Investigating the consequential validity of the *Hanyu Shuiping Kaoshi* (Chinese proficiency test) by using an Argument-based framework

Shujiao Wang

Department of Integrated Studies in Education

McGill University, Montreal

January 2018

A thesis submitted to McGill University

in partial fulfilment of the requirements of

the degree of doctor of philosophy

ACKNOWLEDGEMENTS

I would like to take some time to thank those without whom this project would never have been possible. Although it is just my name on the cover, many have contributed to the research in ways deserving of special thanks.

I am grateful for the financial support I received for my research and funding for my Ph.D. program including but not limited to:

- Paragon Testing Grant, 2017
- Herschel & Christine Victor Fellowship, 2014-2015
- SSHRC- Joseph-Armand Bombardier CGS Doctoral Scholarship, 2014-2017
- McGill's Graduate Entrance Scholarship, 2013

I am also indebted to a long list of individuals who have personally supported me during my studies. First and foremost, I would like to express my deep appreciation and thanks to my supervisor Dr. Carolyn Turner, who has been a tremendous mentor for me. It has been an honor to be her last Ph.D. student. I appreciate all her contributions of time, ideas, and encouragement to make my Ph.D. experience rewarding and to allow me to grow as a researcher. I greatly appreciate the freedom she has given me to find my own path and the guidance and support she offered when needed. I am also thankful for the excellent example she has provided as a successful woman educator and professor.

I would also like to express my gratitude to my co-supervisor Dr. Mela Sarkar. I want to thank you for your brilliant comments and suggestions on my dissertation from a different perspective. Our conversations about academia and life have helped shape my personal outlook and professional career. I further acknowledge the contributions of my dissertation advisory committee member Dr. Liying Cheng. Your dedicated work in washback has always inspired

me. Your critical, insightful and detailed comments based on your expertise have helped me gain deep understanding of the complexity of consequential validity and how to adapt an argument-based framework.

My time at McGill was made enjoyable and enriched by dedicated faculty, staff, and friends that have become a privileged part of my life. I am grateful for the time I spent with all of you. Special thanks go to Education and East Asian Studies teachers and students. Thank you all for your help and encouragement along the way. I have benefited from you at various stages of my studies. For example, I gained valuable research experience by assisting Dr. Beverly Baker. Thank you for your insightful and enriching comments on my proposals, presentations, and papers. I want to thank Mengting Pan for her great help in contacting participants during my data collection process. I also want to thank Cindy Chiang for your proofreading of the dissertation. Thanks also to Hui Wang for your help with the statistical analyses of the data. I am grateful as well to all the participants in this study. Your enthusiasm for teaching and learning Chinese always touched and inspired me to conduct research in this area, and even beyond what I ever could have been able to do. It is my hope and persistence that my study will provide useful information to second language teachers and learners, test developers, and other stakeholders.

Last but most important of all, I would like to thank my family for all their love, encouragement, and support. To my parents and parents-in-law, thank you for supporting all my pursuits. Words cannot express how grateful I am to you for all of the sacrifices that you have made on my behalf. And for my loving, supportive, encouraging, and patient husband Zhenhuan, your faithful support throughout this PhD experience is deeply appreciated. Most of all, special thanks to my beloved son Zeyu. You are such a good boy who always cheers me up. I love you.

ABSTRACT

In recent years, China's rising global power has led to an international increase in Chinese language learning. The national standardized test of Chinese language proficiency for non-native speakers, the *Hanyu Shuiping Kaoshi* (HSK), literally "Chinese Proficiency Test," has played a vital role in certifying language proficiency for higher education and professional purposes. The multiple uses of the HSK have generated growing concerns about its validity, especially the reformed HSK's (post-2009) consequential validity. Employing Bachman and Palmer's (2010) Assessment Use Argument (AUA) framework, this mixed methods sequential exploratory (MMSE) study investigates the HSK's micro- and macro-level consequences, as well as how and to what extent the test affects Chinese as a second language (CSL) teaching and learning. In Phase I of this MMSE study the official HSK documents were analyzed by content analysis; interviews with 12 test stakeholders were then conducted and analyzed by a two-cycle qualitative coding approach (Saldaña, 2009). In Phase II, 136 CSL/CFL teachers and 512 HSK test-takers participated in a questionnaire, and the data were analyzed by using exploratory factor analysis (EFA) and structural equation modeling (SEM); classroom observations were also conducted and analyzed to contextualize the quantitative results. Phase III involved two exploratory questionnaires and interviews with 35 administrative personnel who use the HSK to inform academic and employment decisions, and the data were analysed through statistical (e.g., descriptive statistics) and qualitative methods (e.g., grounded theory). The results of the MMSE study highlighted the complexity of the HSK's consequences and washback effects. They indicated that although the HSK had limited effects on teaching, it was somewhat successful in its goal of promoting CSL/CFL learning. In general, HSK scores and other related information (e.g., score report, level interpretation) also provided users with relevant, useful, and meaningful

data for candidate selection. Overall, based on the HSK's AUA conceptual framework, the findings provided evidence that Claim 1 (Consequences), Claim 2 (Decisions), and Claim 3 (Interpretations) were partially supported, in that the test developers' intended goals for the HSK were only achieved to a certain degree. This study helped unpack the consequential validity of the HSK in the CSL context, shed light on understanding the washback effects of the HSK, fleshed out the values underlying the multiple interpretations and uses of the test, and pointed to implications for the HSK developers and future consequence/impact/washback research.

RÉSUMÉ

Ces dernières années, la puissance mondiale croissante de la Chine a conduit à une augmentation internationale de l'apprentissage de la langue chinoise. Le test standardisé national de maîtrise de la langue chinoise pour les non-natifs, le Hanyu Shuiping Kaoshi (HSK), littéralement « test de compétence en chinois », a joué un rôle essentiel dans la certification des compétences linguistique et professionnelle. Les multiples utilisations du HSK ont généré de plus en plus d'inquiétudes quant à sa validité, en particulier la validité corrélative du HSK (post-2009). Utilisant le cadre d'évaluation de l'utilisation [AUA] de Bachman et Palmer (2010), cette étude exploratoire séquentielle à méthodes mixtes (MMSE) examine les conséquences micro et macro du HSK, ainsi que comment et dans quelle mesure le test affecte le chinois en tant que langue seconde de l'enseignement et d'apprentissage (CSL). Dans la phase I de cette étude MMSE, nous avons examiné les contenus des documents officiels de HSK ; des entrevues avec 12 intervenants ont ensuite été réalisées et analysées selon une approche de codage qualitatif en deux cycles [Saldaña, 2009]. Au cours de la phase II, 136 enseignants du programme CSL/CFL et 512 participants au test HSK ont répondu à un questionnaire, ce qui nous ont permis de recueillir des données à l'aide de l'analyse factuelle exploratoire (EFA) et de la modélisation par équation structurelle (SEM) ; des observations en classe ont également été menées et analysées afin de contextualiser les résultats quantitatifs. La phase III comportait deux questionnaires exploratoires et des entretiens avec 35 membres du personnel administratif qui utilisaient la HSK pour éclairer les décisions académiques et professionnelles, et les données étaient analysées en utilisant des méthodes statistiques (statistiques descriptives, par exemple) et des méthodes qualitatives comme la théorisation ancrée. Les résultats de l'étude MMSE ont mis en évidence la complexité des conséquences du HSK et des effets de *washback.* Ils ont indiqué que bien que le

HSK ait eu des effets limités sur l'enseignement, il a plutôt réussi à promouvoir l'apprentissage CSL/CFL. En général, les scores HSK et d'autres informations connexes (par exemple, rapport de score, interprétation de niveau) ont également fourni aux utilisateurs des données pertinentes, et significatives pour la sélection des candidats. Dans l'ensemble, selon le cadre conceptuel de l'ASA de HSK, les constatations ont démontré que les revendications 1 (Conséquences), 2 (Décisions) et 3 (Interprétations) étaient partiellement justifiées, car les objectifs des développeurs de tests pour le HSK étaient seulement atteints dans une certaine mesure. Cette étude a permis de comprendre la validité et les impacts de la HSK dans le contexte CSL, les effets de *washback* du HSK ; d'étoffer les valeurs sous-jacentes aux multiples interprétations et utilisations du test et de souligner les implications des développeurs HSK et les impacts de la recherche *washback* dans le futur.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 The rationale of the study

In the field of language testing (LT), validity theory has received increasing attention over the past decades (e.g., Brown, 2004; Messick, 1989, 1996; Moss, 1992; Young, 2008). In the early discussion, validity was predominantly defined through discrete forms of validity (i.e., content, criterion, construct); this was referred to as the "trinitarian" view (Guion, 1980, p. 385). Messick's (1989) unified model of validity has broadened our understanding of validity to be a multifaceted concept with encompassing value implications and social consequences. Although attempts have been made to explore the perspectives beyond the narrow sense of validity (e.g., consequences and ethical considerations), efforts have "failed to provide an explicit link between validity and test use" (Bachman, 2005, p.7). More specifically, few empirical studies have investigated test consequences within a coherent validation framework in order to evaluate the validity argument strength for a particular test (Chapelle, Enright, & Jamieson, 2008, 2010).

One aspect central to consequential validity is washback (Weir, 2005). Messick (1994, 1996) regarded it as an "instance of the consequential aspect of construct validity" (p. 242). Washback refers to "the impact of a test on learners and teachers, on educational systems in general, and on society at large" (Hughes, 2003, p.53). The amount of literature on washback has demonstrated the importance of this issue and has provided valuable considerations for language education. However, the washback literature is rather limited with respect to the languages investigated and the range of stakeholders involved. More specifically, researchers have neglected languages other than English (Huang, 2013; Manjarrés, 2005) and perspectives besides those of teachers and learners (Cheng, 2014).

In recent years, China's status as a rising global power has led to an international increase

in Chinese language learning (Odinye & Odinye, 2012; Wang, 2016). The national standardized test of Chinese language proficiency for non-native speakers - *Hanyu Shuiping Kaoshi* (HSK), literally "Chinese Proficiency Test," has played a vital role in certifying language proficiency, especially for higher education and professional purposes. Despite its significance and its test developers' claims of high internal validity and reliability (Luo, Zhang, Xie, & Huang, 2011), it is surprising that very little focus has been paid to the HSK, especially the reformed HSK[1]. In fact, only a few empirical studies relating to the test's consequential validity have been conducted (Huang, 2013; Huang & Li 2009). This observation is ultimately incommensurate with the test's important status.

Moreover, as Kane (2006) argued, a test used to implement education policy should be evaluated in terms of its consequences. The first World Chinese Conference in 2005 marked the recognition of "promoting Chinese language internationally" (PCI) as a national strategic policy (Li, 2012). Since then, a series of advancements have occurred such as the opening of Confucius Institutes, the training of Chinese as a Second/Foreign Language (CSL/CFL) teachers, as well as the launch and reform of Chinese proficiency tests. Thus, the understanding of HSK's development and revision is important in this context.

Subsequently, to investigate the HSK's role in teaching/learning CSL/CFL within the context of PCI, more rigorous empirical studies are needed to explore the test's consequential validity and washback. Specifically, at the micro level (i.e., in classrooms), studies should explore how washback influences teaching and learning, as well as how negative washback

[1] The revised HSK was introduced in November 2009. Compared with the old HSK, the major revisions of the new HSK include: fewer uncommon lexical items, new test formats, re-designed grading system, writing tasks with heavier weighting, and new oral tests. More information about the new HSK will be provided in Section 3.2.

effects can be minimized and how positive ones can be maximized. At the macro level (i.e., social), researchers should identify how test users utilize the HSK, in addition to how they interpret scores, how these scores affect their decision-making process, and how the decisions made may affect/impact the future of test-takers. Because, on the one hand, considering the HSK's goal "to support the interrelationship between teaching and testing, and to facilitate teaching and learning through testing [考教结合，以考促学、以考促教]", evidence of the HSK's washback on teaching and learning is crucial to test evaluation and validation. On the other hand, HSK may also be used for professional purposes, thus evidence from users/stakeholders at the macro level are also considered legitimate and necessary from the socio perspective.

1.2 The Purpose of the Study

To fill this research gap, this study adapted Bachman (2005) and Bachman & Palmer's (2010) assessment use argument (AUA) approach as a conceptual framework for investigating the HSK's consequential validity and washback effects with multiple stakeholders at micro and macro levels. An AUA is "an overall logical framework for linking assessment performance to use (decisions)" (Bachman, 2005, p.1). It consists of a set of claims: 1) The assessment record, which is the score or qualitative description obtained from the assessment; 2) an interpretation on whether the assessment is able to perform its intended evaluative goals; 3) decisions that are to be made based on the assessment record interpretation; and 4) the consequences of using the assessment and the subsequent decisions. The AUA, as either an assessment utilization argument or an assessment validity argument, can be used for more than just test development; it can provide a rationale and a set of procedures for justifying the intended uses of an assessment (refer to Section 2.6.2 for a detailed explanation of AUA). Articulating an AUA within the HSK

context can provide logical and methodological guidance for the current study. In addition, the current study can provide a specific context (i.e., serve as a case study), and add new features and bring broader understanding to the existing knowledge of AUAs. The purposes of the MMR study are threefold:

1) To reveal CSL teachers' and test-takers' perceptions of the HSK content, use, and impact; to explore any potential washback in CSL instruction/learning; and to explore the possible relationships between their perceptions of the test and their teaching/learning practices;

2) To identify the perceptions of other score users (in both academic and non-academic settings) concerning score interpretation, decisions made based on HSK levels/scores, and its intended and unintended uses; and

3) To explore the relationship between the PCI policy and any micro or macro level consequences of test (HSK) use.

1.3 Research Questions

The present study examined the HSK's consequential validity in terms of washback effects and how the test is used in the educational and societal context of CSL/CFL teaching and learning. The overarching research questions of the study are:

> *In the context of promoting the Chinese language internationally, what are the consequences of the HSK at the micro and macro levels? To what extent and in what ways does the HSK affect CSL teaching and learning?*

Employing a mixed methods sequential exploratory (MMSE) design, which will be explained in detail in Chapter 3, this dissertation includes 3 studies, which compose the 3 phases and/or components of the whole MMSE design. For clarity, the term 'mixed methods (MMR) study' will refer to the whole dissertation research (i.e., all three phases), while Study 1, Study 2,

and Study 3 respectively refer to the sub-studies in this dissertation. The research questions of Study 1 are:

- *What are the potential consequences of the HSK use from the test developer's perspective?*

- *What are the actual consequences of the HSK use from a multiple stakeholders' perspective at both the micro (classroom) and macro (society) levels?*

- *Is there any kind of relationship across the PCI Policy, TCSL, and any consequences of the HSK?*

To illuminate the relationships between CSL/CFL teachers' and learners' perceptions of the HSK's washback effect and their teaching/learning behaviors, there are two sets of research questions for Study 2. The first set addresses the test takers and the second set addresses the teachers.

- *What are HSK test takers' perceptions concerning the HSK content, use, and impact? What are their perceptions about whether the HSK score/level reflects their real proficiency? What are the relationships between these perceptions and their test preparation practices?*

- *What are CSL teachers' perceptions concerning HSK content, use and impact? How does the potential influence of the HSK manifest in their classroom practices? What are the relationships between these perceptions and their teaching practices?*

The research question of Study 3 is:

- *How do the HSK score users interpret and react to the consequences and the use of the HSK in both academic settings (i.e., higher education in China) and non-academic settings?*

1.4 Significance of the Dissertation (MMR) Study

This MMR study is the first attempt to investigate the HSK's consequences within the AUA framework. It is a timely response to the recent call in LT and presents a full-scale washback study, focusing on test consequences at both micro and macro levels. It not only attempts to provide findings that are applicable to pedagogical and methodological issues of CSL teaching and learning, but also intends to complement efforts made to broaden the understanding of the relationship between language education and language policy.

First, the study investigates the HSK's consequences by obtaining perspectives of multiple stakeholders (e.g., the test developer, test takers, teachers, and test users in academic and non-academic settings) on washback effects and the use of high-stakes tests. It also attempts to provide a more comprehensive model to enrich the existing washback literature by giving a better overall picture of not only how washback effects occur at the micro level, but also elaborates on how the test influences society at the macro level (McNamara, 2008). More specifically, at the micro level, the current researcher intends to explore the washback effects on teaching and learning behaviors (e.g., test preparation strategies) and beliefs (e.g., perspectives on the test and test use) with respect to the HSK; at the macro level, the study attempts to reveal the HSK's uses (e.g., values, score interpretation, decision-make related issues) in a broader social context.

Second, the MMR study intends to adapt the AUA framework, which could provide conceptual guidelines for an explicit and coherent linkage from test performance to interpretations, as well as from interpretations to uses. By collecting consequential evidence (e.g., intended consequences vs. actual consequences) in this type of coherent validation framework, this study attempts to explore the connections between test developers and test users in order to reveal any useful implications for test development. Researchers (e.g., Meyer, 2014)

argued that the HSK is not a proper measure of "communicative" language competency and may not reflect learners' actual proficiency level. There is therefore a need to analyze the new HSK according to its use and to make a more explicit labeling of its intended purpose.

Third, in the context of PCI, China is a rising global power and an increasing number of people are interested in learning the Chinese language and its culture. In such a context, this study's insights can help CSL teachers and learners reflect on their beliefs, strategies and methods, and trigger a deeper understanding of their teaching and learning. Such a reflection may help them adjust their teaching and learning strategies. In addition, beyond its pedagogical value, this research also has social significance for Chinese language learning and teaching, and the findings may shed light on the relationship between language assessment and language policy.

In sum, this study is significant in that its findings may not only provide pedagogical and methodological implications for CSL teaching and learning, but may also provide practical implications for the HSK's development, specifically in terms of enhancing its positive consequences and reducing its negative ones.

1.5 Organization of the Dissertation (MMR) Study

This dissertation consists of 8 chapters and its organization is as follows.

Chapter 1 offers an introduction to the MMR study including the rationale, purpose, research questions, significance, and organization of the study.

Chapter 2 provides a detailed overview of the literature. It begins with an overview of the consequential validity literature in educational assessment and language assessment, particularly on washback effects. Argument-based validation theories are then reviewed and the AUA framework is articulated in terms of the HSK; this provides a structural framework and

methodological guideline for this study. Following this, validity and washback studies on the HSK are reviewed. Finally, a summary of this chapter is provided, which highlights the research gap and the research problem this study addresses.

Chapter 3 starts with a general description of the educational, sociocultural, and historical context in which this study is situated. The rational of adopting the mixed-method research design is explained and the overall MMSE design of this MMR study is described.

Chapter 4, 5, and 6 respectively present the data collection, data analysis, findings, and discussions for Study 1, Study 2, and Study 3.

Building on the earlier chapters, Chapter 7 includes an overall discussion of the major findings from the 3 studies by synthesizing, integrating, and triangulating the results from different data sets generated from the AUA framework.

Chapter 8 summarizes the major findings and elaborates on their implications. The limitations in terms of technical difficulties as well as the overall scope of the MMR study are addressed. The chapter ends with a proposal of possible directions and recommendations for future research.

1.6 List of Definitions of Terms

Below is a list of definitions for terms and concepts frequently used in this dissertation. They were defined for the purposes of the study and should be interpreted as such within this dissertation.

Consequential validity: This concept is one dimension of validity. According to Messick (1989, 1996), when evaluating the validity of a test, it is essential to evaluate the intended and unintended consequences of its uses.

High-stakes tests: This term is used to describe tests that have major consequences for

students, teachers, and schools for informing major decisions, such as for university admission purposes. High-stakes tests can greatly influence the teaching and learning behaviors of those involved in the tests. Thus, even a minor change in the test can cause strong washback effects on the stake-holders (Shohamy, Donitsa & Ferman, 1996).

Washback/ Impact/ Consequence: In the field of LT, Hamp-Lyons (2000, p.586) argued that the term "washback" refers to "influences on teaching, teachers, and learning (including curriculum and materials)", while "impact" refers to the "wider influences" beyond the classroom. However, consequence is often used in educational assessment and defined in a broader sense; it can include any effect that a test may have on individuals, classroom practices, schools, and policies in the educational system or society. These terms are further defined in the next chapter.

HSK: *Hanyu Shuiping Kaoshi* (HSK), which is translated as the Chinese Proficiency Test, is a national standardized test designed to evaluate the Chinese proficiency of non-native Chinese speakers.

