Verb form	Errors in formation of the verb (phrase) not specific to time or tense marking
Ag	Agreement	Errors in either noun or verb form resulting in lack of agreement between subject and verb
Ne	Noun ending	Missing, unnecessary or incorrect plural or possessive marker / confusion about singular or plural noun ending
Art	Article	Wrong, unnecessary or missing article
Prep	Preposition	Wrong, unnecessary or missing preposition
Pron	Pronoun	Wrong, unclear, unnecessary or missing pronoun (excluding relative pronouns)
C	Connective	Wrong, unclear, unnecessary or missing conjunctions and other connectives
Wc	Word choice	Wrong word, word with unclear meaning in context, missing or unnecessary words
Wf	Word form	Wrong form of the word– i.e., the word is in the wrong lexical category for the context
Wo	Word order	Wrong word order
Exp	Expression	Errors involving multiple words but only affecting one sentence element
P	Punctuation	All punctuation and capitalization errors, correctable by merely changing the punctuation and/or capitalization
SS	Sentence structure	The use of wrong sentence pattern, missing or unnecessary expressions and other complicated disorders which affect more than two or more sentence constituents
Sp	Spelling	Wrong spelling
M	Miscellaneous	Other errors that do not fit into the above categories

Appendix B: The Writing Prompt

You are Jackie Ho, a student at SKFGLR Secondary School. Your class is writing adventure stories for the school magazine. Your teacher has given you the following pictures to help you to write a story.

You may use some of the ideas from the pictures and/or your own ideas in your writing. Write the adventure story in about 150 words in 40 min. Please provide a title for your story.



References

Alderson, C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*, 102–118.

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research, 12*, 409–431.

Bitchener, J., & Knoch, U. (2009). The value of a focused approach to written corrective feedback. *ELT Journal, 63*, 204–211.

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten-month investigation. *Applied Linguistics, 31*, 193–214.

Bonilla López, M., Van Steendam, E., & Buyse, K. (2017). Comprehensive corrective feedback on low and high proficiency writers: Examining attitudes and preferences. *International Journal of Applied Linguistics, 168*, 91–128.

Burt, M., & Kiparsky, C. (1976). Global and local mistakes. In J. H. Schumann & N. Stenson (Eds.), *New frontiers in second language learning* (pp. 71–80). Newbury House.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*, 395–407.

Chan, A. Y. W. (2010). Toward a taxonomy of written errors: Investigations into the written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly, 44*, 295–319.

DeKeyser, R. (2007). Skill acquisition theory. In B. van Patten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 97–113). Lawrence Erlbaum Associates.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*, 353–371.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139–155.

Erling, E., & Richardson, J. (2010). Measuring the academic skills of university students: Evaluation of a diagnostic procedure. *Assessing Writing, 15*, 177–193.

Ferris, D. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition, 32*, 181–201.

Ferris, D. (2011). *Treatment of errors in second language student writing* (2nd ed.). MI: University of Michigan Press.

Ferris, D. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing, 19*, 6–23.

Ferris, D., & Hedgcock, J. (2005). *Teaching ESL composition: Purpose, process, and practice*. Lawrence Erlbaum Associates.

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161–184.

Ferris, D., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing, 22*, 307–329.

Frear, D., & Chiu, Y. (2015). The effect of focused and unfocused indirect written corrective feedback on EFL learners' accuracy in new pieces of writing. *System, 53*, 24–34.

Gass, S. (1997). *Input, interaction, and the second language learner*. Lawrence Erlbaum Associates.

Gunterman, G. (1978). A study of the frequency and communicative effects of errors in Spanish. *Modern Language Journal, 62*, 249–253.

Hairston, M. (1986). On not being a composition slave. In C. W. Bridges (Ed.), *Training the new teacher of college composition* (pp. 117–124). NCTE.

Harris, M., & Tony, S. (1993). Tutoring ESL students: Issues and options. *College Composition and Communication, 44*, 525–537.

Hartshorn, J., & Evans, N. (2012). The differential effects of comprehensive corrective feedback on L2 writing accuracy. *Journal of Linguistics and Language Teaching, 3*, 16–46.

Hartshorn, J., & Evans, N. (2015). The effects of dynamic written corrective feedback: A 30-week study. *Journal of Response to Writing, 1*, 6–34.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hinkel, E. (2002). *Second language writers' text: Linguistics and rhetorical features*. Lawrence Erlbaum Associates.

Hyland, K. (2013). Student perceptions of hidden messages in teacher written feedback. *Studies in Educational Evaluation*, 39, 180–187.

Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System, 34*, 509–519.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*, 83–101.

James, C. (2013). *Errors in language learning and use: Exploring error analysis*. Routledge.

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *Modern Language Journal, 75*, 305–313.

Khalil, A. (1985). Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of Arab EFL learners. *TESOL Quarterly, 19*, 335–351.

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*, 509–541.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275–304.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81–96.

Kurzer, K. (2018). Dynamic written corrective feedback in developmental multilingual writing classes. *TESOL Quarterly, 52*, 5–33.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*, 285–312.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*, 69–85.

Lee, I. (2010). Writing teacher education and teacher learning: Testimonies of four EFL teachers. *Journal of Second Language Writing, 19*, 143–157.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching, 52*, 524–536.

Lee, I., Mak, P., & Burns, A. (2015). Bringing innovation to conventional feedback approaches in EFL secondary writing classrooms: A Hong Kong case study. *English Teaching: Practice and Critique, 14*, 140–163.

Lee, I., Yu, Y., & Liu, Y. (2018). Hong Kong secondary students' motivation in EFL writing: A survey study. *TESOL Quarterly, 52*, 176–187.

Lightbown, P., & Spada, N. (2006). *How languages are learned* (3rd ed.). Oxford University Press.

Llosa, L., Beck, S. W., & Zhao, C. G. (2011). An investigation of academic writing in secondary schools to inform the development of diagnostic classroom assessments. *Assessing Writing, 16*, 256–273.