Chinese as a Second Language (CSL): CSL is the use of standardized Chinese (Mandarin) by speakers of other native languages. In recent years, the rising national power of China has led to a worldwide enthusiasm for learning the Chinese language. It also has a long history. Although reliable numbers concerning CSL learners worldwide are non-existent (Sun, 2009), there is evidence of a strong increase. In recent decades, China has helped 60,000 Chinese language teachers promote CSL internationally (Custer, 2010). Since the HSK was designed for non-native Chinese speakers as well as overseas Chinese, the researcher included Chinese as a Foreign Language (CFL) as well as Chinese as a Heritage Language (CHL) in this CSL definition.

Chinese Language Proficiency Scales for Speakers of Other Languages (Office of Chinese Language Proficiency Scales for Speakers of Other Languages, 2009), abbreviated as *Scales*, is an official document with guidelines for CSL teaching and learning. It was created to meet the needs of Chinese language teaching and learning worldwide. It was developed by language education and testing experts from over 80 universities in China and abroad. Designed for learners of CFL, the *Scales* provide a five-band holistic description of learners' ability to use the Chinese language for communication. It is regarded as an important measure of linguistic proficiency of Chinese language learners.

"Promoting Chinese Internationally" (PCI) policy: In the early 1990s, Hanban (中国国家对外汉语教学领导小组办公室[2]) published a book on promoting Chinese through policies and various agencies. The first World Chinese Conference, held in 2005, marked the recognition of the PCI as a national strategic policy (Li, 2012). This policy has motivated various decisions such as the opening of Confucius Institutes in over 100 countries, the training of Chinese language teachers, as well as the launch and reform of Chinese language proficiency tests.

1.7 Chapter Summary

This chapter opened with the rationale and the purpose of the MMR study. It then introduced the research questions and described the significance of the research. Finally, the chapter outlined the definitions of key terms used in this dissertation and provided a brief overview of the document's structure. In the next chapter, the literature on consequential

[2] Hanban is the colloquial abbreviation for the Chinese National Office for Teaching Chinese as a Foreign Language. It is a non-government and non-profit organization affiliated with the Ministry of Education of the People's Republic of China. Hanban is most notable for its Confucius Institute program, but it also sponsors the Chinese Bridge competition, which is a competition in Chinese language proficiency for non-native speakers. According to its mission statement (Hanban, 2014), Hanban is committed to developing Chinese language and culture teaching resources and making its services available worldwide, meeting the demands of overseas Chinese learners to the utmost degree, and contributing to global cultural diversity and harmony. Generally, Hanban is charged with cultivating knowledge and interest in the Chinese language and culture in non-Chinese speaking countries.

validity, washback studies, and argument-based validation theories will be reviewed.

Chapter 2 Literature Review

2.1 Introduction

To address the research questions in Chapter 1, the present MMR study focused on the consequences of the HSK in Chinese society and higher education in the PCI context by employing the argument-based validation approach. This chapter reviews literature on 1) consequential validity in educational assessment and language assessment, which reconceptualizes validity and washback and investigates the relationship between them; 2) argument-based validation theories, and a framework articulating AUA for the HSK, which provides a methodological guideline for the MMR study; and 3) validity and washback studies on the HSK, which include studies conducted in both English and Chinese. A summary is also provided to highlight the current literature gap and the research problems the MMR study addresses.

2.2 Consequential Validity and Washback

This section reviews the evolution and current issues with validity and washback.

2.2.1 The initial understanding of validity

The concept of validity has evolved over time. In early discussion, validity was defined as "whether a test really measures what it purports to measure" (Kelly, 1927). In the past three decades, the validity theory in educational evaluation and LT has been broadened by the inclusion of the consequences, impact, and uses of tests (e.g., Messick, 1989, 1996; Kane, 2002, 2006). Specifically, there is an increasing recognition that test validity is affected by the under-representation of test constructs and "construct-irrelevant variance" with attending factors, such as educational and social consequences, test-taking experiences, and test uses for different purposes (Haladyna & Downing, 2004). Weir (2005) defined validity as the extent to which a

test can be shown to produce scores that accurately reflect candidate's level of language knowledge or skills. He also stated that washback, also called washback validity, is central to the concept of consequential validity. Furthermore, Moss, Girard & Haniford (2006) argued that validation studies must include multiple stakeholder perspectives in order to expose sources of evidence that would otherwise invalidate test inferences and uses. Numerous empirical studies have linked test validity to its use and consequences (e.g., Cheng, Klinger, & Zheng, 2007; Sun, 2016; Wang, H., 2010; Xie, 2010).

2.2.2 The emerging concept of washback

Washback has also been referred to as test impact or test consequences; however, these terms all refer to facets of the same phenomenon in education, regardless of the stakes of the tests, ranges of disciplines, and backgrounds of the test-takers. In the field of LT, the work of Alderson and Wall (1993) is still considered as a landmark in shaping the construct of washback studies. In their paper, they explored the potentially positive and negative relationships between testing, teaching, and learning in order to address the question of "Does washback exist?" (p. 115). They proposed 15 hypotheses regarding the potential influences of language testing on various aspects of language teaching and learning. This ultimately provided the fundamental guidelines for washback studies over the next two decades.

The definitions of washback have evolved over the years. In the early stages, some scholars believe that washback effects mainly occur in classrooms. For example, Hughes (2003) defined what he referred to as "backwash," which was "the effect of testing on teaching and learning" (p. i). He further asserted that testing can either have a beneficial or a harmful effect on teaching and learning. Prodromou (1995) additionally stated "the backwash effect can be defined as the direct or indirect effect of examinations on teaching methods" (p.13). Later, Wall (1997)

distinguished between test washback and test impact in terms of the scope of their effects (and this distinction became generally accepted henceforward). Under this perspective, "washback" is referred to the effect of tests on teaching and learning, while the "impact" encompasses a broader meaning than washback, as it includes any effects that a test may have on individuals, policies, or practices within the classroom, school, educational system, or society as a whole. Similarly, Hamp-Lyons (1997) and McNamara (2000) criticized the term washback as being too narrow; they pointed out that general education and educational measurements tend to employ the more general term of impact, which includes effects beyond the classroom as well as effects on the educational system and society as a whole. Later, Shohamy, Donitsa-Schmidt, and Ferman (1996) pointed out that the degree of a test's impact is often influenced by several contextual factors, such as the status of the subject matter tested, the stakes of the test, and the use of the test. Moreover, Bachman and Palmer (2010) defined washback as "the broad effects of an assessment on learning and instruction in an educational system" (p. 109). Cheng (2013) further stated that test consequences are influenced by the ideological, social, and political milieu surrounding particular educational systems.

2.2.3 Debates in operationalizing validity and washback

Messick (1994, 1996) regarded washback as an "instance of the consequential aspect of construct validity" (p. 242), which links washback to validity, and covers elements of test use, impact, and the interpretation of results. Bailey (1996) argued that in terms of validity, any test can have positive or negative washback effects, depending on whether it impedes or promotes the accomplishment of learners' and/or programs' educational goals. She focused on the specificity of this phenomenon, which could induce differential impact on different test stakeholders, as they observe how a test serves its purposes and uses from their own points of

view. Bachman (2005) and Bachman and Palmer (2010) proposed an argumentative validity framework with a set of principles and procedures for linking test scores and score-based inferences to assessment use and consequences. Within this Assessment Use Argument (AUA) framework, washback effects were seen as the test's impacts on individuals (teachers and students), educational systems, and society. More recently, Xie (2010) argued that washback is not unrelated to test validity. She presented a conceptual model of the relationships between washback and test validity, which is reproduced below in Figure 2.1. Adapting the argumentative validation approach and Messick's consequential aspect of construct validity theory, her model stated that test design may influence test preparation, which can in turn impact test performance. The validity of score interpretation can be appraised through evaluating test design and test performance, and this validity can affect test use. The use (or misuse) of an assessment may also affect stakeholders' perceptions of the test and can ultimately result in different test preparation behaviors. This model systematically linked washback and test validity by examining the relationships among test design, test preparation, and their performance; however, it only accounted for test-takers' perspectives, and did not address other stakeholders' (e.g., teachers) perspectives or a variety of mediating factors both inside and outside classrooms.

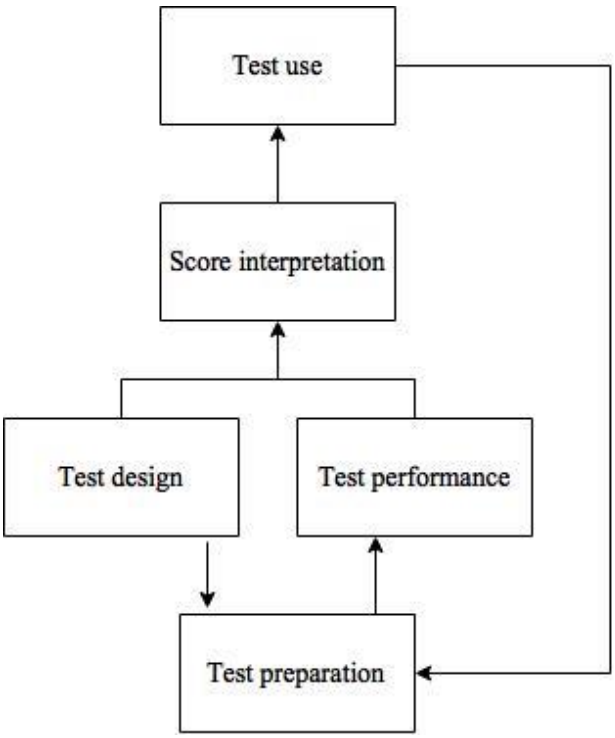


Figure 2.1. Conceptual model: Relationships between washback and test validity (p.68, Xie, 2010)

In sum, the terms and definitions of washback/impact/consequence at the micro and macro levels have evolved over the years and have highlighted several concerns. More specifically, the relationship between washback and test validity needs to be clarified. Hence, validation studies must include multiple stakeholders' perspectives in order to expose evidence that would otherwise invalidate test inferences and uses (Moss et al., 2006).

2.3 The Nature of Washback and Factors Mediating its Process

Several important washback works were published in the past two decades. In 2004, Cheng, Watanabe, and Curtis published their seminal work on washback studies. In this book, the researchers drew on a range of significant language washback studies. The authors also suggested directions for further research in order to respond to the question, "what does washback look like?" (p. ix), which was a step beyond the question "does washback exist?" as

posed by Alderson and Wall in the early 1990s (as cited in Cheng, 2014). In this section, major sources that discuss the nature and the mediating factors of washback effects are discussed. Gaps and future directions in washback studies are also provided.

Although most LT researchers agree on the existence and importance of washback, there still is considerable variation in opinions about how washback works (Bailey, 1996) and whether its effects are positive and negative, or intended and unintended. Due to the potential and actual misuse of certain tests, most discussions of washback have emphasized assessments' assumed negative effects on the quality of teaching and learning (e.g., Andrews et al., 2002; Cheng, 2004; Qi, 2007; Wang, J., 2010); this is especially true for traditional large-scale high-stakes examinations. For example, Qi (2007) conducted a washback study that focused on the writing task in the National Matriculation English Test (NMET) in China. The results revealed the anticipated and actual effects of washback on secondary school teaching. More specifically, while the test did increase the frequency of writing practice in schools, the pedagogical objectives of these practices were not in line with the test creators' intention (i.e., to boost students' communicative abilities). Both teachers and learners neglected the communicative context of writing, as they chose to focus instead on ways to increase test performance (e.g., considering the assumed preferences of the markers). This urge to raise scores in a real test situation suggested that high stakes assessments do not lead to positive improvements in teaching and learning.

There are other researchers who believe assessments can have a more positive impact, especially when a test has been introduced, modified, revised, or improved in order to exert a positive influence on teaching and learning. For example, positive washback can be fostered by providing teachers with ongoing training and guidance on assessment and instructional practices

(e.g., Davison, 2008; Muñoz & Álvarez, 2010; Turner, 2009). Furthermore, the studies on the development of Quebec's ESL high school exit exam (Turner, 2005, 2009) are significant in this aspect, since they demonstrated the exam's positive washback effects in terms of improving instruction quality. Nevertheless, many researchers have argued that it is difficult to determine whether the effects of tests are positive or negative. For instance, Alderson and Wall (1993) stated that in their study, there was no clear relationship between tests and their effects on classroom practices. Green (2013) later argued that the variables shaping washback effects are complex, which make the effects highly variable and contextualized.

As more research is conducted, the understanding of washback in the LT community will continue to improve, particularly in terms of how teachers are affected by the phenomenon. Most research findings have shown that high-stakes tests have a significant impact on L2 teaching such that the tests altered the teachers' pedagogical methods (e.g., Alderson & Hamp-Lyons, 1996; Cheng, 1997, 2005). In addition, the results of some studies have suggested that assessments can change how teachers administer tests (Wall & Alderson, 1993) and how high-stakes exams may increase teachers' level of anxiety and fear (Ferman, 2004). Whereas much of the research has investigated how washback affected teaching, "less emphasis has been given to learners" (Watanabe, 2004, p. 22). There are some studies, however, that focus on learners' perspective; these include Shih's (2007) and Cheng's (1997, 1998) work on the GEPT[3] and HKCEE[4], respectively. In Cheng's early study (1997), she examined HKCEE's washback on English learning through a survey. The results indicated that the test was the most significant

[3]The General English Proficiency Test (GEPT, Chinese: 全民英語能力分級檢定測驗, or 全民英檢 for short) is a English language proficiency test that was commissioned by Taiwan's Ministry of Education.
[4]Hong Kong Certificate of Education Examination (HKCEE, Chinese：香港中學會考) was a standardised examination between 1974 and 2011. It was given to students at the end of their five-year secondary education.

factor involved in motivating students to learn English, more so than their future career plans. In a later study, Cheng (1998) found that although students changed their learning beliefs after the test's content was changed, learners still reported that they retained their original learning processes, learning strategies, and individual motivation to learn English. Additionally, Shih (2007) investigated GEPT's washback on English learning in Taiwan and found that existing theories do not fully explain the washback of this test on the educational context; consequently, a new tentative washback framework was proposed. This model included various factors that helped explain the complexity of the washback effects. It also elucidated how tests influence students' learning, especially in East Asian contexts.[5]

In addition, much of the literature that examined the impact and washback of tests was designed to inform top-down educational reform and thus focused primarily on the educational dimension. Recent studies, however, have begun to examine how tests affect the social dimension, as well as the direct and indirect effects of language tests on language policies within the AUA framework. The majority of previous studies on consequential validity have been conducted from the perspectives of test designers (Bachman, 2000). In fact, these studies rarely considered both the cognitive dimension of language testing (e.g., motivation, anxiety, attitude) and its social dimension (e.g., potential test uses/misuses within a context) (Chalhoub-Deville, 2003; Cheng, 2008). Validity evidence from the test-takers' perspectives is still limited in language assessment. Even fewer studies have included the perspectives of test users, such as that of administrators and employers. The challenge that remains is how data collected from multiple stakeholders can be used to justify test score use. Tests clearly have an impact and strong washback on many areas beyond the classroom (Shohamy, 2007), but more work is

[5] EFL education in China, Korea, and Japan is remarkably similar.

needed, particularly for understanding the multi-dimensional impact of washback and for establishing the link between test validation and test use. As Cheng (2013) advocated, studies need "to go beyond the micro-level of the classroom (washback) to the macro-level of society (impact), to analyze the social factors that lead to assessment practices" (p. 12).

Moreover, there are generally two types of designs in empirical washback studies: 1) examining the effects and consequences of washback on teaching/learning/textbooks by comparing exam preparation courses and regular (non-exam) courses or different teachers in the same setting (e.g., Greene, 2007; Read & Hayes, 2003; Watanabe, 1996, 2004); and 2) collecting and interpreting complex data through mixed methods, longitudinal designs, and proactive participatory approaches (e.g., Cheng, 1997, 1998; Turner, 2005; Tan, 2009). Although the methods, designs, and contexts vary from study to study, most investigated the effects of high-stakes and/or large-scale exams on educational reform/innovation. Along with the increasing interest in learning-oriented assessment (LOA) (Purpura, 2008; Turner & Purpura, 2015) and classroom-based assessment (CBA) (Turner, 2012), it is necessary to investigate the complex relationships between high-stakes tests and CBA from an LOA perspective. If the information gathered from CBA is used appropriately, then it can provide relevant and timely feedback to teachers, help them monitor students' progress, and better support future learning. It might also be a simpler way to minimize negative washback effects and to meet the anticipated goals of the test.

Drawing on some major theoretical and empirical washback studies mentioned above (e.g., Alderson & Hamp-Lyon, 1996; Brown, 1997; Cheng, 2004; Shohamy et al., 1996; Wall, 1997; Wall, 1997; Watanabe, 2004; Shih, 2007; Wang, S., 2013), various factors affecting the process of washback can be identified, such as:

1) Test factors (e.g., test validity, stakes of the test, status of the test within the educational setting, test methods, test content, purpose of the test, skills tested);
2) Individual factors (e.g., teachers' beliefs, teaching methods, teachers' educational backgrounds, students' test-taking strategies, test-takers' nationalities, motivation of the test-takers); and
3) Context factors (e.g., micro-level: the classroom setting in which the test preparation is being carried out; macro-level: the society where the test is used, educational policies).

Overall, the nature of washback is dynamic, and factors associated with a test are complex phenomena that can influence teaching and learning. The washback effects, which are always contingent on the context, could be positive, negative, both, or neither. More empirical studies on washback are still needed to explore how it influences teaching and learning and how negative effects can be minimized while maximizing the positive ones.

2.4 Reconsidering Washback: From the View of Test Use

In 210 BC, the civil service examination in China pointed to the use of tests for selecting qualified workers (Madaus & O'Dwyer, 1999). Thousands of years later, contemporary social theory offers a rich array of conceptual frameworks that can conceptualize the social context and the role of test use in such settings. Gipps (1999, 2002) stated that a test is a social activity with various roles, which can only be understood if the social, cultural, economic, and political contexts are also taken into consideration. Kellaghan and Greaney (2001) concurred that assessments are used for multiple purposes such as describing the students' progress and identifying learning problems; guiding students in their choice of courses or vocational options; motivating them by providing goals and targets; certifying that they have attained an expected level of competence; and admitting students into higher education institutions.

As previously mentioned in this chapter, since the 1990s, the application of validity theory in educational evaluation contexts has established grounds for the inclusion of test impact and use within validation studies (Messick, 1989, 1996; Kane, 2002, 2006). The details of the argument-based validation frameworks involving tests' consequence as a component will be presented in Section 2.5. It is critical that the link between test validity and the consequences of test use are established from multiple stakeholders' perspectives (Cheng, 2014). Furthermore, the use (or misuse) of test scores and the values/stakes attached to a test should be investigated within all relevant contexts (i.e., in society, in classrooms). To better understand the washback effects associated with test use, it is important to review the philosophical, theoretical, and practical frameworks/models of test consequences. Three influential theoretical frameworks are introduced below for this purpose.

Critical language testing

The washback effects have been increasingly discussed from the point of view of critical language testing (CLT). Proposed by Shohamy (1998), CLT developed out of a combination of critical pedagogy and critical applied linguistics. Under this perspective, tests are seen as powerful tools that are embedded in social and political contexts. They are thus related to goals, effects, and consequences and are open to interpretations (Shohamy, 2001). From tools used to measure linguistic knowledge, language tests are now viewed as instruments connected to and embedded in political, social, and educational contexts. Accordingly, the quality of tests is not judged merely by their psychometric traits but rather in relation to their impact, ethicality, fairness, values, and consequences. Shohamy (2001) introduced democratic assessment principles to CLT, which included the need for testers to take a greater responsibility for how their tests were being used. By pointing out the uses of tests to users and the public at large, these

principles provide a critique of test consequence and make stakeholders more reflective of the social dimension of test use.

As shown in Figure 2.2, a number of strategies and mechanisms are used by central authorities to create, perpetuate, and manipulate language policies (Shohamy, 2006). Representing one of such covert mechanisms, tests can be leveraged to influence society and can play a major role in implementing and introducing language policies.

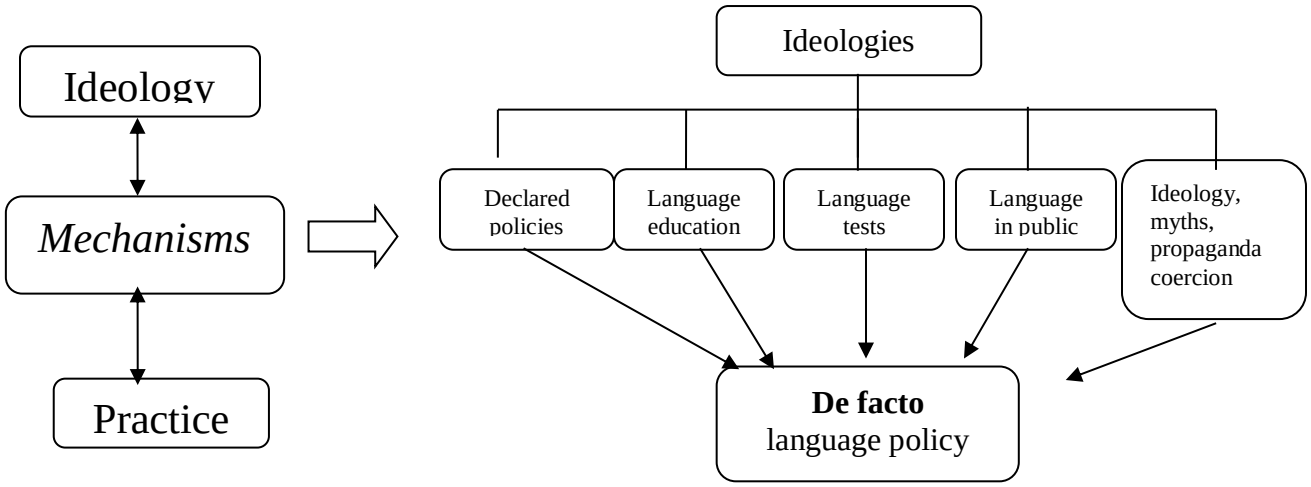


Figure 2.2. Mechanisms within ideology and practice and language tests as a mechanism affecting language policy (Figure developed based on Shohamy, 2006, pp. 53-54)

The second part of Figure 2 shows how tests can then be used as mediators and mechanisms for creating language education policies to control and manipulate *de facto* language policies. Shohamy (2007) further explained that LT should mediate ideologies and practices in more open, democratic, and negotiable ways; this will help prevent the use of tests as powerful mechanisms to impose influential policies that have no empirical basis. Therefore, CLT broadens the field of LT by engaging in a wider sphere of social dialogue and debate on the forms and practices of political uses of language tests and their relation to language teaching and learning.

"Participants-Processes and Products" (PPP) model

Hughes (1993) discussed the mechanisms through which washback operates. He stated that:

> The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work. (p. 2)

Here, Hughes stressed the participants' perceptions and how these factors affected what they did. Bailey (1996) later synthesized Hughes' ideas and proposed the "Participants-Processes and Products" (PPP) washback model to explore the complicated relationship among tests, stakeholders, and various educational processes (see Figure 2.3). She categorized the participants and products into two groups: the learners and the improved learning of the target construct (language proficiency in this case); and the other products that can promote students' learning, such as new curriculums, materials, improved teaching methods, and valuable research findings.

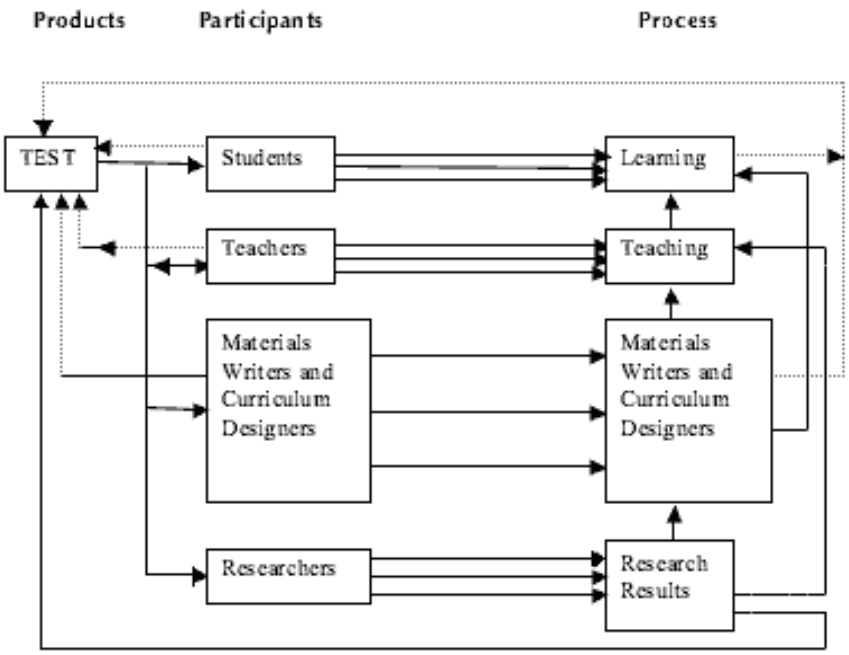


Figure 2.3. Basic model of washback - "PPP" (figure adopted from Bailey, 1996, p. 264)

From this model (see Figure 3), it can be determined that a test has direct influences on the participants who are involved in various processes, and these processes result in products specific to each type of participant. Additionally, unlike the linear relationship between tests and teaching/learning posited by Alderson and Wall (1993), this model shows the multiple interactions between various components.

The hybrid model of English language teaching innovation in Japan

Tests continue to be used as a vehicle for curriculum innovation (Andrews, 2004). Building on the ideas of many researchers in the field of educational innovation studies, Henrichsen (1989) examined data from efforts to reform the Japanese English Language Teaching (ELT) system, which was carried out by the English Language Exploratory Committee over a period of twelve years. He proposed that those who intended to introduce educational innovations must be aware of factors at three stages of the diffusion and innovation processes.

The hybrid model of the diffusion/implementation process (shown in Figure 2.4 below) showed that the awareness and evaluation of those expected to react to the innovation (in this case, teachers reacting to the new test) are influenced by many factors, including the channels of communication used, the characteristics of the innovation, and various features in the educational context. This model provided planners of other reform campaigns with an understanding of what the important factors were, how they were related, and how to deal with them. It also provided an overall idea of all the stakeholders and procedures before, during, and after any innovation in education. This hybrid model can account for the three components (i.e., participants, process, and product) that Hughes proposed. It also included other factors such as the educational system, and the teaching and learning practices within a system. This model of English language teaching innovation in Japan can also act as a reference for improving TCSL and for implementing the new HSK abroad, which are important factors to be considered in this study.

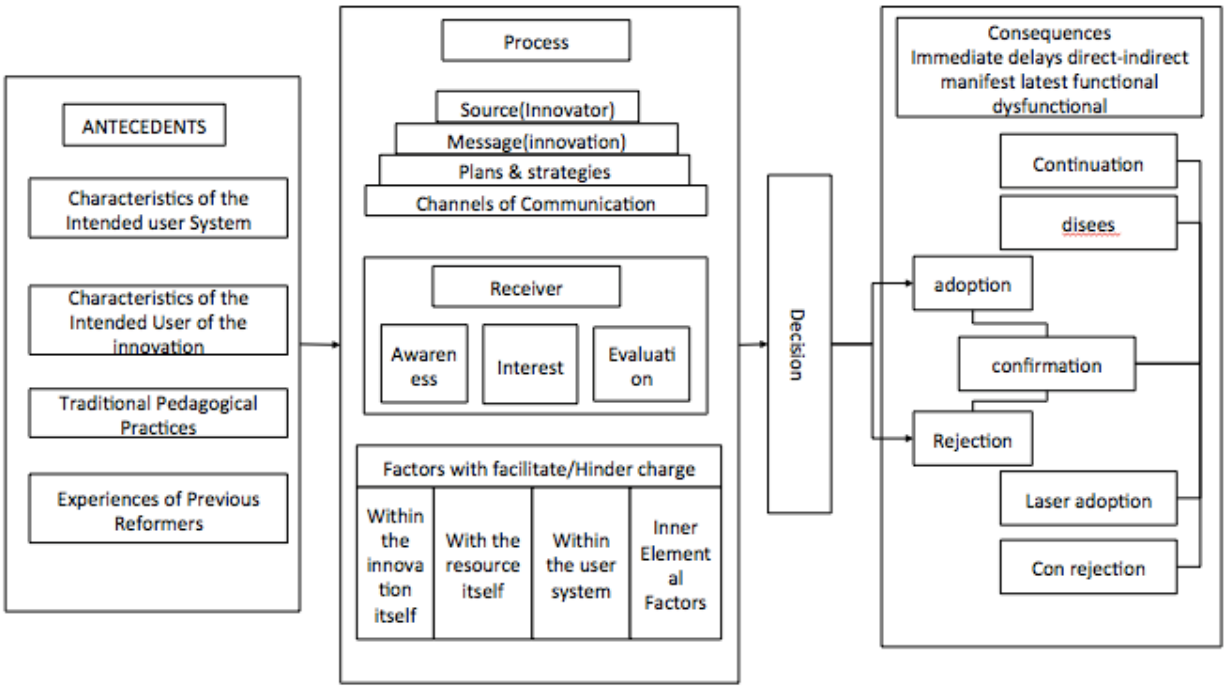


Figure 2.4. The Hybrid Model (Henrichsen, 1989, p.80)

Across these frameworks and models from the perspective of language use, there has been a growing emphasis on collecting washback evidence from multiple stakeholders and using multiple methods in a more comprehensive analysis.

2.5 Overview of the Research Methodologies Employed in Washback Research

During the past three decades, the methodologies employed in washback/impact/consequence studies have evolved. In the examination of research methodologies employed in these studies, the differences between the methodological approaches (i.e., quantitative, qualitative, and MMR), and the pros and cons of each approach are discussed below. At the end of this section, the research tools and designs employed in washback studies to date are summarized. It should be noted that because the tools and designs are closely related to the research methods, the questions addressed in the previous sections will be revisited and expanded on in terms of the advantages and disadvantages of each tool/technique.

Quantitative research has been actively employed in applied linguistics in order to test theoretical hypotheses (e.g., the washback hypothesis). Some research tools for quantitative data collection in the field include tests, interviews, surveys, verbal reports, prompted production, prompted responses, and grammar analysis techniques (Gass & Mackey, 2007). These methods can be found in washback studies. Specifically, research methods were largely drawn from the findings of teachers' and/or students' self-report questionnaires (e.g., Andrew & Fullilove, 1994; Herman & Golan, 1991; Hughes, 1988; Shohamy, 1992). For instance, Xie & Andrew (2013) conducted a washback study to examine the relationships among test preparation practices, beliefs concerning test use, and test performance through pre-/post-tests and a test preparation questionnaire of 1,003 test-takers. Employing multiple regression and structural equation modeling methods, the researchers found that: 1) test preparation did improve test scores; and 2) test preparation patterns were affected by the perceived effectiveness of different strategies. In this case, the quantitative approach was suitable for the large-scale study and to explore relationships among multiple variables. However, there are several limitations to using a purely quantitative design. Firstly, it was not possible for the authors to gain deeper insight into the phenomenon (e.g., determine why the preparatory practices did not improve test results). Some researchers realized that follow-up interviews with the survey respondents should have been conducted in order to obtain "a fuller explanation, understanding, and validation of these survey findings" (Cheng et al., 2004, p. 380). Additionally, information from numerical data can be overly abstract or too general for direct application to specific local situations, contexts, and individuals (Patton, 2002; Tashakkori & Teddlie, 2010).

As opposed to quantitative methods, qualitative research has been regarded as a useful approach for investigating complex natural/social phenomena (Patton, 2002). In the realm of LT,

qualitative methods have been increasingly used to examine issues related to test consequential validity in classroom contexts, such as discourse analysis, introspection, and ethnographic methods (Lumley & Brown, 2005). More specifically, qualitative methods in washback studies have been adopted to provide a contextualized perspective and to discuss issues concerning LT practices. Although discourse analysis is not a monolithic method (i.e., it includes different approaches within its framework), the usefulness of discourse analysis has been employed in relation to a wide range of assessment issues (Lumley & Brown, 2005). For example, Glover (2014) employed discourse analysis to investigate the influence of examinations on teachers' talk by synthesizing the discourse data recorded in class, field notes, and teacher reports. Although she portrayed the method as a salient tool for exploring the difference between teachers' attitude and the way they actually teach, the drawback of this method is that it is extremely time-consuming. To avoid this limitation, some researchers have suggested using checklists instead (O'Sullivan, Weir & Saville, 2002).

Introspection is also often adopted as a qualitative approach. When used in the format of verbal reports or diaries, it allows researchers to situate a particular phenomenon or personal experience in a specific context in a rich and detailed manner. Gosa (2009), for instance, conducted the first diary study to investigate unobservable factors that may affect the presence or absence of washback on students. However, an analysis of self-reported narratives may result in bias and subjective distortion (McLeod, 2009).

Another qualitative approach, ethnography, has been gaining attention among LT researchers as it "elicits phenomenological data that represent the worldview of the participants being investigated and participants' constructs" (Watanabe, 2004, p. 22). Sadeghi (2014) conducted an interpretive ethnographic case study using observations and field notes to

understand how high-stakes testing affected teachers' instruction. The teachers' performances were observed and documented according to the University of Cambridge Observation Scheme for four TOEFL and IELTS preparation courses. Since "washback is not as simplistic as it may seem" (p.18), this approach was used to examine and illustrate the complexity and uniqueness of each teachers' practices over a period of time. However, since the ethnographer/researcher was also one of the teachers, the results may have been influenced by the researcher's personal biases and idiosyncrasies (Patton, 2002; Tashakkori & Teddlie, 2010). Thus, although qualitative research gathers in-depth understandings of human behaviors and the reasons that govern such behaviors, the constraints and drawbacks of this approach cannot be ignored.

Compared to earlier studies, which simply adopted the mono-method and used a single data source, more recent washback studies have utilized a multi-method approach and often draw on various data sets. The multi-methods[6] approach includes the use of more than one method of data collection. A well-known Sri Lankan washback study that focused on the O-Level Examination led the way in terms of employing multiple methods (Alderson & Wall, 1993; Wall & Alderson, 1993; Wall, 1996, 1997, 1999). Its longitudinal design differed from other studies in that it included a baseline study, questionnaires to teachers and teacher advisers, group interviews with teachers, and document analyses. At the time, it was the most comprehensive and sophisticated washback study that had been conducted in LT. Thus, the Sri Lankan study can be regarded as the most essential empirical washback work; not only did it extend the theoretical basis and vision of washback research, but it also set out a research agenda. Additionally, Wall

[6] Multi-method research could be qualitatively driven design (QUAL + QUAL), quantitatively driven design (QUAN + QUAN), interactive or equal status design (QUAL + QUAN), or MMR. MMR is more specific that that it includes the mixing of qualitative and quantitative data, methods, methodologies, and/or paradigms in a research study or set of related studies.

and Alderson (1993) highlighted the complex nature of the washback effects and stressed the importance of incorporating various methods to investigate the existing washback phenomenon. They advocated the importance of complementing classroom observations with teacher interviews, questionnaires, and analyses of materials (specifically, test materials and what teachers had prepared for classes), and helped expand the range of instruments available for this type of research. Furthermore, other researchers (e.g., Alderson & Hamp-Lyons, 1996; Wall, 1996) incorporated a classroom observation component in their studies in order to fully comprehend the behaviors of students and teachers in the classroom under the influence of a high-stakes examination. Through observing teaching and learning processes in the classroom, researchers were able to investigate: 1) the ways in which tests influenced teaching content and teachers' delivery of lessons; and 2) the amount of time that was spent by students in preparing for a test. Without observations, researchers would not have been aware of the inconsistencies between teachers' reports and their actual practices, and between what students perceived they studied and what they actually addressed in the test. The shift subsequently motivated a substantial number of evidence-based and observational washback studies (e.g., Alderson & Hamp-Lyons, 1996; Burrow, 2004; Cheng, 1997, 1998; Qi, 2004; Shohamy, Donitsa-Schmidt & Ferman, 1996; Turner, 2009; Watanabe, 1996).

Table 2.1 presents a summary of the major multi-methods washback studies in international education contexts; it also provides detailed information about the different types of data collection methods and participants involved in each study. However, this is not meant to be an exhaustive list of all studies, but is instead a set of representative studies. Although the methods, designs, and contexts vary from study to study, there are some shared features in terms of research method trends, such that an increasing number of washback studies has collected and

interpreted complex data through mixed methods, longitudinal designs, and proactive participatory approaches, and has involved various research participants.

Table 2.1

Summary of Major Washback Studies on Teaching and Learning

Studies	**Exams Studied**	**Research Contexts**	**Data Collection Methods**	**Participants**
Alderson & Hamp-Lyons (1996)	TOEFL	USA	Interviews Classroom observations	Teachers
Alderson & Wall (1993), Wall & Alderson (1993), Wall (1996, 1997, 1999, 2000, 2005)	Sri Lankan O-Level Examination	Sri Lanka	A baseline investigation, questionnaires, document analysis, classroom observations, group interviews	Teachers, Students
Andrew (1994)	Hong Kong Use of English Exams	Hong Kong	Questionnaires	Test-designers, Teachers
Cheng (1997, 1998, 1999, 2001, 2003, 2004)	The Hong Kong Certificate of Education Examination in English	Hong Kong	Questionnaires Interviews Classroom observations	Test developers, Textbook writers, Teachers, Students
Greene (2006, 2007)	IELTS	UK	Pre, post tests Questionnaires, Classroom observations, Interviews	Teachers, Students, Department heads, Family members
Shih (2007)	General English Proficiency Test	Taiwan	Documents analysis, Interviews, Observations	Students
Shohamy (1991, 1992, 1993, 2001, 2006), Shohamy et al. (1996)	Arabic test, EFL oral test, Education Examination in English	Israel	Teaching material analysis, Interviews, Classroom observations	Teachers, Students, Inspectors
Turner (2001, 2006, 2009)	Quebec Secondary Five ESL Speaking Exam	Quebec, Canada	Questionnaires, Interviews, Classroom observations	Teachers, Students

2.6 Argument-based Validation theories

In the previous section, three conceptual frameworks (i.e., the PPP washback model, the CLT model, and the hybrid model of English teaching innovation in Japan) were discussed to support the core framework of the MMR study. This section subsequently focuses on the development of Kane's interpretative argument as well as Bachman and Palmer's Assessment Use Argument (AUA). A rationale for adopting argument-based approaches in this MMR study is also discussed.

2.6.1 Kane's Interpretative Argument

Aiming to provide "clear guidance on how to validate specific interpretations and uses of measurements," Kane (2006) developed "a pragmatic approach to validation" (p.18). In other words, this was an argument-based approach that drew inferences from test scores by gathering and disseminating evidence supporting intended score interpretations.

Kane's interpretive argument approach was based largely on Toulmin's (1958, 2003) model of argumentation. This approach consisted mainly of a *claim*, for example, the interpretation of test scores. This claim was in turn based on data, scores or other manifestations of performance of a test taker. Then, the relation between the claim and the data must be justified by a *warrant*, and the warrant itself must be supported by the *backing*, which is the empirical data of an investigation. In this argumentative chain, counterproposals against the argument can be brought up by *rebuttals*, which are meant to challenge or weaken the argument. Kane's approach included two steps, 1) to build an *interpretive argument* which involves an argumentative chain; and 2) to build a *validity argument*, in which validation studies and research data are used to rebut or to warrant this argumentative chain. Figure 2.5 illustrates the chain of inferences in this interpretative argument.

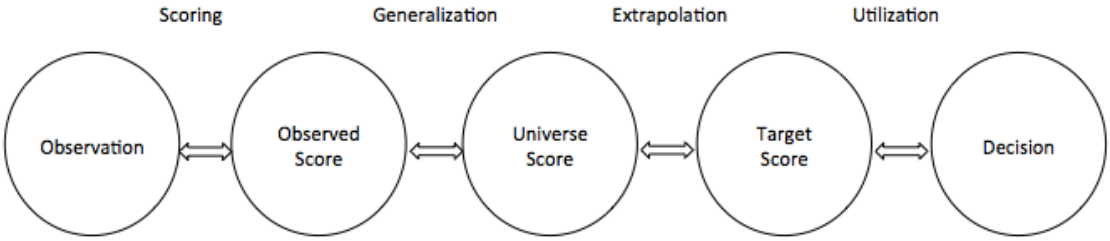


Figure 2.5. Links in an interpretative argument (Kane, 2004, p. 167)

In this model, four inferences including scoring, generalization, extrapolation, and utilization can help researchers link test performance observation to interpretations and test score use. Extrapolation inference goes beyond testing contexts and looks at test takers' target scores or actual performance in real life situations. Utilization inference refers to the use of target score either in terms of construct understanding or decision-making. The prominent advance in Kane's approach was how actual scores were used and how test consequences were considered as part of validity. However, the question of how to investigate test uses and consequences still remain, as Kane failed to develop a well-defined methodology.