Lyster, R., & Saito, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. P. García Mayo, M. J. Gutiérrez Mangado, & M. Martínez-Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). Benjamins.

Mahfoodh, O. H. A. (2017). "I feel disappointed": EFL university students' emotional responses to teacher written feedback. *Assessing Writing, 31*, 53–72.

McCretton, E., & Rider, N. (1993). Error gravity and error hierarchies. *International Review of Applied Linguistics in Language Teaching, 31*, 177–188.

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–242.

Rahimi, M. (2021). A comparative study of the impact of focused vs. comprehensive corrective feedback and revision on ESL learners' writing accuracy and quality. *Language Teaching Research, 25*, 687–710.

Rifkin, B., & Roberts, F. D. (1995). Error gravity: A critical review of research design. *Language Learning, 45*, 511–537.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*, 69–90.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of article*s*. *TESOL Quarterly, 41*, 255–283.

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*, 556–569.

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision or learners' accuracy in using two English grammatical structures. *Language Learning, 64*, 103–131.

Storch, N. (2010). Critical feedback on written feedback research. *International Journal of English Studies, 10*, 29–46.

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition, 32*, 303–334.

Tomiyana, M. (1980). Grammatical errors communication breakdown. *TESOL Quarterly, 14*, 71–79.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*, 327–369.

van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL International Journal of Applied Linguistics, 156*, 279–296.

van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in Dutch multilingual classrooms. *Language Learning, 62*, 1–41.

Vann, R. J., Meyer, D. E., & Frederick, O. L. (1984). A study of faculty opinion on ESL errors. *TESOL Quarterly, 18*, 427–440.

Wang, X., Shen, L., & Lu, X. (2016). Journal between peer learners: An innovative project to motivate EFL writers. *RELC Journal, 47*, 245–252.

Xie, Q. (2019). Error analysis and diagnosis of ESL linguistic accuracy: Construct specification and empirical validation. *Assessing Writing, 41*, 47–62.

Xu, C. (2009). Overgeneralization from a narrow focus: A response to Ellis et al. (2008) and Bitchener (2008). *Journal of Second Language Writing, 18*, 270–275.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics, 36*, 277–284.

Icy Lee is Professor at the Faculty of Education of the Chinese University of Hong Kong. Her publications have appeared in international journals such as the Journal of Second Language Writing, TESOL Quarterly, System, and Language Teaching Research. She is currently Principal Associate Editor of The Asia-Pacific Education Researcher.

Na Luo is Associate Professor of Applied Linguistics at Wuhan University of Science and Technology. Her research interest lies in second language writing, particularly academic writing. This chapter was finished when she worked as a postdoctoral fellow at the Faculty of Education of the Chinese University of Hong Kong.

Pauline Mak is Assistant Professor in the Department of English Language Education at the Education University of Hong Kong. Her research interests include language assessment, second language writing, and second language teacher education. Her publications have appeared in international journals such as System, Language Teaching Research and TESOL Quarterly.

Chapter 14
Assessment Training in the Use of Portfolios: Voices from Writing Teachers

Ricky Lam

Abstract Despite the benefits of writing portfolios, scholars remain unclear about how assessment training influences teacher use of portfolios for writing assessment in China. The chapter investigates the role and effectiveness of assessment training when Chinese teachers attempt portfolio assessment. The study was conducted in a doctorate degree programme in Hong Kong. Three informants from Mainland China registered an 11-session content course on English language assessment. The assessment training consisted of three lectures and two workshops on the principles of language assessment and writing portfolio assessment respectively. Data were collected by an open-ended questionnaire, post-workshop individual interviews and reflection papers, and analysed by qualitative methods. Implications are drawn to suggest future directions of developing teacher assessment literacy in China and beyond.

Keywords Portfolio assessment · Assessment training · L2 writing · Teacher assessment literacy in China

14.1 Background

Portfolios are broadly defined as dossiers to document a learner's efforts, professional growth, and achievements. In language education, portfolios are viewed as a learning-cum-assessment tool. Of various types of portfolios, writing portfolios have been widely used in L1 but not in L2 or EFL contexts. In the past few decades, there has been a body of research exploring the benefits of writing portfolios when applied as an instructional approach or an assessment tool (Burner, 2014). Yet, there is relatively little research to reveal what and how teachers learn to implement portfolio assessment (Lam, 2018). In studies of assessment literacy, scholars state that most teachers spend up to one-third of their professional time to evaluate students, but

R. Lam (✉)
Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China
e-mail: rickylam@hkbu.edu.hk

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_14

receive limited or no training in assessment which possibly bring about harmful effects on student learning (Stiggins, 2014). Thus far, not much is known about how systematic assessment training plays a role in enhancing teacher use of performance assessments, especially in the context of Chinese learners of English. Because of this, the focus of the paper is to look into how Chinese teachers attempt portfolio assessment to promulgate teaching and learning of writing alongside standardised testing. Its purpose is to test out whether a training approach could enhance Chinese teachers' assessment competence in EFL writing. More specifically, the paper aims to identify the role and effectiveness of assessment training when EFL teachers attempt writing portfolio assessment. The paper starts with a literature review section, followed by a methodology section. Results and discussion sections are then presented. The paper ends with an implication section on how to facilitate the development of teacher assessment literacy.

14.2 Literature Review

The literature review has three parts, comprising (1) portfolio assessment in L2 writing, (2) the role and effectiveness of assessment training, and (3) the overall (writing) assessment landscape in China.

14.2.1 Portfolio Assessment in L2 Writing

Utilising portfolios in writing classrooms corresponds with the process writing movement, where teaching writing emphasises multi-drafting, self- and peer-editing, and self-reflection. Studies on writing portfolios reveal that students become self-regulated in learning writing, and have considerable learning gains in accuracy and idea development (Mak & Wong, 2018). Portfolios can be said to reduce writing anxiety and to provide students with ample opportunities to revise works-in-progress (Lee, 2017). Portfolios for teaching are also likely to foster active agency and metacognitive capabilities when students collate their works reflectively (Curtis, 2018). Despite this positive evidence, there are studies reporting logistical issues, which would discourage both teachers and scholars from trying out portfolios, including standardised content coverage, the conflict between direct and indirect tests, and a lack of assessment training. For content coverage, portfolios originally promote variety, learner choice, and reflectivity, but some portfolio programmes require students to include prescribed portfolio entries to stifle creativity and learner autonomy (Scott, 2005). Regarding the conflict between direct and indirect tests, Hamp-Lyons (2002) stated that teachers might find it taxing to use indirect tests (portfolio assessment) to evaluate writing, given that direct tests (large-scale essay testing) warrant test fairness and scoring consistency. As to assessment training, Jiang and Hill (2018) discover that teacher learning of classroom-based assessment (e.g., portfolios) remains inadequate, particularly among teachers in the Asia-Pacific region.

Although portfolios have become popular, Hamp-Lyons (2007) stated that their use as assessment to evaluate writing is still problematic, because portfolios involve giving feedback, using feedback to inform teaching, and monitoring student learning formatively. Thus far, these aspects of portfolio assessment are seldom taught in teacher education programmes. On this note, Weigle (2007) suggested incorporating assessment into writing/ELT method courses to instruct writing teachers about assessment. She further described how portfolio assessment could be effectively introduced in L2 contexts. While Weigle (2007) has provided EFL writing teachers with proper assessment training input, more has to be specified concerning how teachers can learn to integrate teaching and assessing writing with constructive feedback in portfolio assessment. To echo the importance of assessment training, Hamp-Lyons (2006) found that the instructor was unable to give revisable feedback to Esing (the only informant in the study) or tell Esing about the strengths and weaknesses of her writing. Because the teacher was not skillful to assess writing, Esing was trapped in a negative feedback loop, showing no improvement in her later drafts. In Lam's (2019) study, while the two teacher informants were considered assessment-capable, they could merely mimic the *form* not the *essence* of portfolio assessment when asking their students to perform self-reflection. Based upon the above review, the following section discusses the role and effectiveness of assessment training.

14.2.2 Role and Effectiveness of Assessment Training

In research, assessment training refers to one form of professional development, which equips teachers with knowledge, skills, and principles about large-scale and classroom-based assessments. Undoubtedly, it plays a major role in facilitating the development of teacher assessment literacy (Popham, 2011). Recent studies reveal that school-level and university-level teachers are underprepared to perform assessment-related tasks, including preparing students for large-scale examinations adequately and synergising formative and summative assessments to promote learning (Xu & Brown, 2016). They find that teachers are particularly less proficient in performing the latter tasks. Notwithstanding the proliferation of language assessment textbooks, Davies (2008) warned that because the contents of these textbooks were chiefly ready-made and followed a cookie-cutter approach, teachers were unable to tryout those learnt testing theories with students. Some teacher education programmes in Hong Kong and Canada only offer assessment courses as an elective not a core course, so a certain number of pre-service teachers may not benefit from assessment training (Deluca & Klinger, 2010; Lam, 2015). Similarly, the teacher respondents in Europe reported that they learnt about assessment from colleagues and on the job (Vogt & Tsagari, 2014). What makes the assessment training picture more complex is that a majority of pre-service teachers' mentors, veteran in-service teachers, and language teacher trainers equally lack assessment capability (DeLuca & Johnson, 2017).

Despite an apparent lack of assessment literacy among teachers, the effectiveness of assessment training in classroom-based assessment remains mostly positive. In the US, around 75% of the respondents (mainly university-level instructors) received proper assessment training and were ready to implement alternative assessments (Crusan et al., 2016). Nevertheless, it did not necessarily mean that the respondents knew how to use writing portfolios to improve pedagogies. In China, Xu (2017) examined four novice EFL teachers' assessment literacy in a 3-year longitudinal study. Not until the third year of their practicum, did the two participants develop an enhanced knowledge of performing improvised formative assessment. It was concluded that assessment training together with personal learning and reflection proves to be the most effective. Zhang and Yan (2018) investigated the quality of multiple-choice test items used in a regional English language test in China. The results indicated that the two teachers could write reliable test items, had good intuitions of the level of difficulty of the test, but failed to have sufficient quality control of EFL tests like some ungrammatical items. Given these encouraging results, selected participants in Vogt and Tsagari's (2014) and Lam's (2019) studies demanded more assessment training in conducting writing portfolio assessment, as they felt less competent to do this. The data implied that teachers might know about preparing students for large-scale, standardised tests, so training them in that did nothing to help them use writing portfolios as a tool for improving teaching, learning, or assessment. In fact, the teachers expected to learn how to fulfill both learning and grading functions of assessment with portfolios more effectively. The next section takes a closer look at the assessment landscape in China.

14.2.3 (Writing) Assessment Landscape in China

In China, there has been a long history of utilising writing assessment to select civil servants. The prompts and contents of this archaic writing assessment were analogous to those of nowadays impromptu essay testing, where the examination conditions were highly standardised (Cheng & Curtis, 2010). This deep-seated testing culture has ideologically shaped the current examination system – Gaokao – a nation-wide college entrance examination. Gaokao is said to be a legacy of Confucian-heritage culture, where emphasis is put on effort, test performance, and a competitive learning mode (Carless, 2011). Consequently, high-stakes writing examination like Gaokao is commonly viewed as a means of upward social mobility, allowing students to become elites and professionals in the country. This predominant examination-oriented culture runs counter to the implementation of quality-oriented education reform in China, which promulgates experiential learning, critical thinking, and formative assessment (Tan & Chua, 2015). To obtain the best results in

Gaokao, students generally resort to studying the examination syllabus by rote, and teachers mostly adopt the didactic approach to conducting their lessons.