2.6.2 Bachman's Assessment Use Argument

Bachman (2005) and Bachman and Palmer (2010) expanded upon the argument-based approach and proposed the assessment use argument (AUA) framework. This section will first touch upon Bachman's early work before highlighting the advances and merits of an AUA.

Influenced by Messick's validity theory, in his book Bachman (1990) outlined the implications of validity as a unitary concept pertaining to test interpretations and use; he also emphasized that the inferences made on the basis of test scores and their use are the object of validation rather than the tests themselves. Based on this work, Bachman and Palmer (1996) later

proposed an overarching notion of test usefulness as a manageable validation framework, which encompassed five qualities: reliability, construct validity, authenticity, interactiveness, and practicality. While this framework elicited considerations regarding construct validity and the impact of usefulness, it did not explicitly establish a link between validity and test use. Indeed, as McNamara & Roever (2006) pointed out, the cost of manageability comes with a certain loss of theoretical coherence.

In order to fill the gap in the literature concerning the lack of comprehensive methods that allow for deeper analyses, Bachman (2004, 2005; Bachman & Palmer, 2010) utilized Toulmin's (2003) and Kane's (2002, 2006) approaches as a basis for articulating an AUA. The AUA is a conceptual framework that consists of a series of inferences that link test taker performance "to a claim about assessment records, to a claim about interpretations, to a claim about decision[s], and to a claim about intended consequences, along with warrants and backing to support these claims" (p. 103). The relationships among assessment, measurement, tests, and test uses are illustrated in Figure 2.6.

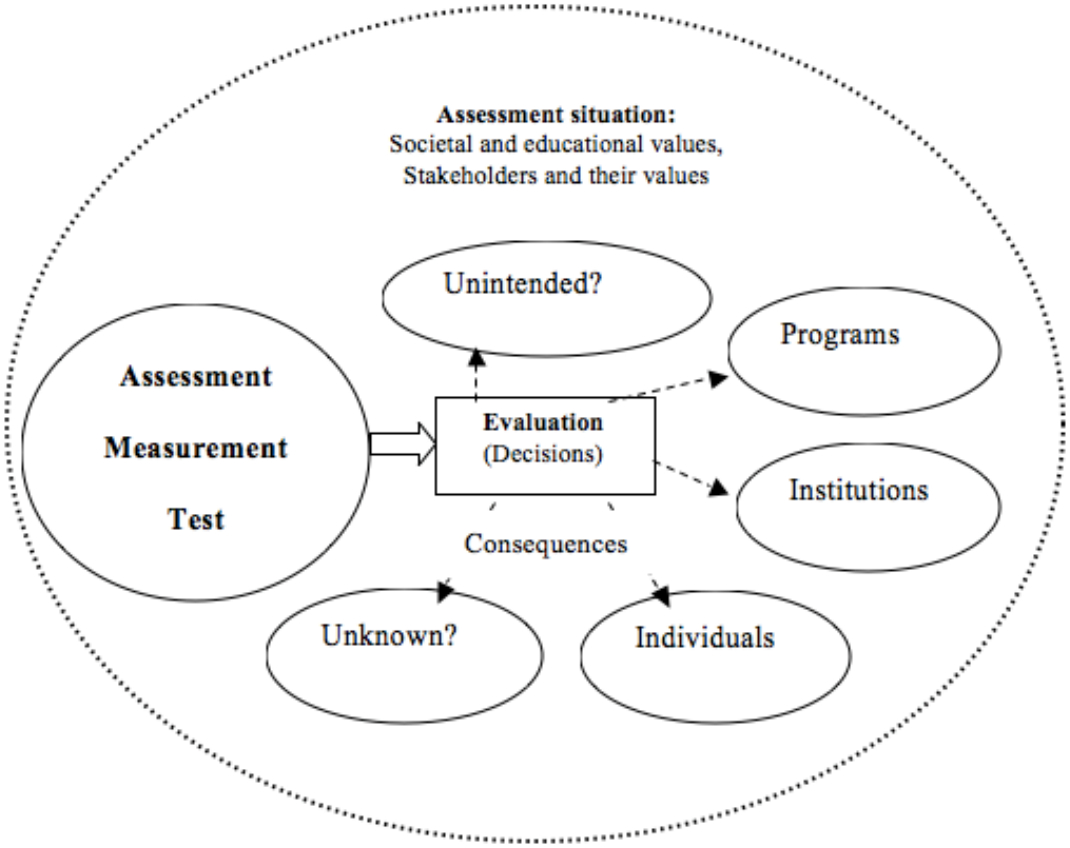


Figure 2.6. Relationships between assessments/measurements/tests, their use for evaluation, and the consequences of assessment use (Bachman & Palmer 2010, p.22)

According to Bachman & Palmer (2010), the AUA includes two parts 1) a utilization argument that links an interpretation to a decision and 2) a validity argument that links assessment performance to interpretation. The current MMR study places greater emphasis on the first part. Figure 2.7 presents an overview of the AUA framework, illustrating its structure, elements, links, and qualities.

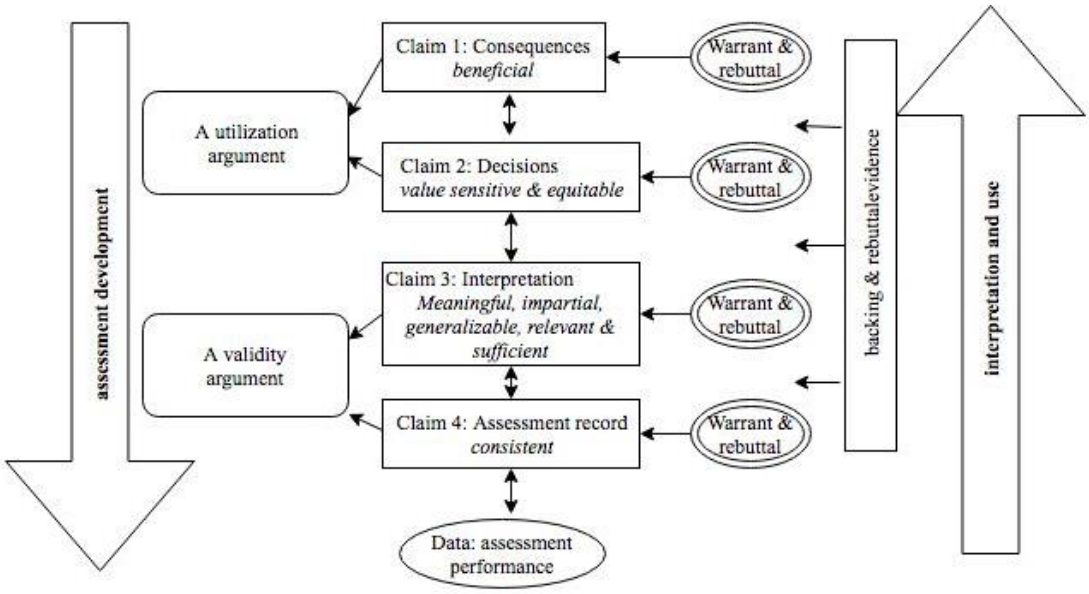


Figure 2.7. An AUA framework (adapted from Bachman & Palmer, 2010, p.91, p.104)

As argued by Bachman and Palmer (2010), using a language assessment involves obtaining samples of individuals' language performance, recording their performance quantitatively or qualitatively, interpreting these records to indicate test-takers' language development indicators of aspects of test-takers' language ability, making decisions based on these interpretations, and considering the consequences of the assessment or any related decisions. As shown in Figure 2.7, an AUA consists of a series of statements (claims, warrants, rebuttals) about the outcomes (i.e., consequences, decisions, interpretations, assessment records) of a given assessment and the qualities of these outcomes. It provides a means for defining the qualities that are associated with specific assessment outcomes as well as for understanding the relationships among various qualities. In addition, test consequences in an AUA study are linked to validity issues via a series of coherent inferences which thus prompted later argument-based approaches to include arguments for test use. Moreover, rather than seeking evidence according to the traditional "checklist" of validities, an AUA offers sufficient flexibility in collecting the backing evidence relevant to validity or test use claims.

The two aforementioned argument-based approaches provide several advantages for the MMR study. First, they highlight an overarching structure and provide a conceptual guidance for an explicit and coherent linkage from test performance to interpretations, and from interpretations to uses. Furthermore, test score interpretations are not "valid" unless they are connected to specific purposes. This advantage is crucial in investigating the consequential validity and washback of the new HSK. Although the HSK test developers have stated the goals of this test, both the test takers and CSL teachers (especially those from the Western nations) still argue that the HSK is not a proper measure of "communicative" language competency and actual proficiency level (Meyer, 2014). Hence, the new HSK should be interpreted according to its use and a more explicit labeling of the HSK's intended purpose is needed.

The second fundamental advantage concerns the consequences of testing. The outcomes of testing must be considered when using an assessment, since they can reveal potential weaknesses or negative aspects of the test development process. Moreover, they are directly addressed by the decision inference. In testing, it is important to consider the consequences various stakeholders are confronted with. As well, it is also important to consider the extent to which these consequences suit the purpose of the test. Are the outcomes intended (and positive), or unintended (and negative)?

2.7 The Validity and consequence/impact/washback studies on the HSK

A detailed review of the three most prominent journals in the field of Teaching Chinese as a second/foreign/heritage language (TCSL) in the past 30 years (i.e., *Language Teaching and Linguistic Studies, Chinese Language Learning, Chinese Teaching in the World*) showed an increasing number of research studies on Chinese language assessments. Three characteristics can be observed. First, the studies in the 1980s and1990s mainly revolved around framework

development in language tests, test design, validity, and equivalence. The trend after the 2000s followed the interests in new test technologies and ethical issues. Second, studies on the HSK account for 53.4% of all LT articles in the three journals. This statistic shows the preference for valuing exams and neglecting other forms of assessments (e.g., classroom-based assessment). Finally, as most of these articles are descriptive, there is a lack of empirical research in this area, particularly in classroom.

Inspired by washback research conducted in the ESL/EFL context, an increasing number of washback studies have been conducted in the Chinese context over the past decade, mostly focusing on EFL instruction (e.g., Gu, 2005; Qi, 2004, 2007; Wang, J., 2010). Compared to the abundant research on EFL tests in the Chinese context, it is surprising that there is very little focus on the washback effects of the Chinese language tests. The first study on washback was conducted by Yang and Liao (2000). It consisted of a case study of the old HSK's impact on 67 Thai college-level students majoring in Chinese. The results revealed that the students with positive beliefs about learning and testing received better scores than those with negative beliefs. It also indicated that the HSK had provided a negative influence on CSL teaching, as more attention was paid to linguistic knowledge and communicative skills were neglected.

Ten years later, Huang and Li (2010) conducted a mixed methods study on the old HSK's washback effects from the teachers' perspective by surveying 150 CSL teachers from eight Chinese universities. They found both positive and negative washback of the HSK in CSL education. On the one hand, teachers reported that the HSK objectively and accurately reflected students' language proficiency. Since "the HSK is the baton of CSL teaching and learning" (p. 26), teachers also reported that the test could inform them about students' strengths and weaknesses, help them adjust their teaching objectives and pedagogy, and promote students'

self-learning. On the other hand, the teachers criticized how the old HSK overemphasized the memorization of vocabulary and grammar rules at the expense of enhancing communicative skills. They also pointed out that to some extent, the HSK was not fair for Western test-takers when compared to those from Chinese-character cultural circles. This was because several Asian languages are written using Chinese characters, which meant that it was easier for students from those countries to read and write Chinese. Based on these results, Huang (2013) further explored the HSK's washback effects on learning behavior through a quantitative survey study. The findings showed that 64% of the test takers believed the HSK aided their learning. At the same time, they also believed the HSK needed improvement (e.g., reducing the number of questions, lowering the difficulty level, etc.). While this was a good step forward in terms of developing a research base for the HSK's washback effects, the study relied on an over-simplified questionnaire design, and it did not clearly show how the HSK affected learning.

In order to investigate the washback effects of the new HSK in classrooms, Huang, Y. (2014) adopted Shih's (2007) classroom observation scale and interview protocols to compare two types of TCSL classes – an HSK preparatory course and a regular course. The results indicated that the regular course placed emphasis on "communication," while the preparatory course was exam-oriented and only focused on teaching to the test. The author concluded that the washback effects were not significant, which contradicts other research findings in TCSL contexts.

Other researchers have focused on investigating the HSK's impact on aspects beyond teaching and learning, such as on the selection of CSL textbooks and HSK preparation materials. Influenced by structural linguistics theories, the old version of the HSK used CSL textbooks that focused on linguistic accuracy, such as by emphasizing grammar and sentence structure practice.

However, Li and Zhang (2011) pointed out that the new 2009 revision of the HSK had a more communicative emphasis. They argued that CSL textbooks should reflect this change from a focus on linguistic knowledge to language communication development. In addition, they stated that CSL textbooks/materials should correspond to different learners' characteristics, such as nationality. For example, it was reported that half of all the HSK candidates were from Korea.[7] Li and Zhang subsequently suggested that CSL textbooks and HSK preparation materials that target the Korean market should be further developed.

A review of the methods adopted in washback studies on the HSK in both teaching Chinese as a second and foreign language contexts is provided below. The findings are illustrated in Table 2.2. When compared to the washback studies listed in Table 2.1, it can easily be noted that most studies were mono-method (e.g., questionnaires) and had one category of participants, rather than utilizing other quantitative, qualitative, or MMR methodologies. In order to gain a comprehensive understanding of the HSK's washback effects, more research on multi-methods and from the perspectives of multiple stakeholders is needed.

[7] From the test-takers distribution around the world (in the year of 2010), Asia is the best represented continent and Koreans contributed 54.37% to all the participants (*Report on Overseas Enforcement of New Chinese Proficiency Test*, 2011).

Table 2.2

Summary of Major Washback Studies on the HSK

Studies	**Research Contexts**	**Data Collection Methods**	**Participants**	**Results**
Huang (2013)	TCSL (old HSK)	Questionnaire	Test-takers	The old HSK needed to be improved; the HSK had a positive influence on learning.
Huang & Li (2010)	TCSL (old HSK)	Questionnaires Interviews	Teachers	Both positive and negative washback of the HSK existed in CSL education
Huang, Y. (2014)	TCSL (new HSK)	Classroom observations, Interviews	Test-takers Teachers	The regular course places emphasized "communication", yet the preparatory course was teaching the test. The washback effects were not significant.
Wang (2013)	TCFL (new HSK)	Questionnaires Interviews, Documents analysis	Test-takers	Both positive and negative washback effects were found.
Yang & Liao (2000)	TCFL (old HSK)	N/A	Teachers Students	The students with positive beliefs about learning and testing obtained better scores than those with negative beliefs. They also indicated that the HSK had a negative influence on CSL teaching.
Zhang (2011)	TCSL (new HSK)	Documents, Materials analysis	N/A	CSL textbooks should reflect the transfer of linguistic knowledge to language competency development, and should correspond to different learners' characteristics.

The above review of the Chinese literature shows three limitations concerning the HSK and washback studies in international CSL educational contexts. The first limitation is associated with defining or conceptualizing the term washback. Several studies discussed the issue of washback (e.g., Huang, 2013; Huang, Y., 2014), but they did not have a thorough understanding

of the theoretical underpinnings of this concept, especially in terms of addressing the macro level. The second limitation is that in some cases (e.g., Yang & Liao, 2000), researchers did not seem to provide sufficient data to back up their claims. Furthermore, it is crucial to note that most HSK washback studies were based on the old HSK (before 2009). Thus, the washback effect of the updated version cannot be taken for granted without new empirical evidence. The last limitation relates to the research methods employed by the researchers. A number of studies only relied on a single method (e.g., questionnaire), rather than other quantitative, qualitative, or mixed methods methodologies. In order to ascertain a comprehensive understanding of the washback effects, more research using multi-methods and incorporating perspectives from multiple stakeholders is needed. In light of the aforementioned limitations, more rigorous research is necessary in the area of CSL and the HSK.

2.8 Articulating an AUA Framework within the HSK context

In a validity review article of the HSK, Chen (2006) listed several shortcomings of the validation research on the CSL proficiency assessments. For example, he criticized the common validation procedures, which only focused on the validity of the scores and did not interpret validity as a more complex concept. He insisted that "validation has to primarily take into account theoretical considerations that investigate whether the use of an assessment, the interpretations of test scores, and the inferences drawn from them are valid" (Chen, 2006, p.204). Chen pointed to the core issue: validity needs to have a purpose, a function, and a frame in which it can be applied. Combined with the PPP washback model, CLT theory, and Henrichsen's model, the argument-based approach constitutes a framework for guiding test development and test use. Not only can it provide the HSK stakeholders with useful insights to help deepen their

understanding of test validity from the perspective of test use, but it can also shed light on the CSL/CFL test washback/impact/consequence literature.

Among the various argument-based validation approaches in general education and LT (e.g., Bachman, 2005; Bachman & Palmer, 2010; Chapelle et al., 2007; Kane, 2006, 2013; Toulmin, 2003), the AUA was selected for this study for many reasons. First of all, an AUA establishes logical and coherent inferential links from the test taker's performance to the test developers' intended consequences. It guides the process of assessment development or assessment justification. This framework can also be helpful in providing a methodological guideline for collecting evidence and for mapping qualitative data regarding stakeholders' perceptions towards test use. In addition, an AUA advances the argument-based approaches to validation by including an argument for test use (Liu, 2013). More specifically, it can reflect the iterative and fluid nature of test use in teaching and learning (Doe, 2015).

As indicated in the previous section, the generic AUA framework includes two parts 1) a utilization argument that links an interpretation to a decision and 2) a validity argument, which links assessment performance to interpretation. Due to the nature and purpose of the current MMR study, a greater emphasis was placed on the former and a top-down approach was used to justify the test's intended purposes. In addition, since confidential data was extremely difficult to obtain and the researcher had no access to the central database of HSK scores needed to support the consistency claim (Claim 4) with its warrants (refer to Figure 2.7), the main focus was on Claim 1, Claim 2, and Claim 3. As stated by Bachman and Palmer (2010), it was unnecessary to articulate all the illustrative warrants they listed. Therefore, considering the practical constraints, the process of articulating the AUA was adapted to the HSK and the purpose of the current MMR study. The claims and warrants for this MMR study are listed below in Table 2.3.

Table 2.3

Articulating AUA Framework into the HSK Context

Claim 1: Consequences (beneficial)

The consequences of using the HSK and of the decisions that are made are beneficial to all the stakeholders.

Warrants:

1. The consequences of using the HSK that are specific to immediate stakeholder groups (students, teachers, programs) will be beneficial.
2. In language instructional settings, the HSK promotes desirable instructional practice and effective learning, and the use of the HSK is thus beneficial to students, teachers, programs, etc.

Claim 2: Decisions (values, equitable)

The decisions made based on the basis of the interpretations of the HSK take into consideration educational and societal values and relevant regulations, and are equitable for the stakeholders affected by said decisions.

Warrants:

1. Decisions made on the HSK scores take into account the existing educational and societal values and relevant legal requirements in both academic and non-academic settings.
2. Test takers are classified only according to the cut-off scores and decision rules, and not according to any other considerations; test takers and other affected stakeholders are fully informed about how the decisions are made and whether decisions are actually made in the way described to them.

Claim 3: Interpretations (meaningful, impartial, generalizable, relevant, sufficient)

The interpretations of test takers' overall Chinese proficiency are meaningful with respect to the *Scales*[8], the curriculum objectives and the test specifications. The interpretations are fair to all test takers, realizable to the Chinese language use domain in which the decision is made, and are relevant to and sufficient for the decisions that are to be made.

Warrants:

1. The HSK is meaningful and generalizable for its content representativeness and content relevance in accordance with the Scales and the curriculum objectives.
2. The assessment tasks do not include content that offend or favor test takers, and the individuals are treated impartially during the whole procedure of the assessment.
3. The assessment-based interpretation provides relevant and sufficient information to make decisions.

[8] The *Chinese Language Proficiency Scales for Speakers of Other Languages* (Office of Chinese Language Council International 2015), abbreviated as *Scales*, serves as a reference for creating a syllabus for teaching Chinese to speakers of other languages, for compiling Chinese textbooks, and for assessing the language proficiency of CSL learners. More information about the *Scales* will be provided in Section 3.2.