There are studies revealing how Gaokao negatively impacts teaching and learning in English language classrooms. In Gu's (2014) study, the teacher participant, Shelley, lamented that she struggled to strike a balance between following the curriculum reform initiatives (using formative assessment) and accommodating student needs to perform well in the public examination. Shelley added that her instructional approach was mostly governed by the Gaokao syllabus. Likewise, Yan (2015) reported that there were implementation gaps between new English curriculum requirements and teachers' classroom practices. The teacher participants preferred the product-based pedagogy to the process-oriented pedagogy owing to numerous barriers, including psychological challenges to teachers, students' resistance, lack of school support, and the backwash effect of the prevalent examination culture. To lower the stakes of Gaokao, Gu (2012) suggested that teachers take an eclectic stance of assessment by aligning teaching and testing with formative assessment and adopting multiple methods of assessment. Hamp-Lyons (2016) also noted that a transition from test use for bureaucratic purposes to test use for learning-enhancing purposes requires a high level of teacher assessment literacy, especially in an examination-dominated culture like China.

Thus far, the use of alternative assessments in the new English curriculum is high on the agenda in China, namely writing portfolio assessment. Nevertheless, from the reviewed literature, assessment training about the use of writing portfolios for teaching and assessment appears to be scarce and less effective, particularly in the context of Chinese learners of English. Also, there are clear implementation gaps between the assessment reform policies and actual classroom practices when teachers innovate their writing assessment practices. To better understand these dilemmas, the study intends to address the following two research questions:

1. What is the perceived role and effectiveness of assessment training in writing portfolio assessment?
2. To what extent does the assessment training help resolve individual, institutional, and cultural issues when the participants plan to attempt the portfolio approach?

14.3 Methodology

14.3.1 Research Design

The study adopted a qualitative methodology, enabling the author to gain an in-depth perspective of the role and effectiveness of assessment training in writing portfolios. Using a case study approach, the author could specifically examine how the participants experienced the assessment training, and whether the training would facilitate or inhibit the possibility of introducing portfolio assessment in their workplaces. The case study approach was likely to generate unique insights into the importance of assessment training, especially within the current assessment reform landscape in China. More importantly, it deepened various stakeholders' understanding of how the assessment training fostered the development of teacher assessment literacy.

14.3.2 Participants

Three key informants participated in the study, including Joan, Rebecca, and Taylor (pseudonyms). They were females, attending a first-year doctorate programme at one comprehensive university in Hong Kong. Joan, Rebecca, and Taylor had 3–7 years' teaching experience in China. Joan taught speaking and writing at a private tutorial school in the southern part of China. Rebecca taught general English in a Hong Kong government-funded secondary school, and Taylor taught translation and interpretation at a top-tier Guangdong university. Prior to the study, the informants claimed that they had not received any language assessment training.

14.3.3 The Assessment Course

The three participants received assessment training via an 11-session content course about English language assessment. One topic strand of the course included three lectures on basic knowledge of language assessment, and two workshops on the application of writing portfolio assessment in L2 environments. Each lecture and workshop lasted for three hours. The contents of the lectures covered: basic assessment principles (e.g., validity and reliability); various assessment purposes; theories of classroom-based assessment; and language assessment literacy. The contents of the workshops consisted of: principles, issues and recommendations of writing portfolio assessment, and feedback provision and enactment in L2 writing classrooms.

14.3.4 Data Collection and Analysis

Three data sources were used to collect qualitative data: (1) a pre-workshop open-ended questionnaire; (2) a post-workshop individual interview; and (3) a reflection paper. The questionnaire aimed to understand the participants' views and practices of writing assessment prior to the training. The interview elicited their insights into the usefulness of the assessment training. The reflection paper identified how the participants could mediate individual, institutional, and cultural issues when they planned to attempt writing portfolio assessment. The questionnaire had 3 parts and 17 items, including background, perceptions of language assessment, and assessment training (see Appendix 1). It was administered in Week 2 of the course before the lectures and workshops commenced. An interview guide comprising 8 questions was adopted (see Appendix 2). The three individual interviews were conducted in Week 9 of the course after the assessment training completed. The reflection paper required the participants to write about why, how, and what should be changed in writing assessment practices in their work contexts. The participants were expected to critique the change process with theories, observations, and site-based evidence.

Data were analysed with the following procedures: assembling, coding, comparing, and interpreting (Burns, 2010). Assembling the data is about reading and re-reading all data sources before coding. Deductive coding was adopted to blend in the two themes under study: (a) the role and effectiveness of assessment training, and (b) the ways assessment training mediates multi-level issues when portfolios are put into practice. Questionnaire and interview data were compared during the coding process. Partial interview and documentary data (reflection papers) were also juxtaposed to check whether the participants' views and actions converged or diverged. After comparing, the author could develop insights by interpreting the processed data relating to the findings of current scholarship on assessment literacy and his own research experience.

14.4 Results

14.4.1 Research Question 1

To address the perceived role and effectiveness of assessment training, this section reports the three participants' pre-training and post-training perceptions.

14.4.1.1 Pre-training Perceptions (Questionnaire Data)

Before the assessment training, Rebecca, Joan, and Taylor said that they received no training in language assessment or any forms of alternative assessment. From the questionnaire, the three key informants were eager to learn about L2 writing assessment, since assessing writing was complicated. Neither did the participants apply writing portfolio assessment in their teaching contexts previously although Rebecca has heard about writing portfolios when working as a teaching assistant in Hong Kong. When asked about whether portfolio assessment could replace one-off, impromptu writing assessment in China, Rebecca emphasised that this idea was not likely to happen due to the issues of practicality and scoring consistency, and Joan mentioned that time would be a major barrier to use portfolio assessment. Interestingly, Taylor was somewhat enthusiastic about using portfolio assessment to replace existing standardised testing, but she proposed that more empirical research was needed to substantiate its large-scale application.

Although the participants have not learnt about portfolio assessment, they expressed its relevancy to their teaching jobs and showed interests in giving portfolio assessment a go, especially for Taylor who planned to introduce e-portfolio in her university. Concerning the levels of understanding, even without proper training, Rebecca and Taylor came to grips with some rudimentary concepts and principles of writing portfolio assessment. For instance, Rebecca was concerned with the practicality and scoring issues when portfolios were applied. She further jotted down a phrase 'low reliability' as a challenge in portfolio implementation. To Taylor, she distinguished the differences between large-scale and classroom-based assessments and categorised portfolio assessment as one form of the latter. However, for Joan, she seemed to have limited knowledge about educational assessment. In her questionnaire, she mainly discussed the role of large-scale testing like Test for English Majors 8 and showed little understanding of classroom-based assessment like writing portfolios.

When asked about their expectations towards the assessment training, the participants had different views. For example, Rebecca wanted to learn about giving effective written corrective feedback, because it could help resolve students' immediate writing problems. She believed that written corrective feedback might facilitate the development of self-assessment skills. While Rebecca preferred a quick-fix approach to assessment training, she remained inquisitive to learn how to boost student motivation for keeping portfolios and use feedback to inform teaching and learning of writing. Similarly, Joan stated that she was keen on acquiring some hands-on experience of portfolio-based lessons, including authentic classroom examples and down-to-earth implementation procedures. She felt that these examples could equip her with adequate knowledge and skills in carrying out portfolio assessment. Unlike Rebecca and Joan, Taylor wished to learn about theories and classroom applications of writing portfolio assessment, because she considered both theory and practice were significant for her to conduct research and improve pedagogy.

14.4.1.2 Post-training Perceptions (Interview Data)

Generally, the participants were positive about the role of assessment training, given that they had learnt about the principles, features and procedures of writing portfolio assessment. By attending the lectures and workshops, they developed a deeper understanding of what portfolio assessment entailed. All three participants found the lectures, academic readings, discussion forums, and mini-project task very beneficial, which might enhance their awareness and conceptual understanding of portfolio assessment. Despite its facilitative role, the participants advised the instructor to invite guest speakers (preferably frontline teachers) to share their portfolio tryout experiences. Rebecca proposed to include a workshop on scoring in writing portfolios with well-defined rubrics. Further, Taylor suggested that the weekly reading task should be graded and made compulsory, so that the participants became motivated to read up the assessment literature regularly.

Prior to the interviews, Rebecca and Joan had reservation about the usefulness of the assessment training although they expected to learn about how to put writing portfolio assessment in action. Before the training, Rebecca and Joan thought that portfolios could only serve the formative purpose as its application in large-scale testing remained unproven. After receiving the training, Rebecca changed her mind and believed that portfolio assessment could serve both summative and formative purposes, provided that teachers were able to score student portfolios impartially and accurately. Moreover, before the training, Rebecca misunderstood that portfolio implementation would increase teacher workload. Yet after the workshop, she realised that the portfolio approach, advocating learner autonomy and self- and peer-assessment, might reduce teacher workload accordingly, because students could share assessment responsibility with their teachers together.

For Joan, even after training, she did not have an obvious change in her belief – a trust in high-stakes testing. During the interview, Joan was very skeptical about the benefits of writing portfolio assessment, as most teachers in China did not know this new trend. She added that because of an examination-driven culture, students would ignore the importance of writing development and simply focus on the assessment results. Joan also emphasised that portfolio scoring was subjective and the issue of fairness remained unresolved. She said, '*I want to know how to set up reliable criteria to assess students in a fair way*'. As to Taylor, she reported that after the training, she developed a better understanding of the principles and practices of portfolio assessment, and decided to research on this approach. Her plan was to set up an e-portfolio system in her affiliated university. Then, she investigated her own portfolio application together with her colleagues via an action research study. Taylor's proactive initiative to change was borne out by this quote, '*They (Taylor's colleagues) are talking about how to change assessment in their lectures. Yeah, they want to bring in formative assessment. And I talked with them about portfolio assessment and they are interested.*'

In sum, the assessment training served as a form of professional development for the participants, especially when all of them received no training in language assessment. The training played a *facilitative* role in enhancing the three participants' understanding of the principles and practices of writing portfolio assessment. Based upon the data, it seems that Rebecca and Taylor benefited more from the assessment training than Joan due to the fact that Rebecca was reflective upon how she assessed student writing pedagogically (i.e., written corrective feedback) and Taylor was open-minded and passionate about researching a new assessment approach (i.e., attempts to initiate e-portfolios). Joan also gained new knowledge after the training, but still held a deep-seated view that conventional standardised testing was superior to portfolio assessment.

14.4.2 Research Question 2

To address the extent to which the assessment training resolves multi-level issues, this section details the three participants' post-training perceptions and their assessment reform plans in the reflection papers.

14.4.2.1 Post-training Perceptions (Interview Data)

When asked about in what ways the assessment training mediated individual, institutional, and cultural issues, the three participants had different perspectives. At the individual level, Rebecca believed that the assessment training could deepen her understanding of using portfolios as a classroom-based assessment method. At the institutional level, Rebecca thought that the assessment training was able to change school leaders' mindsets, enabling them to be more receptive to innovations. For instance, school leaders might encourage teachers to attempt various alternative assessment approaches. To Rebecca, the assessment training might not successfully mediate a wider cultural issue if portfolios were adopted as a large-scale assessment. She stated that writing portfolio assessment might reduce the stakes of standardised testing and student study pressure. However, she felt that to measure student writing via portfolios remained complex and subjective. Rebecca concluded that the assessment training might change teachers' and school leaders' beliefs in the usefulness of portfolio assessment, but not its large-scale application, because the latter seemed to be logistically problematic and empirically unproven.