2.9 Chapter Summary

A large number of empirical studies have examined the various washback effects on teaching and learning (e.g., Andrews, Fullilove & Wong, 2002; Cheng, 1998, 2005; Hamp-Lyons, 1997; Qi, 2005; Shih, 2007; Turner, 2009;Wang, S., 2013; Watanabe, 1996); stakeholders other than teachers and learners (e.g., Cheng, Andrews & Yu, 2011; Hawkey, 2006; Pan & Roever, 2016); construct validity (e.g., Bachman, 2005, 2010; Xie, 2010); language policy (e.g., Shohamy, 2001, 2006; McNamara & Roever, 2006); and ethics and fairness (e.g., Davies, 1997; Kunnan, 2000; Shohamy, 2004). This chapter subsequently reviewed the literature on consequential validity, washback effects, and the methodologies employed in washback/impact/consequence studies. In the literature these terms are sometimes distinctively defined, and at other times they are used simultaneously meaning the same thing. For the purposes of this dissertation, the terms will be used simultaneously. Following this, argument-based validation theories were reviewed along with a framework articulating AUA for the HSK, which served as a methodological guideline for the current MMR study. The following chapter will provide a detailed description of the research context and present the methodological design of the study.

Chapter 3 Methodology

3.1 Introduction

This section provides an overview of the methodology employed in the current MMR study. It starts by describing the research context and the rationale for the research methodology, before moving to discuss the research design and ethics issues.

3.2 Research Context

3.2.1 CSL Education and the "Promoting Chinese Internationally" Policy

As mentioned in the earlier chapters, the rising national power of China has led to a worldwide enthusiasm for learning the Chinese language in recent years. Although reliable numbers concerning CSL learners worldwide are non-existent (Sun, 2009), the evidence of a strong increase has been witnessed. Since the influence of the Chinese language is rapidly rising, some have even argued that it could overtake English as the international lingua franca in the 21st century (Odinye & Odinye, 2012).

The rise in CSL can actually be traced back two thousand years. The prosperous economy and civilization during the Han and Tang dynasties led to the emergence and establishment of the Chinese-character cultural circle[9]. As one of the world's oldest established civilizations, China's 5000-year history has produced unique traditions, including Chinese medicine, philosophy, Kung Fu, and cuisine. More recently, since the reform and the opening up of China to the world in 1978, the Chinese economy has been growing rapidly, and the country has made much progress in building a highly cultural and ideological civilization.

[9] The Chinese-character cultural circle refers to the nations and districts that currently or previously used and have inherited Chinese-character traditions. Geographically, this cultural field includes China, the Eastern Indochinese Peninsula, the Korean Peninsula, and Japan. In religious terms, the field is also called the "Confucian cultural circle" or the "Chinese Buddhist cultural circle." The basic elements of the circle include Chinese characters, Chinese literature, Confucian tradition, Chinese Buddhism, China-pattern laws and institutions, China-pattern production technology, and China-pattern customs. In short, the Chinese-character cultural circle is a real, reflexive, and vigorous cultural phenomenon.

The history of the teaching and learning of Chinese is a long one marked by recent expansion (Tsung & Cruickshank, 2011). As early as the seventh century, there have been records of teaching Chinese to foreigners in China. With regard to teaching Chinese as a foreign language in schools and universities in the western hemisphere, it was introduced over a century ago in Paris as of 1840, at Yale in 1871, and in London at the School of Oriental Arabic and Semitic Studies in 1917. At that time students were mainly missionaries and sinologists (Tsung & Cruickshank, 2011). Research in this field is fairly recent and has been conducted in various contexts, such as Chinese as a SL/FL in China and other countries, as a SL to ethnic minority groups in China and post-colonial contexts (e.g., Singapore and Hongkong), and as a heritage language in diasporas across the world. The various contexts of Chinese teaching and learning have led to a diversity in CSL curriculums, teaching approaches, as well as in learners and their identities.

China's role as a world economic power in recent years has led to a growing interest in CSL. China's current "Promoting Chinese Internationally" policy is yet another critical aspect of its support of CSL teaching and learning (TCSL). In the early 1990s, Hanban published a book to discuss the promotion of Chinese, focusing on policies and agencies responsible for language promotion. The first World Chinese Conference, held in 2005, marked the promotion of the Chinese language internationally as a national strategic policy (Li, 2012). In addition, the National Medium and Long-Term Educational Reform and Development Plan (2010-2020) stated that to further expand the scale of foreign students coming to China, the "studying in China" program would be implemented and the number of foreign students would reach 500,000 by 2020. Since then, a series of advancements has occurred, such as the opening of Confucius Institutes in over 100 countries, the training of Chinese language teachers, as well as the launch

and reform of Chinese tests. Since 2012, the name of the undergraduate program -“对外汉语 (teaching Chinese as a Second Language),” which has existed for over 30 years in Chinese universities officially changed its name to “国际汉语教育(International Chinese Language Education)” by the Ministry of Education (MOE) (Lu, 2014). As such, to support this effort, it would be worthwhile to investigate the comprehensive relationship between language policy and language testing in a specific context.

3.2.2 The Development of the HSK

In line with the global “Chinese fever” phenomenon, growing numbers of CSL learners worldwide have participated in Chinese language proficiency tests for CSL/CFL. The number of Chinese proficiency tests available to Chinese learners has risen as well (Meyer, 2014). These tests play different roles and fulfill various purposes in certifying proficiency levels. For instance, they have served as a mandatory requirement for entering a Chinese university program (e.g., the HSK, the Taiwanese Test of Chinese as a Foreign Language 華語文能力測驗); for placing students into appropriate language course levels (e.g., placement tests in university CSL programs); for recruiting employees with Chinese business communication abilities (e.g., Business Chinese Test); and for encouraging foreign young students to learn Chinese (e.g., Youth Chinese Test). Among all these tests, the HSK (汉语水平考试), the official Chinese proficiency test from the People’s Republic of China, has the largest test population, prompted the most research, and had a major impact on test takers’ lives (Meyer, 2014).

According to Sun (2009), the development of the HSK can be divided into three phases: 1) the initial phase ranging from 1980 to 1990; 2) the expansion phase ranging from 1990 to 2000; and 3) an innovative phase ranging from 2000 onward. The development of the HSK began in 1984 at the Beijing Language and Culture University (BLCU). Its development was

strongly influenced by the dominant English language proficiency tests (e.g., TOEFL), which prompted it to shift its focus from language knowledge to language ability. In 1992, the HSK became the official national standardized test and was launched outside of China. In 2000, the number of test takers reached over 80,000, of whom 36.5% were "foreigners" and the remaining were members of Chinese ethnic minorities. Shortly after, in 2004, the Ministry of Education of China withdrew the HSK authorization from the BLCU, shifting all the rights to Hanban. Incorporating aspects of its previous version while also drawing on the latest findings in global language testing (e.g., developing computer/internet based tests), the new format of the HSK was introduced in November of 2009. The test can be paper- or internet-based, depending on the test-takers' choice and what the specific test center offers. In the internet-based test, the writing task (for HSK 3 onwards) can be completed by using the Chinese input system[10]. In other words, test-takers only need to write in Pinyin and pick the right character from the keyboard. Learners who take the paper-based test are not afforded this luxury as they need to remember all of the strokes for various Chinese characters and then write them down manually. The current test structure is presented in Table 3.1, and more detail on the tests' questions/items/tasks are provided in Appendix 1.

The new HSK is designed based on the *Chinese Language Proficiency Scales for Speakers of Other Languages* (Office of Chinese Language Council International 2009), which is abbreviated as *Scales*. It is an official document with guidelines for CSL teaching and learning, and serves as a reference for designing CSL/CFL syllabi, for compiling Chinese textbooks, and for assessing the language proficiency of CSL learners. The *Scales* have been established on "the

[10] Chinese input systems, also called Chinese input methods, are methods that allow a computer user to input Chinese characters. Mostly, they fall into one of two categories: phonetic readings or root shapes.

principle of drawing on the strengths of other language proficiency scales already developed internationally, taking theories of communicative competence as their foundation, focusing on the learner's actual use of the language and reflecting the characteristics of the Chinese language" (p. iii). It provides a five-band all-round description of learners' ability to use the Chinese language for communication. Hanban stated at the time that the HSK's six levels corresponded to the five-bands of the *Scales*, and the five levels of the Common European Framework of Reference for Languages (CEFR). The estimated equivalence among the New HSK Tests, the *Scales*, and the CEFR is presented in Table 3.2.

Table 3.1

The New HSK Test Structure

Level	**Vocabulary**				**Written test**			**Description**	**Oral test (HSKK)**
	Words (Cumulative / new)		**Characters (cumulative / new)**		**Listening**	**Reading**	**Writing**		
1	150	150	174	174	20 questions, 15 minutes	20 questions, 17 minutes	Not tested	Designed for learners who can understand and use some simple Chinese characters and sentences to communicate, and prepares them for continuing their Chinese studies. In HSK 1 all characters are provided along with Pinyin.	Beginner (27 questions, 17 minutes)
2	300	150	347	173	35 questions, 25 minutes	25 questions, 22 minutes		Designed for learners who can use Chinese in a simple and direct manner, applying it in a basic fashion to their daily lives. In HSK 2 all characters are provided along with Pinyin.	
3	600	300	617	270	40 questions	30 questions	10 items	Designed for learners who can use Chinese to serve the demands of their personal lives, studies and work, and are capable of completing most of the communicative tasks they experience during their Chinese tour.	Intermediate (14 questions, 21 minutes)
4	1200	600	1064	447	45 questions	40 questions	15 items	Designed for learners who can discuss a relatively wide range of topics in Chinese and are capable of communicating with Chinese speakers at a high standard	
5	2500	1300	1685	621	45 questions	45questions	10 items	Designed for learners who can read Chinese newspapers and magazines, watch Chinese films	Advanced

								and are capable of writing and delivering a lengthy speech in Chinese.	(6 questions, 24 minutes)
6	5000	2500	2663	978	50 questions	50 questions	1 composition	Designed for learners who can easily understand any information communicated in Chinese and are capable of smoothly expressing themselves in written or oral form.	

(Retrieved June 14, 2015 from http://en.wikipedia.org/wiki/Hanyu_Shuiping_Kaoshi)

Table 3.2

The Estimated Equivalence among the New HSK Tests, the Scales, and the CEFR

New HSK	**CEFR**	***Scales***
HSK Level 6	C1	Band 5
HSK Level 5		
HSK Level 4	B2	Band 4
HSK Level 3	B1	Band 3
HSK Level 2	A2	Band 2
HSK Level 1	A1	Band 1

(The Office of Chinese Language Council International, 2010, p. 1)

A number of researchers (e.g., Liu et al., 2006; Yang & Liao, 2000) argued that in the past decades, TCFL overemphasized grammar and sentence structures, both of which are strongly based on structural linguistics perspectives. They believed that the inadequate attention that was paid to the social functions of language and its use in real-life settings negatively influenced CSL learners' communicative competence. Similarly, the old HSK had also been criticized for being impractical and for forcing students to learn specific language knowledge (e.g., grammar) solely for the purpose of passing the exam rather than to acquire practical linguistic abilities. Liu (1994) stated that the language materials underlying the construction of HSK items should focus more on communicative language functions. This means that the materials should cover authentic language situations that people encounter in daily life. More recently, a mixed-methods study on the reformed HSK conducted by Wang (2013) confirmed that the inclusion of a compulsory speaking component strengthened the development of learners' communicative competence; this increased their interest in building oral

communicative abilities in CSL classrooms and helped them apply the acquired skills in real-life situations.

With this in mind, HSK test developers had become interested in measuring communication ability in relation to language knowledge. They insisted that this relationship should positively influence TCSL, and the new HSK purported to address these problems. They also claimed that the new HSK was able to more accurately measure language use including producing spoken and written Chinese. More specifically, the old HSK had 11 levels, of which the most advanced levels were measured through multiple-choice items. The new HSK includes written (in intermediate and advanced levels) and spoken sections, in addition to listening and reading comprehension, and grammar components. In comparison to the old HSK, the new HSK reduced the number of characters (8% fewer) and vocabulary/character combinations (38% fewer). For example, the top level requires 5,000 words rather than 8,000 words. Some researchers (Bellassen, 2011; Meyer, 2014; Xie, 2010) criticized the new HSK, saying that it lowers the standards in CSL and prematurely linked the test to the Common European Framework of Reference for Languages (CEFR). The HSK developers, however, insisted that the assessment was constructed using scientific principles with the intention of improving teaching and learning through testing (Sun, 2009), and it conveyed the spirit of the *Scales*, which is an official guideline document for TCSL. Communicative ability is ultimately the core focus.

In sum, the research context is based on two aspects, namely: (1) The TCSL and Promoting Chinese Internationally Policy, and (2) the HSK. Due to the reform of the test (e.g., reduction of the number of questions, lowering of the level of difficulty, inclusion of oral exams and Chinese input systems) and a massive promotion campaign led by the Hanban, the number of test takers has risen substantially since 2010. Moreover, CSL teaching and learning

worldwide has created a breeding ground for further developing and conducting research on the HSK. In fact, inspired by the abundant language testing (LT) research conducted in the ESL/EFL context, researchers have been conducting numerous validity, reliability, and equivalence studies in the field of Chinese language assessments. Nonetheless, it is surprising that there is very little focus on the reformed HSK and its use, especially from the perspective of washback effects within the context of the policy on Promoting Chinese Internationally - an observation that is incommensurate with the test's important status. In addition, the goal of the reformed HSK is to support the interrelationship between teaching and testing, and to "facilitate teaching and learning through testing" [考教结合，以考促学、以考促教]. Consequently, more rigorous empirical studies are needed to explore the nature of its consequential validity and washback, that is, at the micro (classroom) level, how it functions to influence teaching and learning, and how to minimize the negative effects of washback and maximize the positive ones; at the macro (society) level, to explore how test users understand the test use, score interpretation, and decisions made based on HSK levels/scores.

3.3 Mixed Methods Sequential Exploratory Design

3.3.1 Rational for the MMR methodology

Drawing on the literature review in Chapter 2, it is evident that there is increasing awareness that MMR can provide a holistic picture of a research problem and can provide valuable insight into the deeper and wider understanding of complex phenomena (Creswell & Plano Clark, 2011; Green, 2007; Teddlie & Tashakkori, 2010). Consequently, this study was conducted employing a multi-phase MMR framework due to the considerations described below.

This approach can not only help overcome limitations and solve problems associated with mono-method studies (Kelle, 2006), but is an advanced research strategy in many ways, such as

that it allows for 1) complementarity, in that overlapping and different facets of a phenomenon may emerge; 2) sequencing, wherein the first method is used sequentially to help inform the second method; 3) initiation, under which contradictions and fresh perspectives emerge; 4) expansion, where mixed methods add scope and breadth to a study (Greene et al., 1989), and 5) triangulation, in which seeking convergence, corroboration, and correspondence of different methods' results can be used to explain unexpected results. For example, a post-stage interview after a questionnaire survey can add strength to data triangulation and provide useful explanations for quantitative findings.

The value of MMR in washback studies also lies in its flexibility to mix aspects of the qualitative and quantitative paradigms or several methodological steps in the design (Creswell & Plano Clark, 2007). For example, in Cheng (1997, 1998, 2005), Turner (2009, 2013) and Wall (1999)'s studies, the QUAN and QUAL data sources are combined in data collection and data analysis, whereas in Watanabe (1996) and Alderson and Hamp-Lyons (1996), the two approaches are mixed only in the phase of data analysis. Further, in order to optimize the research findings of washback studies, choosing and implementing the most appropriate research design in accordance to the research questions is one of the crucial components of employing MMR (Creswell & Plano Clark, 2007, 2011; Greene, 2007; Teddlie & Tashakkori, 2009, 2010, Turner, 2013). A salient example is Tan's (2009) study that employed a sequential exploratory triangulation design to examine a change in the language of instruction for Mathematics and Science (MS) subjects from Bahasa Malaysia to English in the context of Malaysia's PPSMI[11]

[11] PPSMI policy refers to Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris (Teaching and Learning of Science and Mathematics in English). This policy, implemented by the Malaysian Ministry of Education in 2003, mandates a change in language of instruction, from Bahasa

policy. The multiple sources of data (i.e., document analysis, observations, interviews, field notes, and documents) in this longitudinal study were analyzed both quantitatively and qualitatively. The information was then triangulated in different stages of the MMR design, which provided diverse types of information to validate the analysis and to address her research questions. As Tan noted, "this iterative process made it possible to generate new questions for various participants as the year progressed and also allowed me to find contradictions and puzzles in terms of the [implementation of] the PPSMI policy" (p. 85). In light of this practical evidence provided by different researchers, it is clear that the methods they used were pragmatically chosen and allowed them to generate helpful and valid answers to their research questions.

Another major strength of an MMR design is how its tailor-made, exclusive nature can better reflect the context and characteristics of the target issue. In order to capture the whole picture of the consequence/impact/washback phenomenon, an increasing number of researchers have realized the importance of investigating issues in a specific context (e.g., Cheng, 1997, 1998, 2001, 2004; Cheng and Sun, 2015; Davison, 2006; Qi, 2005; Shohamy et al., 1996; Tan, 2009, 2011; Turner, 2009; Wall, 1999; Watanabe, 2004). Watanabe (2004), for instance, emphasized the significance of "context" in washback studies by iterating that

> It is crucial to describe the context (both at micro and macro levels) as explicitly as possible, not only to help readers understand the role of the test in that context, but also to establish transferability or the demonstration of the generalizability, or applicability of the results of a study in one setting to another context, or other context. (p.25)

Malaysia to English for all Mathematics and Science subjects taught in Malaysian primary and secondary schools.

These studies' findings suggest that in addition to test-related factors (e.g., stake, status, purpose, format, content, etc.), there are other factors that affect washback effects, such as context-related factors (e.g., classroom size, timing of the course, available resource, and professional support) and stakeholder-related factors (e.g., educational background, experience, language ability, and training). The HSK's consequence/impact/washback is also a complex phenomenon with multiple dimensions. The divergent perspectives of test users, complex individual (students and teachers) behavior, and HSK's contextual characteristics make it difficult to understand the situation using a single method. To answer a more complete range of research questions and to acknowledge the multi-faceted nature of the phenomena inherent in systemic environments as well as dynamic classrooms, MMR is the most appropriate and essential methodology for this study.

3.3.2 The design of the current MMR study

This dissertation study employed a mixed methods sequential exploratory (MMSE) design, whereby a qualitative study was conducted to help identify theoretical concepts/core issues and to develop measurement instruments and hypotheses for the subsequent quantitative study. A quantitative study was then carried out to identify whether concepts/issues established from a comparable small number of cases could be described and explained in a greater domain (Creswell, 2015; Kelle, 2006).

The dissertation includes 3 studies which are the 3 phases and components of the MMR design. To be specific, Study 1 is the first phase (i.e., qualitative phase), and then Study 2 and 3 are the second and third phases (i.e., quantitatively orientated phases). The methodology of each study will be discussed separately in the following chapters, together with the findings and discussions for each. A visual diagram of the MMR design of Studies 1, 2, and 3, which "pull

together all of the components of the study" (Creswell, 2015, p. 63), is presented in Figure 3.1. It includes the methods, data collection, data analysis procedures, and products of each study, and the overall design of the dissertation.