For Joan, the assessment training could equip her with fundamental knowledge on portfolio implementation. She believed that the assessment training could enhance her confidence when attempting new assessment methods. However, at the institutional level, she wondered how much school leaders would support teachers when they initiated assessment change. Joan explained that not every school or district in China received sufficient resources to pilot writing portfolio assessment, given that assessment reforms involved additional teacher training, student commitments, school management endorsement, and parent support. She expressed her concerns whether the assessment training could resolve the cultural-related issues, because the current assessment practices in China were heavily examination-driven and governed by bureaucratic education policies. Despite her willingness to attempt portfolio assessment, Joan thought that the assessment training took up a minor role (around 30%) in mediating these multi-level issues.

In the interview, Taylor reckoned that the assessment training was effective to change teacher beliefs about the value of portfolio assessment. Institutionally, Taylor was hesitant, saying that changing school leaders' mindsets to adopt new assessment methods was a long-term endeavour. Also, the assessment training would have more direct impact on teachers than on school administrators. With that being said, Taylor was somewhat hopeful that the assessment training could mediate cultural-related issues. She further added that owing to Gaokao, change in assessment practices might take time and need communal support. But, at the tertiary level, she could promote writing portfolio assessment more steadily, because university instructors had greater autonomy than school teachers concerning educational reforms.

14.4.2.2 Assessment Reform Plans (Documentary Data)

The use of assessment reform plans served to find out the extent to which the assessment training mediated individual, institutional, and cultural issues when the three participants introduced writing portfolio assessment. In Rebecca's paper, she critiqued journal writing as an assessment tool in one Hong Kong secondary school. Rebecca argued against evaluating student writing by journals due to the following reasons: (a) comprehensive marking; (b) no involvement of students in the assessment process; (c) emphasis on linguistic accuracy; and (d) no timely feedback (journal entries returned to students rather late). After Rebecca identified these feedback issues, she proposed a new assessment plan with eleven steps. She went on to justify why she made such changes. For instance, she planned to promote active learning, greater involvement of students in the assessment process, and use of portfolios to encourage reflectivity. Near the end of the paper, Rebecca suggested that teachers should consolidate their assessment literacy by giving students revisable/timely feedback and by marking student writing more accurately. Rebecca advised that instructed training should be given to students before they were asked to perform self- and peer assessment. From Rebecca's paper, it was clear that she had a thorough understanding of feedback for learning. She was able to identify assessment issues and propose changes with classroom evidence. She has built clear pedagogical insights into the assessment problem that happened in her work place. Although she only briefly mentioned portfolio assessment, she incorporated the notion of continuous feedback into writing portfolio assessment. The assessment training could effectively help Rebecca to mediate individual and institutional issues.

Joan's paper focused on evaluating the likelihood of introducing writing portfolio assessment in Chinese secondary schools. In the paper, Joan displayed a basic understanding of the rationale and principles of portfolio assessment. Additionally, Joan pointed out that teachers and administrators may encounter constraints when introducing writing portfolio assessment, including student weaknesses in writing; packed teaching schedules; and low levels of assessment literacy. Nonetheless, when she discussed three classroom examples of writing portfolios, she only cited three common ELT practices, which were unrelated to portfolio assessment such as displaying student good works on bulletin boards; jotting down useful phrases and vocabulary items; and keeping grammar plus vocabulary correction books. When it came to suggesting ideas on wider portfolio application, Joan was unable to provide concrete recommendations except on the point of school support. The tone of Joan's reflection paper appeared to be less affirmative probably due to limited teaching experience and a lack of exposure to portfolio application. Hence, the assessment training might moderately mediate the individual issue (change of mindsets) rather than institutional and cultural issues, given that Joan firmly believed in the significance of large-scale assessment.

Taylor's paper focused on proposing a change in the assessment practices of a consecutive interpreting course. In her work, Taylor demonstrated an advanced understanding of formative and summative assessment and articulated why change in assessment was necessary. A well-defined gap to innovate assessment formats was identified from the research literature, namely the benefits of e-portfolios. Because of the availability of resources and accessibility of technology, Taylor could steadily introduce e-portfolios in her programme. Her assessment plan included the newly-added contents of an e-portfolio programme like self-assessment reports, reflective diaries, and selection of best interpretation recordings. For evaluation, Taylor constructed a criteria-referenced rubric relating to these contents. Based upon Taylor's proposal, she was quite determined to innovate the current assessment practices with theoretical justifications and pedagogical rationale. Taylor has even set a 1-year timeline to introduce the assessment change. Given that the assessment training empowered Taylor to be *a change agent*, it enabled her to mediate individual, institutional, and also cultural issues (willingness to challenge the assumption of the psychometric paradigm of assessment).

In brief, Rebecca and Taylor appeared to be more optimistic about using the assessment training to mediate multi-level issues than Joan who had great faith in high-stakes testing. Having analysed their reflection papers, the author finds that Joan could only use the assessment training to mediate individual but not institutional and cultural issues due to her lack of experience in alternative assessments, whereas Rebecca utilised the assessment training to mediate both individual and institutional issues by reflectively challenging the existing corrective feedback practices. For Taylor, she was probably the most assessment-competent participant, who best used the assessment training to mediate all levels of issues when she was about to launch her e-portfolio programme. Having said that, all three participants, indeed, learned about portfolio assessment well enough to think more deeply and usefully about it. Their self-assured feedback confirmed their willingness to innovate writing portfolios regardless of challenges.

14.5 Discussion

This section characterises the three participants' roles within an assessment training landscape in the use of writing portfolios, followed by a discussion on the usefulness, quality, and needs of assessment training. Rebecca was seeking best written corrective feedback practices which could be applied in her school. She was knowledgeable about the dynamic interplay between the formative and summative purposes of writing assessment. She also had a solid understanding of the assessment principles in general and the theory of writing portfolio assessment in particular. She cautioned the importance of practicality when evaluating student writing with written corrective feedback. With these in mind, Rebecca can be said *an inquisitive practitioner*, who utilised assessment training to enrich her assessment repertoire.