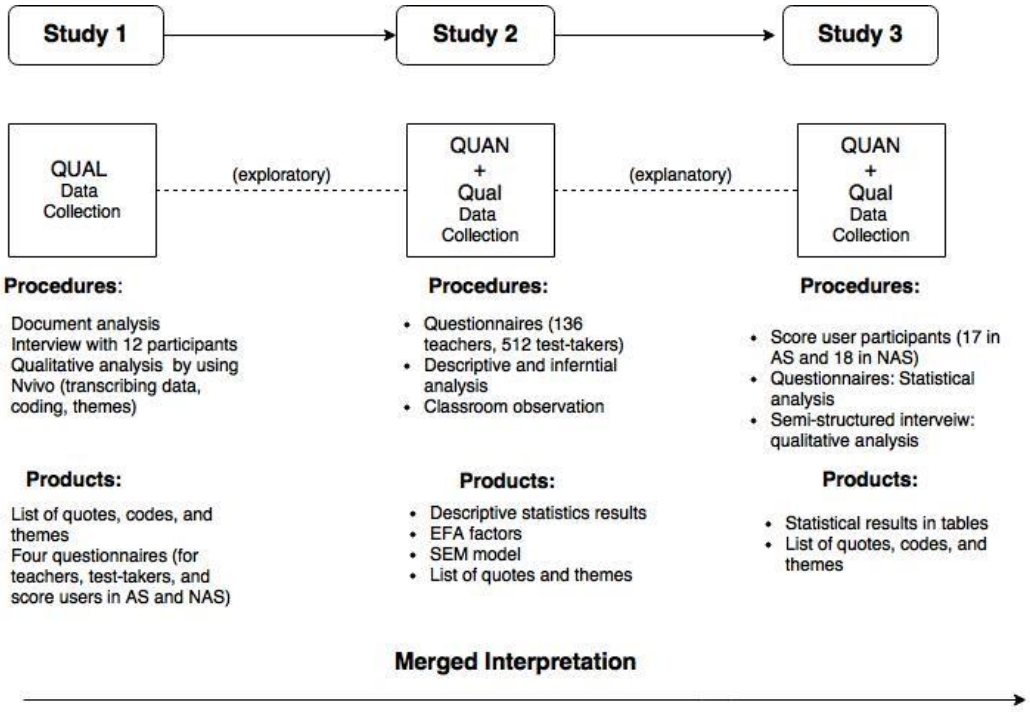


Figure 3.1. The visual diagram of the MMR design of the dissertation

As mentioned in Section 2.8, an AUA framework was employed in the MMR study and in the sub-studies providing both the theoretical framework and the methodological guidelines for collecting evidence regarding stakeholders' perceptions towards HSK use. For example, the AUA framework is integrated into Study 1, and the results and discussions of Study 1 are presented in relation to the three AUA claims (i.e., consequences, decisions, and interpretations). Study 2 concentrates on the consequence claim, particularly at the micro (classroom) level. Study 3 focuses on the consequence claim at macro level, which includes how the test affects various stakeholders' decision-making process and test score interpretations. Chapter 7 is an overall discussion expanding on the major findings from the 3 studies by synthesizing,

integrating, and triangulating the results from the different data sets generated from the AUA framework.

3.4 Ethical issues

In compliance with the McGill Research Ethics Board (REB) rules, I obtained access to the research site and received consent from the participants before conducting this study. Since there were no sensitive or personal questions in the questionnaire or interviews, no ethical concerns were identified in completing either the survey questionnaires or the interviews. No psychological, emotional, economic, cultural, and/or social risks were foreseen for the participants.

In Study 1, the interview participants were given a consent form to read and sign before being interviewed. Depending on the participants' preference, the consent form was either emailed or provided before the interview; consent forms were signed before participants were able to progress into the interview phase. Regarding to the telephone interview, written consent was not be able to obtain. Instead, participants were read the oral consent form before the interview was initiated. If they consented to participate, the interview proceeded. Participants were all aware that they had the option to withdraw from the study at any time and that all information collected up to the time of withdrawal would be destroyed.

In Study 2, each of the paper and online version questionnaires contains a lead-in page that consists of a consent form. Regarding the online version, participants must click on an agreement button to participate in the study before proceeding to the actual questionnaire. They were asked to keep this page for their records. Considering that the test-takers who participated in the current research were from different countries and had different language proficiency levels in Chinese, I prepared two versions of the Letter of Consent and the online questionnaire:

one in English, and the other in Chinese. Classroom observation participants were given a second Letter of Consent in Chinese, which specifically discussed their rights during the observation.

In Study 3, the consent process for the questionnaires were the same as for Study 2, and the interviews were the same as for Study 1.

In all 3 studies, the participants of the survey and the interview were assured that their responses to the questionnaire or the interview would not be released to anyone (including the university authorities) without their consent and that their responses would be used only for the stated research purposes. They were assured that the issue of confidentiality was taken seriously by McGill REB rules. To demonstrate appreciation for the participants' time, all of the participants in these three studies could enter/report their name and contact information at the end of the questionnaire/interview in order to receive a $5 (≈ 20 RMB) gift card. If any participants chose to decline to answer a question or withdraw from the project, there was no compensation for participating. The total amount of compensations was approximately $3,000. Information collected to draw for the prizes were not linked to the study data in any way. This identifying information was stored separately, and was destroyed after the prizes were provided.

3.5 Chapter Summary

Due to the HSK reform and a massive promotion campaign led by the PCI policy, the number of CSL/CFL learners and HSK test-takers has risen substantially. The status of CSL/CFL worldwide has also prompted a breeding ground for further development and research on the HSK. The pragmatic principles of MMR make it powerful and efficient in two ways. Firstly, it enables researchers to investigate the general picture of a specific educational/societal context on a macro-scale; and secondly, it allows researchers to explore and gain detailed insights on specific cases on a micro-scale level. To be more specific, MMR approaches appear useful for

research on the impact of a test on teachers and students in classrooms as well as other stakeholders of the test in society. The mixed method sequential triangulation design employed in this MMR study allowed me to obtain multiple perspectives on consequence/impact/washback effects of high-stakes tests, explore my research questions more widely and deeply provide more convincing findings than through a monolithic methodology, and contribute a more comprehensive method that can enrich the existing washback literature. Having described the research design and ethical issues, the research methods, findings, and discussions of Study 1, 2, and 3 are presented separately in the following chapters (Chapters 4, 5, and 6).

Chapter 4: Study 1 (Investigating the consequential validity of the HSK by using an Argument-based framework)

4.1 Introduction

As mentioned in the previous chapter, the dissertation research employed a multi-phase mixed methods sequential exploratory (MMSE) design. Study 1 was the first phase. It involved a qualitative study that was conducted to explore the intended and actual consequences of the HSK, and to develop measurement instruments (i.e., questionnaires and classroom observation guide) and hypotheses (i.e., washback hypotheses) for the subsequent quantitative studies (i.e., Studies 2 and 3).

4.2 Methodology

The research questions are as follows:

RQ1: What are the intended consequences of the HSK use from the test developer's perspective?

RQ2: What are the actual consequences of the HSK use from multiple stakeholders' perspectives at both the micro (classroom) and macro (society) levels?

RQ3: In this context, is there any kind of relationship across the PCI Policy, TCSL, and any consequences of the HSK?

According to the AUA framework stated in Chapter 2, test users (including decision makers and those affected by the former's decisions) were identified first for this study. Table 4.1 presents the corresponding relationship among decisions made concerning HSK scores, stakeholders, and decision makers. The first row displays decisions made by the Ministry of Education (MOE) and test developers on HSK scores, which are publicized to different groups of stakeholders via official statements and the HSK specification documents. The second row is

related to decisions made at classroom and institutional levels (i.e., the micro and macro level). The third row focuses on the HSK certificate as a prerequisite for employment at the macro (society) level.

Table 4.1

Major Types of Decisions Made on HSK Scores and the Stakeholders

Multiple decisions		**Stakeholders to be affected**	**Decision makers**
Set cut off scores (180 out of 300 as a passing score)		Students, teachers, university academic affair office	Ministry of Education, the HSK developers
Educational Decisions	Micro level (classroom)	Students and teachers	Students and teachers
	Macro level (Institution)	Students, teachers, etc. (to be investigated)	University administrative officers, program coordinators, etc.
HSK certificate as a prerequisite for employment		Test takers	Employers

4.2.1 Participants

Twelve HSK stakeholders were recruited to provide a multifaceted understanding of the consequential validity of the HSK. Table 4.2 provides information about the participants, including their affiliations, positions, and job description and/or TCSL experience. These test user participants included an officer from the Education Office of the Consulate General of People's Republic of China (PRC), four CSL teachers, four test-takers, an administrative officer from a Chinese university, a human resources (HR) manager from a multinational enterprise, and one director of a HSK test center. These participants can help address the research questions at both the micro and macro levels. Firstly, the CSL teachers, 3 (out 4) test-takers, and the university administrator can represent the stakeholders from the micro (classroom) levels, while

the final test-taker and the HR manager provide insight from the macro (society) perspective. Furthermore, the government officer's opinions can help unravel the PCI Policy and its impact on TCSL and the HSK. Lastly, the representative from a test center can provide a deeper understanding of the test development and administration situation.

Maximal variation sampling (Creswell, 2008) was used in recruiting these participants. Based on the categories of Table 4.2, they were recruited from several research sites, which were varied in terms of location type (CSL in China vs. CFL/CHL outside of China), organization type (universities vs. others), and the type of context (educational vs. societal).

Table 4.2

Profile of the HSK Stakeholders

Name	Affiliation	Position	Job description/Background information
Goyin	The Education Office of the Consulate General of PRC	Consulate officer	Responsible for foreign exchange and cooperation in education; promoting Chinese-related affairs
EDTTKi	A university in China; A multinational enterprise	CSL learner; Employee	Has learned Chinese for 3 years; passed HSK5
EDTTZh	A university in China	CSL learner; Junior year international student	Has learned Chinese approximately 10 years; passed HSK3, HSK5, HSKK Advanced
EDTTGe	A university in Canada	CFL learner; Senior year student	Has learned Chinese for 3 years; passed HSK5
EDTTLu	A university in Korea	CFL learner; Senior year student	Has learned Chinese for 1 year; passed HSK3; failed HSK4
EdTPa	A university in China	CSL teacher	Has taught CSL courses for 3 years
EdTYa	A university in Canada	CFL teacher	Has taught CFL/CSL courses over 20 years
EdTHu	A university in China	CSL teacher	Has taught CSL courses for 8 years
EdTDo	A university in Thailand	CFL teacher	Has taught CFL courses for 10 years
BuWa	A multinational enterprise	Human resource manager	Responsible for recruiting employees
EdAYa	A university in China	Administration officer	Responsible for admissions for international students
TcZh	A HSK test center in Canada	Director	Responsible for test administration

* Pseudonyms were created for all participants to protect their identities. The initial letters of each pseudonym indicate the contexts where the participant was recruited. Go: government; TT: test-takers; T: education teacher; EdA: administration officer; Bu: business; Tc: test center.

4.2.2 Data collection and analysis

Figure 4.1 presents the research design of this study. First of all, the data collection began with an extensive literature review centering on: 1) the HSK-related documents posted on the HSK official website and technique reports issued by Hanban, 2) past HSK exam papers, 3) the HSK specifications documents, and 4) the HSK-related journal publications of the developers. The data obtained from these documents were regarded as official sources that reflected the test developers' intentions. The purpose of this phase was to understand the test development, to identify the new characteristics of the revised HSK, to find out what the HSK claims to measure (e.g., linguistic knowledge and/or language use ability), whether the HSK represents the *Scales*, and to understand the test developers' intended objectives. The data was coded according to these 5 purposeful categories.

In the second phase, in-depth individual interviews were conducted with the stakeholders to collect evidence. The collected information was then used to respond to the research questions, and to either support or refute the underlying warrants of the AUA claims. The interview format was semi-structured. It lasted 30-45 minutes for each participant in Chinese, English or a combination of both Chinese and English. Semi-structured interviews provided the opportunity to probe beyond the answers to the prepared questions. (Bogdan & Biklen, 1998). They were audio recorded and then transcribed for further analysis.

A two-cycle approach (Saldaña, 2009) for data analysis was deemed appropriate. First, all the HSK-related documents and interview data were coded based on the three AUA claims to locate either backing or rebuttal data for the warrants (as explained in Chapter 2). A second round of coding then took place, and new codes and themes were identified where appropriate. The coded data were explored using multiple query techniques in QSR Nvivo for Mac (e.g., text

search, word frequency, coding query, and matrix-coding query) as well as manually through analytical memos.

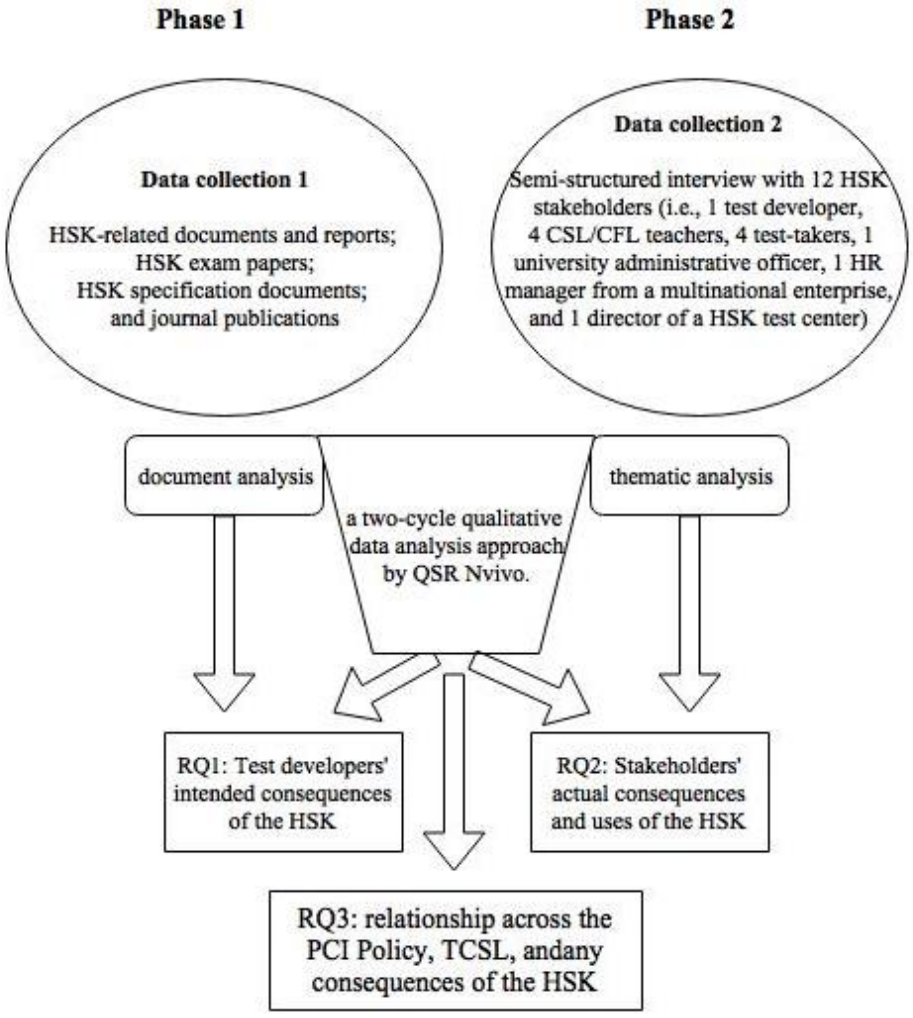


Figure 4.1. Research design of Study 1

4.3 Results and discussion

According to the AUA framework and the research questions, the results of Study 1 are presented and discussed corresponding to RQ1 "the intended consequences of HSK use", and RQ2 "the actual consequences of HSK use". The results of RQ 3, the relationship across the PCI Policy, TCSL, and any consequences of the HSK, are integrated with the above two research questions, rather than presented separately.

Table 4.3 represents the coding themes based on the AUA framework from the Nvivo analysis. The initial coding results from Phases 1 and 2 revealed 20 categories (also called theme nodes in Nvivo). Based on these nodes and further decision tree analyses, 10 key themes were identified. Seven of them corresponded to the 3 AUA claims (as shown in the row of the first coding cycle). In the second cycle, 2 new themes were seen to enrich the AUA framework.

Table 4.3

AUA Claims and Corresponding Themes

AUA Claims	Themes (Cycle 1)	Themes (Cycle 2)
Consequences	• Washback on teaching • Washback on learning	
Decisions	• Decisions for educational purposes • Decisions in business setting	
Interpretations	• The *Scales* and curriculum objectives • Fairness • As an indicator of Chinese proficiency	
		• HSK reform • Context factor

4.3.1 RQ1: Intended consequences of the HSK use

A review of the HSK-related documents (see Figure 4.2, data collection 1) indicated that the HSK test claimed to be a scientific, objective, accurate, and fair measure of students' Chinese proficiency (Office of Chinese Language Council International, 2010). Based on this assumption, the test developers indicated that the test results could serve several purposes: 1) a reference for educational institutions' decision-making concerning recruiting students, assigning students to different classes, allowing students to skip certain courses, and granting academic credits to students; 2) a reference for employers' decision-making concerning the recruitment, training, and promotion of test takers; 3) a method for Chinese language learners to assess and improve their Chinese proficiency; and 4) a method for Chinese language training institutions to evaluate

training results.

In addition, the test developers pointed out that the HSK was initiated based on a philosophy of testing "comprehensive language and communication ability" in order to improve teaching and learning through testing. This purpose was consistently stated across all the documents disseminated by the HSK developers. For example, they criticized the previous HSK version for being impractical and for forcing learners to concentrate on learning specific language knowledge (e.g., grammar) rather than to acquire practical language abilities. Due to inadequate attention paid to the social functions of language and its use in real-life settings, this design negatively influenced learners' communicative language competence. They thus suggested the following:

> "把语言形式和语言社会功能在教学中有机地结合起来，正确地处理好语言能力与交际能力的关系，以达到较全面地培养运用语言能力的最终目标，这是我们对外汉语教学所持的立场，也是设计汉语水平考试的依据。"[12] (Liu, Huang, Fang, Sun, & Guo, 2006, p.12)

With this direction in mind, the reformed HSK claimed to have addressed these problems, said to be a more accurate measure of learners' language use (including producing spoken and written Chinese with a focus on communicative language functions).

Moreover, considering the recent trends in promoting Chinese internationally, the HSK developers stated they made more efforts to attract and encourage CSL/CFL learners to take the HSK test. For example, the HSK's revision included the reduction of the number of questions, the inclusion of oral exams, and the inclusion of Chinese input systems in internet-based tests.

---

[12] Translation: [Cultivating a full linguistic command is the ultimate goal of TCSL. In order to achieve this, attention needs to be drawn to improve integration of language within its social functions, and the relationship between comprehensive language ability and communicative ability, which is the objective of TCSL and the cornerstone of the HSK test design.] (All translations from Chinese are by the author.)

Furthermore, in an official report on researching and producing the new HSK, the HSK developers mentioned the following:

> "…在中低等级考试中，则强调沟通理解，不苛求标准与规范，以一定程度上对不规范、不标准的容忍来换取考生能用汉语完成交际任务后所获得的成就感与自信心，换取汉语考试考生数量、汉语学习者数量的增加。" [13] (Zhang, Xie, Wang, Li, and Zhang, 2010)

Although some researchers (Bellassen, 2011; Meyer, 2014; Xie, 2011) have criticized the new HSK for lowering CSL/CFL standards, claiming that it lowers the standards in CSL and prematurely links the test to the Common European Framework of Reference for Languages (CEFR), the developers insisted that "绝不是单纯地将 HSK 难度整体下调，而是在保持高端难度的前提下，添补低端空白"[We are not simply lowering the level of difficulty of all the HSK levels, but on the premise of maintaining the difficulty of advanced levels, we added lower levels to the HSK (as compared to the previous version).] More specifically, the old HSK had 11 levels but the new HSK has only 6 levels; both included written and spoken sections, in addition to listening and reading comprehension, grammar, and writing (in intermediate and advanced levels) components. Furthermore, in comparison to the old HSK, the new HSK reduced the number of characters (8% fewer) and vocabulary/character combinations (38% fewer). For example, the top level of the new HSK requires learners to have a knowledge of 5,000 words rather than 8,000 words.

Overall, two interrelated categories regarding the intended HSK consequences can be summarized as: 1) Promoting CSL/CFL teaching and learning in the PCI context; and 2)

---

[13] Translation: [... In the elementary and intermediate HSK levels, emphasis is on communication rather than perfect linguistic performance. In order to ensure that test-takers obtain a sense of confidence and achievement after completing the communicative tasks, non-standard and non-normative language use could be accepted to some extent. In that way, the number of HSK test-takers and CSL/CFL learners will hopefully increase.]. (All translations from Chinese are by the author.)

Providing a useful reference of the test takers' Chinese language proficiency for making educational and social decisions.

4.3.2 RQ2: Actual consequences of HSK use

In this section, the results and discussions of the actual consequences of HSK use are discussed. The findings are based on the interview data with the 12 participants (e.g., CSL/CFL teachers and test-takers) (see Figure 4.2, data collection 2) and are described according to the AUA's claims and corresponding themes (see Table 4.4).

4.3.2.1 Claim 1: Consequences (beneficial)

The consequences of using the HSK and of the decisions made are beneficial to students, teachers, and programs.

As described in Chapter 2, washback is a complex phenomenon and involves various intersecting factors in the educational context (Alderson &Wall, 1993; Cheng, 2001, Watanabe, 2004). This section places emphasis on the reporting and discussion of washback effects on test preparation behaviors and teaching practices, in order to provide evidence regarding the intended and unintended HSK consequences from the HSK test-takers' perspective and CSL/CFL teachers' perspectives.