Joan was keen on learning about the basic principles of portfolio assessment. However, she was concerned with its ethical issues, such as test fairness (e.g., non-standardised assessment conditions) and scoring consistency (e.g., rater subjectivity). Joan believed that students and parents were typically examination-oriented, only focusing on the outcomes of Gaokao but ignoring the advantages of portfolio assessment. Owing to her limited exposure to L2 writing assessment, Joan did not benefit much from the assessment training and remained hopeful about standardised testing. She can be said *a disciple of high-stakes testing*, who considered portfolio assessment not suitable to be adopted in public examinations. Taylor confidently mastered the rationale behind portfolios after the assessment training, which inspired her to launch the e-portfolio programme. Taylor was fervent about applying the principles of portfolio assessment into practice. She discussed the new assessment mode (e-portfolios) and planned ahead the logistics of implementation with her colleagues and the author. She also looked forward to seeing more assessment innovations in China, such as China's Standards of English Language Ability which is a Chinese equivalent of Common European Framework of Reference for Languages. Taylor can be said *a game-changer of writing assessment* as she professionally initiated reforms in the assessment practices.

From the results, the assessment training provided in this study was generally effective although it did not change all the participants' mindsets in the use of portfolios to improve teaching and learning of writing. The three participants were rather positive about the usefulness of the assessment training, because they had not received formal training in L2 writing assessment formerly. Notwithstanding its positive impact, the participants felt that the training should better narrow the theory-practice divide by providing more hands-on experience, examples of authentic portfolio applications, and practical sharing by guest speakers. Except Taylor, it appears that the assessment training might not assist the participants to mediate multi-level issues when they planned to introduce portfolio assessment in their schools. For instance, Joan still had a misunderstanding towards the classroom-based portfolio implementation and did not feel convinced of its use as summative assessment. She also failed to suggest actionable recommendations regarding how her affiliated institution could support her when she introduced the alternative assessment. Therefore, the assessment training may not essentially serve as a panacea for the development of teacher assessment literacy.

In fact, the quality of assessment training matters most if we want to enhance teacher assessment literacy in L2 writing (Lam, 2019). The quality of training entails the scope of meetings, course syllabi, practice opportunities, or authenticity in course materials. There are other factors including teacher commitments, teacher beliefs, institutional support, and a larger socio-cultural setting, which may facilitate or impede practitioners' uptake of assessment knowledge, skills, and principles in the mandate training (Xu & Brown, 2016). Institutionally, the quality of assessment training requires constant updates by hiring seasoned scholars to run short-term to middle-term professional development courses although these initiatives need financial resources. Nationally, the Ministry of Education encourages school-university collaborations via action research like in Taylor's case in order to promote a bottom-up approach to assessment training. Concerning the needs of assessment training, policymakers could survey frontline teachers' needs by identifying their perceptions towards beliefs, knowledge, and skills about L2 writing assessment (cf. Crusan et al., 2016). Based upon the questionnaire data, service providers could design context-specific assessment training manuals for teachers who evaluate their student writing by portfolios in diverse educational settings and geographical locations.

14.6 Implications and Conclusion

The study sheds new light on the importance of assessment training in the use of portfolios, especially within a context of Chinese learners of English. The findings of this study further advance our understanding that assessment training is a necessary but *insufficient* condition to make portfolio application successful in EFL environments. The three participants were qualified, eager, and academically able to try out portfolio assessment. Nonetheless, to allow successful integration of portfolios into the classroom and to use them as a means for both formative and summative evaluations require more than systematic training. Institutional support (e.g., teacher-to-teacher mentoring) and contextual support (e.g., financial support from the government) all play a part in shaping why some teachers are more motivated to implement portfolio assessment than others. Thus, it is indispensable for administrators to scale up the assessment training in portfolio assessment. For instance, our data imply that the participants want to learn how to score writing portfolios summatively, given that scoring portfolios is a highly skilled activity (Hamp-Lyons, 2006). Second, our data also imply that besides setting up portfolio systems, the participants need the skills to evaluate their own portfolio implementation through reflective practices, such as teacher reflection groups, journals, or exploratory practice (Hanks, 2015). Exploratory practice is a form of continued professional development, in which teachers reflect upon and investigate their practice, and improve the quality of teaching life through less rigorous research procedures. Third, in average assessment training courses, there should be a healthy balance between theory and practice. Our participants told us that they came to grips with the principles of writing portfolios, but lacked adequate hands-on experience to attempt the new approach. Future assessment training may include portfolio grading tasks, self-reflection tasks, and online discussion tasks on sharing good portfolio practices by and with frontline teachers. Despite its theoretical contributions, the study has its limitations. It has a small sample size and the findings primarily draw upon self-reported data. However, with data triangulation and objective interpretations, the results of the study remain dependable albeit not generalisable to a larger EFL writing context.

Acknowledgements This work was funded by the Language Fund under Research and Development Projects 2018-19 of the Standing Committee on Language Education and Research (SCOLAR), Hong Kong SAR. The project reference number is (EDB(LE)/P&R/EL/175/4).

Appendices

Appendix 1

Open-ended questionnaire

Part A: Background

1. What is your teaching context? Tick as appropriate.
   □ Kindergarten □ Primary school □ Secondary school □ Vocational training school □ Training school □ University
2. What is your teaching experience?
3. What is the location of your school/university? (e.g., the name of town, city, or province)
4. Besides teaching English, are you responsible for other administrative duties? Fill in 'Yes' and what position do you hold? Or 'No'.
5. What is the last employment before you join the EdD Programme at University A?