The interview data showed a wide range of opinions among test-taker participants in their learning beliefs and practices towards HSK preparation. For example, as EDTTGe said, HSK was just a feedback tool used to recognize one's current proficiency level and can help learners to identify strengths and weaknesses in their language abilities. When preparing for the test, this learner still resorted to standard learning methods, such as taking Chinese courses and talking to Chinese speakers, rather than focusing on the test prep materials. He considered the test as part of the learning continuum rather than an end-goal. However, the Korean participant EdTKi had a

different view. He believed that spending time on specific test preparation activities (e.g., taking an HSK test-prep course, hiring a tutor, taking mock tests) could significantly help increase his scores. He thus hoped his Chinese teacher could cover more test-taking strategies and test-related knowledge in his course. After passing the HSK Level 5 certificate, he was successfully admitted to a Chinese program in one of the most prestigious universities in China. It is noted that the greater the perceived importance of a test, the more impact the test will carry (Alderson &Wall, 1993, Shohamy, Donitsa-Schmidt, Ferman, 1996). Despite test-takers having varying opinions towards the washback effects of the HSK on their learning, in general, they believed that the HSK test can assist test-takers to improve their language skills and enhance their competitiveness in future study or in the job market.

However, some unintended consequences were discovered from the test-takers. For instance, as the HSKK, (the oral test), is an elective test separate from the HSK, the test-taker participants in Study 1 paid much less attention to speaking practices than the other three skills (i.e., reading, writing, and listening). As EDTTGe indicated, the lack of spoken tests makes the test less relevant to measuring students' communicative language skills. EDTTZh, the only test-taker participant who participated in the HSKK in this study, expressed the following:

> The test-takers cannot engage in any conversation in the HSK. You hear a sentence or a question from your headphone and then you repeat or answer the question via a microphone. That is not a conversation. I don't think it can reflect one's real communicative competence.

As advocated in Wang's (2016) washback study on the reformed HSK, the inclusion of a compulsory speaking component in the HSK would strengthen the development of learners' communicative competence, promote their interest in building oral communicative abilities in CSL classrooms, and help them apply the acquired skills in real-life situations. The intended

consequences cannot be achieved unless efforts are made to evaluate learners' speaking ability in a way that will reflect their communication skills.

Unlike the varied perspectives from the test-taker participants, interview results with the CSL/CFL teacher participants concerning their teaching practices appeared relatively consistent. All of them contended that CSL/CFL teaching and learning should not be test-oriented and it is inappropriate to include substantial test-related practice in class. Instead, they asserted that teachers should help students build a solid foundation of linguistic knowledge, increase their interest and motivation, and help them become autonomous learners. For example, EdTYa commented that,

> I personally am frustrated with test-oriented classes, and I also find it is inappropriate to use test-related instruction as part of the curriculum. (As a Chinese proverb says), "give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime"…. Sometimes, I adapted test-related contents into my classroom assessment, or even included HSK test questions in the mid-terms and finals, but it would not be the major focus in my class.

In addition, it is worth noting that the test developers have claimed that developing HSK-related textbooks and coaching materials is a significant component of test development. The purpose is to promote the integration of testing within teaching and learning as well as to change a simple examination system into a comprehensive system of "CSL/CFL teaching and learning assisted by the HSK" (Zhang et al., 2010). Aligned with the curriculum, a large-scale assessment can effectively inform classroom practices (Pellegrino, 2014; Phelps, 2012; Porter & Smithson, 2000). However, in this context, the test did not appear to be a major factor influencing teachers' teaching practices and teaching beliefs. Although this could be considered positive washback (e.g., will not result in teaching to the test), it was not entirely consistent with the developer's intended use of the test.

4.4.2.2 Claim 2: Decisions (values, equitable)

The decisions that are made based on the basis of the interpretations of the HSK take into consideration the educational and societal values and relevant regulations, and are equitable for the stakeholders affected by said decisions.

4.4.2.2.1 Decisions for educational purposes

According to the official documents issued by the Ministry of Education (MOE) and Hanban, the HSK can be used to motivate students if it becomes a prerequisite to degree programs at Chinese universities or for obtaining scholarships. In the test-taker participants EdTKi's and EdAYa's university, international students are required to pass HSK 4 upon entering their programs of studies. Regarding these HSK requirements, test-takers' interview results indicated that such requirements motivated them to learn Chinese; however, in some cases, the test was only an incentive for short-term goals and did not increase students' interest in studying Chinese. As EDTTKi said, "many students in our CSL class choose to take the HSK for these requirements, not because they really want to know their proficiency level nor were they more motivated in learning the Chinese language." He continued by saying that, "When I entered my program, I found the courses and textbooks were so different than the test. They were too hard for me. Although I passed HSK 5 (I forgot a lot after that), it doesn't mean I can survive in my courses." This finding not only showed that learning to the test could lead to artificial proficiency gains, but also raised the question of whether it was appropriate to use HSK scores in the universities' recruitment process.

The interview with the administrative participant EdAYa revealed that there were generally three reasons for establishing the HSK certificate as a prerequisite for international students in their degree programs. First, it complied with the MOE and Hanban policies.

According to their guidance, EdAYa's institution used HSK 4 as the standard requirement for most programs. For certain programs (e.g., Law) that required higher linguistic proficiency, the programs raised the cut-off score. Second, the HSK score could become a gatekeeper or a threshold to screen candidates. As EdAYa explained,

> In line with the global "Chinese fever" phenomenon and the PCI policy, an increasing number of international students have entered in the Chinese universities. For example, over 4000 international students have applied to our institution this year but our positions are limited. Due to the large number of applicants, the admissions experienced heavy workload. In order to reduce this workload, we used the HSK score as a threshold to select qualified candidates. Thus, we started to use the HSK Level 4 above 180 as the minimum requirement.

This argument is consistent with HSK's purposes as stated by the test developers, which is "to provide a reliable reference educational institution's decision-making concerning recruiting students" (Office of Chinese Language Council International, 2010, p.1). Third, using the test as a threshold could motivate students to study Chinese and to increase their success in future study or work. For example, EdAYa noted that her institution provided the test-prep course for HSK 6 as a required course for senior year students. The purpose was to prepare them to fulfill their future needs, such as applying for CI scholarship, Chinese government scholarship, and/or job hunting.

In sum, there were direct and indirect educational applications of the HSK. The direct use involves making high-stakes decisions about students based on their HSK levels, as it provides a practical and convenient tool for the admissions staff. The indirect use highlights the motivational role the test can play in the educational context.

4.4.2.2.2 Decisions in business setting

The rapid development of China's economy and its openness to the world has boosted the number of multinational enterprises or organizations in China. In workplaces that require

Chinese communication with clients, the employees' ability to speak and read in Chinese is essential, especially for managerial or executive positions. However, compared with uses in educational contexts, the HSK certificate was not often listed as a prerequisite for employment.

The interview data with the test-taker participant EDTTKi and the test user participant in business context BuWa revealed that having satisfactory Chinese proficiency in the workplace has several implications, such that employees would 1) have a competitive edge, 2) be able to competently fulfill their job responsibilities, and 3) have a higher likelihood of promotion. However, although adopting the HSK certificate as a recruitment prerequisite is considered an advantage, it is not mandatory for recruitment in some companies. This implies that an HSK-requisite policy did not assist employers with their selection procedures. An example was given by BuWa,

> In recent years, our company is seeking to expand the market in China. In light of this, employees who are good at Chinese are more desirable. Besides, Chinese is a basic requirement for employees seeking to be promoted to supervisory positions. Although Chinese is not pertinent to some positions (e.g., sales) at the moment, possessing a certain level of Chinese is considered a key advantage. We provide education aids for all the employees to encourage them to learn Chinese.

Furthermore, regarding to evaluation tools of the Chinese proficiency, BuWa stated that,

> Certificates, such as the HSK and/or other Chinese tests, are not required when they are applying to our company, but they are a plus. Applicants who have them might have more opportunities than those who don't when it comes to being hired or for future promotion. A certificate in Chinese is definitely beneficial.

When asked about the credibility of the HSK, he stated without concern that the HSK accurately reflected test-takers' language abilities. EDTTKi also confirmed the authority of the HSK in job applications, such that "some companies use their internally developed tests to assess applicants' Chinese proficiency, but I think they also acknowledge HSK scores."

Overall, employers were well aware of the importance of their employees' Chinese proficiency. However, it was not a major concern for recruitment purposes, although certified applicants may have an advantage. Thus, it can be concluded that employment and promotion decisions in the business context were not noticeably affected by individuals' HSK scores.

4.4.2.3 Claim 3: Interpretations (meaningful, impartial, generalizable, relevant, sufficient)

The interpretations of test takers' overall Chinese proficiency are meaningful with respect to the Scales, the curriculum objectives and the test specifications. The interpretations are fair to all test takers, realizable to the Chinese language use domain in which the decision is made, and are relevant to and sufficient for the decisions that are to be made.

In the context where language tests are used as policy instruments, test constructs are authorized by policy and context factors (McNamara, 2010; Weir, 2005). The results of this study showed that the implementation of the "Promoting Chinese Internationally" policy had significantly influenced the HSK reform. Regarding why a new HSK was developed, Xie (2011) explained that,

> The structure and scoring system of the old HSK is very complex, the length of the test is long, and the content is out of the scope of the syllabus. Most importantly, it is isolated from CSL/CFL teaching and learning. It had become a hurdle for promoting Chinese internationally, particularly for less proficient learners. The reform is imperative. (p.13).

Accordingly, a series of reforms have been carried out in terms of the test's content and format. In comparison with the old HSK, the new HSK has only 6 levels (5 levels fewer). It also reduced the number of required characters (8% fewer) and vocabulary/character combinations (38% fewer). The number of questions and the level of difficulty in the new HSK were also reduced. The Internet-based version was developed to include Chinese character input systems in the written tasks. In addition, the assessment's focus shifted from language knowledge to a broader definition of language abilities, particularly in the listening and speaking sections. These changes

have met the new PCI situation and greatly encourage learners with lower Chinese proficiency to participate in the test. As Li (2012) stated, "the revised HSK has played a positive role in PCI, and it also led to changes in the CSL/CFL teaching and learning models" (p.192).

However, the findings from the interview data with teachers and test-takers challenged these claims. The CSL teacher participant EdTHu asserted that "the levels of the new HSK is limited, especially for the advanced levels; it cannot be an effective evaluator that shows the differences in test-takers' proficiencies, which was something the old HSK was competent in doing." Another CSL teacher participant, EdTPa, also showed her concern toward the HSK as an indicator of Chinese proficiency. She asserted that there was an insufficient number of questions that assessed students' ability to read and write Chinese characters at the elementary levels. Regarding the higher levels, she added that,

> While students passed HSK 5, even HSK 6, they still find it is extremely difficult to survive while studying (or working) in a pure Chinese environment. They may have a good listening and reading skill, but speaking and writing is very challenging for them.

The test-taker participants demonstrated their contrasting points of view towards Claim 3 in terms of their regional differences. EDTTKi and EDTTLu were from the same country, which has a deep-rooted test-driven culture. They hold the opinion that the HSK scores cannot indicate one's Chinese proficiency because they believed that high scores could be obtained in a short period of time through test-taking skills. They also noted that some test-takers only wanted to show their superiority over applicants without HSK certificates in the job market. However, EDTTKi also stated that a colleague possessed a HSK 3 "but he cannot communicate in Chinese at all." The other interviewees, on the other hand, believed the new HSK could objectively reflect learners' proficiency level and was fairer to test-takers than the old HSK. For example, in the past, test-takers from countries with Chinese-influenced writing systems (e.g., Japan) had a

greater advantage in the old exam. This is because it was very easy for them to understand the meaning of the vocabulary and to write the characters. As the new version includes a Chinese input system, test-takers from Western countries can also easily and accurately write in Chinese during the test. This can be seen in EDTTZh's response, who said that "the input system saved me a lot of time on writing, and my writing score significantly improved in the new HSK."

In this section, findings on the actual uses of the test confirm the prominent role played by the HSK in implementing the PCI policy. The findings also provided further evidence (i.e., backings and rebuttals) concerning the interpretation of the HSK score and its use. The HSK developers have made efforts to achieve the intended consequences, such as through revising and innovating of the test. However, decisions based on inaccurate interpretations are ethically, legally, and socially unacceptable and tend to bring about adverse consequences (Kane, 2013; Messick, 1996). To address this issue, the test developers and test users should have shared responsibilities (Xi, 2010). In other words, the test developers need to provide accurate information to test users, for instance, by enhancing the validity and reliability of the test and ensuring that the level/score reflect the accurate proficiency of test-takers; while test users are accountable for the consequences of the decisions, for example, they must understand the implications of applying an inappropriate cut-off score for admissions, issues relating to adjustments in the admission criteria.

4.5 Conclusion

Due to the HSK reform and a massive promotion campaign led by the PCI policy, the number of CSL/CFL learners and HSK test-takers has risen substantially. The status of CSL/CFL worldwide has generated further development and research on the HSK. Overall, by adapting Bachman and Palmer's (2010) AUA framework in the HSK context, this study closely examined

the consequential validity of the HSK in educational and societal contexts. The study also helped identify the values underlying the multiple interpretations and uses of the test.

The test developers designed the HSK to achieve two interrelated goals, namely: 1) to act as a reference for educational and social decisions centered on individuals' Chinese language proficiency; and 2) to promote CSL/CFL teaching and learning. This study's findings provided evidence that this has been achieved to some degree. For the students in this study, the HSK seemed to play a motivational function. However, for teachers, the test did not appear to affect and influence their teaching practices and beliefs in any major way. Furthermore, the HSK achieved its intended consequence to a great degree in terms of providing test users useful information for making decisions in the educational context. The HSK reform also reflected the test developers' attempt to enhance the test's quality. However, there still are limitations with the HSK test design (e.g., the limited range of language use contexts). These limitations may contribute to unintended negative consequences for students (e.g., focusing only on test-taking skills), and may also give rise to validity concerns as well as ethical concerns about the test (e.g., focusing only on multiple choice questions, fairness issues). Ultimately, the intended consequence of promoting CSL/CFL teaching and learning has only been achieved to a limited extent. Based on the findings of Study 1, themes according to the AUA framework were generated (see Table 4.4). Accordingly, these themes were used as the foundation for developing research instruments in the subsequent studies. More specifically, the findings in Study 1 laid the groundwork for the development of a classroom observation guide, two hypotheses on the effect of washback on teaching and learning, and two questionnaires (one for teachers and another for test-takers) in Study 2, which elicited participants' opinions of the test, test use, and impact and teaching/learning strategies. The themes from Study 1 also informed the creation of the two

questionnaires (one for score users in academic settings, and one for those in non-academic settings) in Study 3, which elicited HSK score users'[14] perspectives on score interpretation, and decisions made based on HSK levels/scores. Details of these instruments will be discussed in the next two chapters.

[14] According to Bachman and Palmer (2010), "test users" mean "those making decisions based on the assessment" (p.135); in the current study, this refers to administrative staff who use the HSK as a criterion to make admission/employment decisions, in both academic and non-academic settings.

Chapter 5 Study 2 (Improving language teaching and learning through language testing: A washback study on the HSK)

5.1 Introduction

According to the visual diagram of the current MMR study (see Figure 3.1), Study 2 was the second phase of this research. By using the argumentative conceptual framework for the HSK established in Section 2.8 and the instruments designed from Study 1's themes, this second study explored the complex nature of washback effects at the micro (classroom) level in the educational context from both the teachers' and test-takers' perspectives.

5.2 Background

In the area of second language education, washback, "the influence of a test or other evaluation procedures on teaching and learning" (Turner, 2001, p.138), has been the focus of an increasing number of theoretical and empirical research studies concerning high-stakes tests. Existing studies revealed that washback is not a monolithic phenomenon; it is instead a highly complex process involving various mediating factors among multi-stakeholders. Although attention has been primarily paid to teachers, teachers' beliefs, and their pedagogical practice in the classroom, more work still needs to be carried out in order to understand how these factors interact with each other. In addition, compared to the widely investigated washback effects on teaching, "less emphasis has been given to learners" (Watanabe, 2004, p. 22). In fact, only a few studies have been conducted on select exams on learners, mostly about the College English Test Band 4 (CET 4), International English Language Tests (IELTS), but rarely about other tests.

As the central national standardized test of Chinese language proficiency for non-native speakers, the HSK plays a vital role in certifying language proficiency for higher education and professional purposes. However, despite HSK's status, very few empirical studies have been

conducted to verify its consequential validity, particularly in terms of the washback effect (Huang, 2013; Huang & Li 2009). Among the limited washback studies, most were based on the old HSK (before 2009), and they failed to achieve a thorough investigation of how washback occurs in this context. Considering that the HSK is a testing program with the philosophy of "以考促教，以考促学" [improving teaching and learning through testing], evidence about washback on teaching and learning is a legitimate, necessary, and crucial part of the test validation process in the PCI context.

Consequently, in order to fill this research gap using the argumentative conceptual framework for the HSK, Study 2 (see Figure 3.1) aims to 1) explore the HSK test-takers' and CSL/CFL teachers' perceptions concerning the HSK, its washback effects, and its use; 2) uncover the relationships between these perceptions and the teaching/learning practices; 3) model HSK's washback effects on teaching and learning; and 4) investigate HSK's washback effects in the real classroom setting. The above objectives will allow the present researcher to evaluate the consequential validity of the HSK in educational settings.

5.3 Methodology

This study has two sets of research questions. Each set addresses issues that concern a different stakeholder group: RQ1 for HSK test takers and RQ2 for CSL/CFL teachers.

RQ1: What are HSK test takers' perceptions concerning the HSK content, use, and impact? What are their perceptions about whether the HSK score/level reflects their real proficiency? What are the relationships between these perceptions and their test preparation practices?

RQ2: What are CSL teachers' perceptions concerning HSK content, use and impact? How does the potential influence of the HSK manifest in their classroom practices? What are the relationships between these perceptions and their teaching practices?

A sequential explanatory mixed methods design was adopted (Creswell & Plano Clark, 2011, 2015) in Study 2 of the current MMR study. Figure 5.1 presents the research design of this study. The primary emphasis of the design was on the quantitative aspects (Phase 1 of Study 2) through questionnaire data collection. It was used to quantify opinions and behaviors and examine the relationships among factors to generalize results from a larger sample population. In Phase 2 of Study 2, the qualitative data were collected through classroom observation to help explain the quantitative results.The design allows for the cross-validation of the methods and a strengthening of the inferences of the results (Creswell, 2009, 2015).

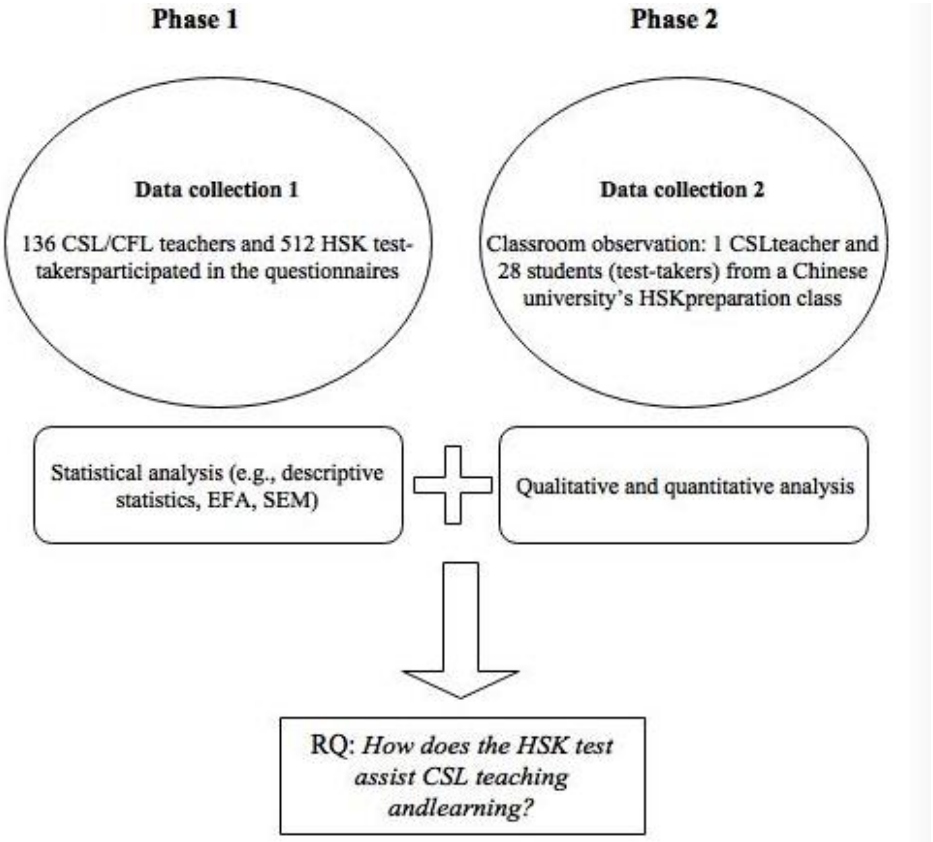


Figure 5.1. Research design of Study 2

5.3.1 Hypothesized washback models of test-takers and teachers

In order to explore the nature of HSK's washback effects on teaching and learning, two structural models of washback effects on test-takers and teachers were hypothesized. The models were developed based on Study 1's findings and on the previous washback research in LT, as well as attitude and behavior theory in educational psychology (e.g., Wang, J, 2010; Xie & Andrew, 2013). As illustrated in Figure 5.2, the test-takers' model hypothesized the relationship between their perceptions of the test (e.g., the test design, use, functions, impact, and test-taking expectations) and their test preparation practices (e.g., test-taking strategies used). The null hypothesis of this model was that test-takers' perceptions of the test do not influence their test preparation. On the other hand, the teachers' model (see Figure 5.3) was used to hypothesize to examine the relationships between teachers' perceptions (e.g., the test design, test use and impact) and their teaching practices (e.g., teaching methods). The null hypothesis was that the teachers' perceptions towards the test do not influence their teaching practices.