Part B: Perceptions of language assessment

6. What is your understanding of writing assessment? And could you give ONE example of classroom-based writing assessment in the Chinese context?
7. What is the relationship between large-scale essay testing and classroom-based writing assessment?
8. Have you heard about writing portfolio assessment? Did you use writing portfolios when you were a school/university student in China? If yes, please elaborate on your experience. If no, please proceed to Q10.
9. What is the rationale behind writing portfolio assessment?
10. Do you think writing portfolios can be used to replace standardised writing assessments like those in TEM or classroom-based writing assessments (composition writing)? Why or why not?

Part C: Assessment training

11. Have you received any writing assessment training such as coursework, seminars, lectures, or online courses? If yes, what have you learnt? If no, proceed to Q12.
12. What is your expectation about EDUD XXX? What do you expect to learn after taking the course? Feel free to elaborate on your response.

13. Do you think learning about writing portfolio assessment is relevant to your job? Why or why not?
14. To what extent does the assessment training help resolve individual (teacher beliefs), institutional (workload or school support), and cultural (an examination-driven society) constraints when you introduce the portfolio approach in the Chinese context?

Individual issues: □ very likely □ likely □ neutral □ not likely □ not very likely

Explanation:

Institutional issues: □ very likely □ likely □ neutral □ not likely □ not very likely

Explanation:

Cultural issues: □ very likely □ likely □ neutral □ not likely □ not very likely

Explanation:

15. What do you want to learn regarding classroom-based portfolio assessment and why do you want to learn about those aspects?
16. Since portfolios have become a trend in L2 writing assessment, what factors will facilitate or inhibit its wider application in China?
17. Other comments:

Appendix 2

Interview guide:

1. What is your understanding of L2 writing assessment?
2. What do you think about the usefulness of lectures and workshops on writing portfolio assessment?
3. Do you have a better understanding of writing portfolio assessment after the workshops? Why or why not? Please give ONE example.
4. Which aspects of the assessment training do you like most and why? And which aspects do you feel less satisfactory and why?
5. What assessment knowledge and skills do you need if portfolios are used to replace the current writing assessment in the Chinese context?
6. To what extent does the assessment training help resolve individual, institutional, and cultural issues when you attempt portfolios as an alternative to writing assessment?
7. In your opinion, how likely do you think teachers/lecturers in China will adopt portfolios to achieve both formative and summative purposes of assessment?
8. Thus far, what form and content of writing assessment training do you prefer and why? Lastly, do you have any comments on the assessment training provided in EDUD XXX?

References

Burner, T. (2014). The potential formative benefits of portfolio assessment in second and foreign language writing contexts: A review of the literature. *Studies in Educational Evaluation, 43*, 139–149.

Burns, A. (2010). *Doing action research in English language teaching: A guide for practitioners*. Routledge.

Carless, D. (2011). *From testing to productive student learning: Implementing formative assessment in Confucian-Heritage Settings*. Routledge.

Cheng, L., & Curtis, A. (Eds.). (2010). *English language assessment and the Chinese learner*. Routledge.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43–56.

Curtis, A. (2018). Portfolios. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (1st ed.). Wiley. https://doi.org/10.1002/9781118784235.eelt0326

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347.

DeLuca, C., & Johnson, S. (2017). Developing assessment capable teachers in this age of accountability. *Assessment in Education: Principles, Policy & Practice, 24*(2), 121–126.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419–438.

Gu, P. Y. (2012). English curriculum and assessment for basic education in China. In J. Ruan & C. B. Leung (Eds.), *Perspectives on teaching and learning English literacy in China* (pp. 35–50). Springer.

Gu, P. Y. (2014). The unbearable lightness of the curriculum: What drives the assessment practices of a teacher of English as a foreign language in a Chinese secondary school? *Assessment in Education: Principles, Policy & Practice, 21*(3), 286–305.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*(1), 5–16.

Hamp-Lyons, L. (2006). Feedback in portfolio-based writing courses. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing contexts and issues* (pp. 140–161). Cambridge University Press.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 487–504). Springer.

Hamp-Lyons, L. (2016). Purposes of assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 13–27). De Gruyter.

Hanks, J. (2015). 'Education is not just teaching': Learner thoughts on exploratory practice. *ELT Journal, 69*(2), 117–128.

Jiang, H., & Hill, M. F. (Eds.). (2018). *Teacher learning from classroom assessment: Perspectives from Asia Pacific*. Springer.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169–197.

Lam, R. (2018). *Portfolio assessment for the teaching and learning of writing*. Springer.

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System, 81*, 78–89.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.

Mak, P., & Wong, K. (2018). Self-regulation through portfolio assessment in writing classrooms. *ELT Journal, 72*(1), 49–61.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265–273.

Scott, T. (2005). Creating the subject of portfolios: Reflective writing and the conveyance of institutional prerogatives. *Written Communication, 22*(3), 3–35.

Stiggins, R. (2014). Improve assessment literacy outside of schools too. *The Phi Delta Kappan, 96*(2), 67–72.

Tan, C., & Chua, C. S. K. (2015). Education policy borrowing in China: Has the West wind overpowered the East wind? *Compare – A Journal of Comparative and International Education, 45*(5), 686–704.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402.

Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194–209.

Xu, H. (2017). Exploring novice EFL teachers' classroom assessment literacy development: A three-year longitudinal study. *The Asia-Pacific Education Researcher, 26*(3-4), 219–226.

Xu, Y. T., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149–162.

Yan, C. (2015). 'We can't change much unless the exams change': Teachers' dilemmas in the curriculum reform in China. *Improving Schools, 18*(1), 5–19.

Zhang, G., & Yan, X. (2018). Assessment literacy of secondary EFL teachers: Evidence from a regional EFL test. *Chinese Journal of Applied Linguistics, 41*(1), 25–46.

Ricky Lam is Associate Professor in the Department of Education Studies at Hong Kong Baptist University. His publications have appeared in Assessing Writing, Language Testing, TESOL Quarterly, and other international journals. He has recently published a book entitled 'Portfolio assessment for the teaching and learning of writing'. His research interests include digital portfolios and language assessment literacy.

Index

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