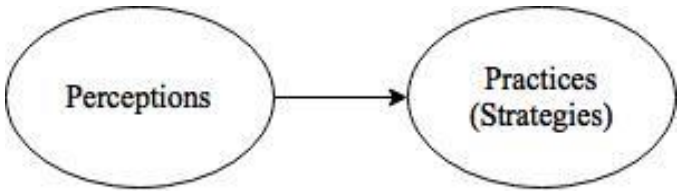


Figure 5.2. Hypothesized structural model of washback on test-taking strategies

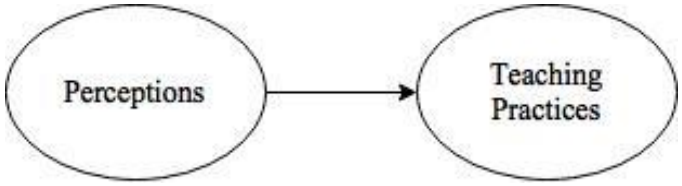


Figure 5.3. Hypothesized structural model of washback on teaching practices

5.3.2 Participants

As shown in Figure 5.1, in Phase 1, 136 CSL/CFL teachers and 512 HSK test-takers participated in the questionnaires. Based on the "maximum variation sampling" technique (Patton, 2002) and the "purposive sampling" method (Maxwell, 1996), most participants (over 90%) were from 7 university CSL programs, while the remaining were from CFL programs (such as university credit courses, HSK oversea test centers, and Confucius Institute programs), language schools, and HSK preparation classes. The variety of school settings and students' proficiency levels was of considerable importance, as individuals from different groups may experience varying levels and types of washback effects (Alderson & Wall, 1993). Among the 512 test-takers, 27.9% were male and 72.1%% were female. Of the 136 teachers, 16.9% were male and 83.1% were female. More details on these participants will be provided in the Results section. In the second phase, the participants for the classroom observation consisted of 1 CSL teacher and 28 students (test-takers) from a Chinese university's HSK preparation class[15]. The rationale of choosing this research site was that 1) all of the students were advanced-level CSL learners and were going to pass HSK6 to fulfill their graduation requirement (i.e., the test had a potential direct impact on their learning); and 2) the course was designed for HSK preparation rather than a regular CSL course, which meant that the teacher was knowledgeable about the test and that the test had a potential direct impact on the class. Thus, participants of this course were recruited for Study 2, as teachers and students in the HSK preparation class would have more knowledge about the HSK than their peers in a standard Chinese language program.

5.3.3 Instruments

[15] This HSK preparation course is a university credit course that is offered to undergraduate students whose major are Chinese language and literature. This university's policy requires all the international students in this major to pass HSK 6 when they graduate.

To obtain the CSL teachers and test-takers' perceptions of the HSK design, test use, test impact, and the participants' teaching and learning practices (e.g., test preparation strategies), two questionnaires (one for teachers, and the other for test-takers), were developed based on the results from Study 1 and previous washback research (e.g., Bailey, 1996; Green, 2007; Sun, 2016; Turner, 2005; Xie, 2010; Wang, J., 2010; Wang, S., 2013) (please refer to Appendices 2 and 3 for the survey instruments). The test-takers' questionnaire comprised 3 sections. Section 1 included demographic questions that elicited participants' biographical data. Section 2 asked participants to report on their perceptions of the test (e.g., test design, test use, test expectation, difficulty level, test functions, test impact), and Section 3 asked them about their test preparation strategies. Open-ended questions were included at the end of both Sections 2 and 3. These sections elicited participants' comments on test aspects and strategies. A strength of this open-ended question format was that it elicited responses that might not to be anticipated (Bachman & Palmer, 2010). Section 2 and 3 were then scored on a 6-point Likert scale of agreement from 1 for strongly disagree to 6 for strongly agree. The teachers' questionnaire also had 3 sections: Section 1) demographic questions; Section 2) the perceptions towards the test (e.g., test design, test use, test functions, test impact); and Section 3) their teaching methods and practices. Similar to the test-takers' questionnaire, it also used 6-point scale items for the last two sections with open-ended questions at the end of each sub-section. The items in Section 2 of both questionnaires overlap, which could facilitate the comparison of test-takers' and teachers' perception on HSK's uses and washback effects.

The following tables (Table 5.1 and Table 5.2) illustrate the questionnaires' constructs and their links to the research questions. The constructs consisted of 4 scales and 9 sub-scales in each questionnaire. The number of items related to each sub-scale is listed in the far right-hand

column. As shown, there were 43 perception items on the HSK, 26 items on test preparation practices in the test-takers' questionnaire, and 31 perception items on the HSK, 14 items on classroom instruction in the teachers' questionnaire. Some of the questions were cross-referenced in both questionnaires.

Table 5.1

Linking Questionnaire for Test-takers to RQ1

RQs	**Scales**	**Sub-scales**	**Item numbers in the questionnaire**	**Items**
RQ1-1	Test design	Test format	A1- A5	5
		Test content	A6- A8	3
		Test nature	A9- A12	4
	Test use	Test goals	P1- P7	7
		Test functions	V6- P9	4
	Test impact	Test effects	E1- E6	6
RQ1-2	Validity	Reflecting real proficiency	V1- V5	3
RQ1-3	Relationships	Perceptions	All items from A, P, V, T, D, and TE sections	26
		Test preparation strategies	TPP1-TPP26	26
		Test outcomes	Q11, Q12-1, Q12-2, Q12-3, Q12-4, Q14	5

Table 5.2

Linking Questionnaire for Teachers to RQ2

RQs	Scales	Sub-scales	Item numbers in the questionnaire	Items
RQ2-1	Test design	Test content and format	TA1-TA7	7
		Test nature	TA8- TA12	5
	Test use and impact	Test goals	TH1- TH6	6
		Test functions	TV1- TV6	6
		Test effects	TE1- TE7	7
RQ2-2	Classroom instruction	Teaching methods	TM1- TM9	9
		Teaching practice	TP1- TP5	5
RQ2-3	Relationships	Perceptions	All items from TA, TH, TV, TE, and TM sections	31
		Classroom instruction	TM1- TM9, TP1- TP5	14

The other instruments pertinent to Study 2 were the Classroom Observation Guide (see Appendix 4), field notes, and post-observation chats. The Classroom Observation Guide was developed based on three sources, namely Watanabe (1996), the findings of Study 1, and the analysis of the questionnaire data on teaching, learning, and assessment practices in the CSL classrooms. Also, the classroom observation helped determine how test-takers and teachers perceived the test, test use, and test impact, as well as whether their understanding of their test preparation or pedagogical strategies was reflected in their actual practices.

5.3.4 Procedure

To enhance the validity and reliability of the questionnaires, the following procedures were carried out before data collection. First, CSL/CFL teachers (n=3) who had teaching experience in both China and abroad and CSL/CFL learners (n=4) who had taken the new HSK

were invited to be part of a focus group. In this group, participants read and responded to the questions, and also commented on the clarity and practicality of the items. Revisions were made based on their feedback. For example, the option "not familiar with it" was added to some items (i.e., TP3 and TA12 in test-takers' questionnaire, and TA1- TA12 in teachers' questionnaire). The focus group discussions also indicated that CFL teachers were less familiar with the HSK as compared with the CSL teachers, and that fewer CFL students took the HSK than their counterparts in CSL programs. Accordingly, due to feasibility, the main research sites were from CSL contexts. The revised questionnaires were then pilot tested. Using an iterative process, 10 CSL teachers and 20 HSK test-takers answered the questionnaires. The Cronbach alpha internal consistency coefficients were .922 and .892. Based on the pilot test results, further minor revisions were made before the questionnaires were finalized.

As the researcher of the study, my experiences as a CSL and CFL teacher in China and North America and as an HSK item writer have given me insider knowledge concerning the research site contexts and potential participants. This knowledge was useful for recruitment purposes. Using my personal contacts, a call for volunteers to complete two surveys via e-mail were sent to CSL/CFL program coordinators, teachers, students, and HSK test centers. The recruitment criterion for HSK test-takers was that the participant needed to have taken the HSK in the past two years (i.e., their test scores were still valid[16] in the survey studies); the criteria for teacher participants were that they needed to be familiar with the new HSK test and had taught test preparation courses. From September to December 2016, the paper version of the two questionnaires was administrated in CSL/CFL classrooms, and the online version of the questionnaires was also made available on Survey Monkey upon request. Ultimately, 136

[16] HSK scores are valid for two years after the test date.

CSL/CFL teacher questionnaires and 512 HSK test-taker questionnaires were completed. The sample size was determined based on the data analysis methods used in this study (see Section 5.2.5).

To help interpret the quantitative results, a HSK preparation class was observed 6 times during April 2017. Each observation lasted 45 minutes. After each observation, a 10- minute post-observation discussion took place; notes were made both during and after these discussions. The audio-recorded observation and the post-observation chats offered a snapshot of classroom activities and provided insight on the participants' HSK perceptions and their actual classroom practices.

5.3.5 Data Analysis

The questionnaire data were analyzed quantitatively with the use of SPSS 24.0 and Amos 24.0. First, the data were entered into the SPSS program manually. Then, SPSS' Missing Value Analysis (MVA) program was used to check missing rates and missing patterns, while the Expectation Maximization (EM) algorithm was used to impute missing values. The results revealed that the test-takers' questionnaire had less than 3% missing values and the missing pattern did not show an obvious pattern. In the teachers' questionnaire, items (n=7) that exceeded 5% missing values were excluded, and the mean substitute method was used for items that has the option "not familiar with." Outliers were examined at the item level through stemleaf, boxplot, and z-score for univariate outliers, as well as through Regression Mahalanobis Distance for multivariate outliers. Furthermore, by examining skewness, kurtosis statistics, and the graphics, no extreme non-normal distribution was found in both questionnaire data sets. After these procedures, 484 test-taker cases and 133 teacher cases were retained for further analysis. Subsequently, the reliability of the questionnaire data was tested. The Cronbach's α for all the 6-

point Likert-scale items in both questionnaires were 0.971 and 0.935, respectively, which attested to the satisfactory internal consistency of the questionnaires.

Quantitative analysis of the current study included descriptive and inferential statistics. Descriptive statistics were used to summarize teachers' and test-takers' perceptions towards the HSK's design, use, and impact, along with their teaching and learning practices. The item-level exploratory factor analysis (EFA) was then performed to identify how the questionnaire items clustered, which allowed the researcher to explore patterns of correlations among items and to verify whether the items' loading correspond to the intended scales and subscales (see Table 5.1 & 5.2 above). More specifically, EFA can be used for item reduction in questionnaire analysis and was especially useful for the teacher's questionnaire, which had a smaller sample size. The factors from the EFA were then used as observable indicators in modeling the latent measurement models via structural equation modeling (SEM) in AMOS 24.0. Afterwards, the SEM, a technique that assesses the structural interrelationships among observed and latent variables through a confirmatory and hypothesis testing approach, was performed (Byrne, 2001). This was because unlike experimental language teaching and learning studies conducted in lab environments, the causal relationships of potentially complex washback effects can be difficult to establish. Thus, SEM was identified as the appropriate model method "with the fewest possible variables, but to achieve the maximal explanation power of reality" (Xie, 2010, p.132).

In terms of analyzing the data from classroom observations, frequency counts were first applied based on the categories developed in the observation guide. The percentage of time spent on each category was then calculated for each classroom activity. Third, the interview (post-observation chat) data were synthesized by focusing on the themes pertaining to the research questions and the findings of Phase 1. Finally, both types of data (quantitative and qualitative) in

the two phases were triangulated to better understand the phenomenon.

5.4 Results

In this section, the results of the data analysis are presented to address the two research questions.

5.4.1 Results of test-takers' questionnaires

5.4.1.1 Descriptive statistics

Among the 484 cases, 72.3% were female and 27.7% were male. Furthermore, for 87.6% of the total cases, ages ranged from 18 to 30, and over 70% of them had been learning Chinese over 2 years. Most of the test-takers have participated in at least two types of CSL/CFL courses; for example, 70.7% of them had taken university-level credit courses, and 20.7% of them had taken the HSK test prep courses. In terms of test-takers' motivation for learning Chinese and taking the HSK, 52.3% stated it was because they wanted to learn a new language, while 84.7% did so in order to work/travel in Chinese-speaking areas.

The descriptive statistics of the perception items are grouped below in Table 5.3 based on the aforementioned intended sub-scales.

Table 5.3

Descriptive Statistics of the Test-takers' Perceptions on the HSK

Sub-scales	Minimum	Maximum	No. of items	Mean	SD
Test format	1	6	5	4.37	1.03
Test content	1	6	3	4.35	0.94
Test nature	1	6	4	4.41	0.91
Test goals	1	6	7	4.66	1.20
Test functions	1	6	4	4.73	0.98
Test effects	1	6	6	4.47	1.03
Reflecting real proficiency	1	6	5	4.59	1.05

Table 5.3 shows that the mean of the items on the test design are similar. In general, these statistics indicate that test-takers held a moderately positive view of the format, content, and nature of the HSK. They were relatively satisfied with the overall content and format of the test and noted that the current HSK version focused more on communicative functions of the language than linguistic knowledge, which is aligned with the goal of the new HSK development. However, they held different opinions regarding the new elements of the revised HSK. For example, they thought the inclusion of the Chinese input systems[17] weakens learners' ability to write Chinese characters and that the HSKK should be included in the HSK.

In terms of the HSK's use and impact, the mean of the perception items ranged from 4.33 to 4.80. From the test-takers' perspective, they strongly agreed that the HSK can measure their Chinese proficiency and that the test provided useful feedback for their Chinese language learning. They thought that taking the test motivated them to learn Chinese and also encouraged them to use Chinese in their daily life. On the other hand, they believed that the HSK overemphasized the memorization of vocabulary and language rules, such that the HSK forced students to study to the test (m=4.48) and forced teachers to teach to the test (m=4.39). They agreed with the values of the HSK, namely that by obtaining a HSK certificate, they will enhance their competitiveness in future endeavours (e.g., scholarship applications, and job seeking or promotion).

A one-way between-subjects ANOVA was conducted to compare the different proficiency groups' (i.e., elementary level, intermediate level, and advanced level) perceptions concerning whether the HSK score/level reflected the test-takers' real proficiency. The results

[17] Chinese input systems, also called Chinese input methods, are methods that allow a computer user to input Chinese characters. Mostly, they fall into one of two categories: phonetic readings or root shapes.

indicated that all groups deemed the test level and score to be a generally appropriate indicator of their overall Chinese ability. However, they believed the HSKK (the oral test) could not fully reflect their speaking ability. There were significant differences among proficiency levels on reflecting their listening ability [$F(2, 468) = 4.23$, $p < .05$], reading ability [$F(2, 468) = 4.44$, $p < .05$], and writing ability [$F(2, 468) = 4.20$, $p < .05$]. A Tukey post hoc test revealed that "the test level/score is an appropriate indicator of their listening and reading Chinese ability". This was found more in the advanced group than the groups with lower proficiency levels.

The results of the descriptive statistics analysis of test preparation strategies (see Table 5.4) showed that test-takers spent more time on regular learning[18] than on test-specific learning[19]. For example, they did not take HSK prep courses or hire HSK private tutors (m=4.07); for speaking, they spent more time on communicating with Chinese native speakers whenever possible (m=4.54) rather than practicing HSKK topics from the past HSKK test papers or mock test papers. Similarly, when preparing for listening, they watched Chinese TV and/or listened to Chinese radio broadcasts (m= 4.55). However, their perception of the HSK also affected the strategy they used. For instance, in order to choose appropriate learning methods, they analyzed HSK papers to identify the question types and analyzed score distributions to judge the relative importance of each section. They also used test-taking strategies during the test to achieve high scores. For example, during the writing test, they tried to avoid grammar and writing mistakes (m=4.41), and adopted more advanced vocabulary and structures (m=4.37). The results showed

[18] The researcher defined regular Chinese learning to include CSL courses, learning from others, or self-learning on a daily basis.

[19] Test-specific learning reflected participants' learning largely on the basis of the test's impact and included strategies and processes that affected their success on the HSK, such as taking test-prep courses and practicing simulated exam papers.

that speaking had the highest standard deviation, indicating there was a large variation in strategies used in this section.

Table 5.4

Descriptive Statistics of the Test-takers' Test Preparation Strategies

Strategy	Minimum	Maximum	No. of items	Mean	SD
General strategies	2	6	9	4.31	0.96
Speaking	1	6	2	4.34	1.22
Listening	1	6	5	4.43	1.08
Reading	1	6	6	4.37	1.02
Writing	1	6	4	4.31	1.06

5.4.1.2 Exploratory factor analysis (EFA)

5.4.1.2.1 EFA on test-takers' perceptions of the test

A principal axis factoring (PAF) extraction with a Varimax (orthogonal) rotation methods and the mini-eigenvalue equals one criterion method was conducted on the perception items, which are 45 questions from the "perceptions towards the test" section of the questionnaire. An examination of the Kaiser-Meyer Olkin measure suggested that the sampling was adequate (KMO= .925). The Bartlett's Test of Sphericity reached statistical significance, thus indicating that the correlations were sufficiently large for EFA (p= .000). The initial EFA produced a 12-factor solution. Along with the inspection of the scree plot, the following criteria were employed in data reduction: 1) more than 3 items in each factor; 2) factor loadings above 0.5; and 3) excluded complex cross loadings items (e.g., cross loading on two or more items with the value above 0.3). Ultimately, 6 meaningful factors accounting for 66.65% of the total variance were generated, achieving both structural simplicity and substantive meaningfulness. The loading of each item on individual factors is listed in Table 5.5. Five items loading on Factor 1 represented test content and nature, 4 items that load on Factor 2 were related to test goals, 3 items loading

on Factor 3 accounted for language proficiency, 4 items loading on Factor 4 were devoted to test effects, 5 items loading on Factor 5 were linked to difficulty levels, and 3 items loading on Factor 6 were related to perceptions of the HSK as a prerequisite/exit requirement. Table 5.6 presents the names of the factors, items in each factor, number of items, and their reliability (alpha value). Compared with the intended scales, the factoring results corresponded well with the intended scales of the original questionnaire design.

Table 5.5

Factor Loading of Test-takers' Perceptions of the Test

	Factor					
	1	2	3	4	5	6
A11	**.691**	.209	.147	.062	.129	.081
A10	**.675**	.188	.149	.183	.032	.143
A12	**.618**	.085	.091	.122	.110	.227
A9	**.587**	.133	.156	.183	.111	.056
A8	**.562**	.159	.092	.253	.023	.169
P6	.154	**.704**	.034	.091	.097	.178
P4	.113	**.699**	.104	.099	.037	.021
P7	.192	**.694**	.074	.095	.151	.114
P5	.140	**.652**	.070	.126	.083	.094
V2	.149	.140	**.760**	.161	.096	.108
V1	.121	.078	**.742**	.133	.147	.065
V3	.238	.135	**.645**	.156	.161	.085
E5	.097	.150	.153	**.695**	.146	.246
E6	.181	.095	.095	**.641**	.187	.230
E2	.273	.185	.028	**.625**	.110	.077
E4	.305	.080	.197	**.557**	.101	.064
D4	.095	.053	.039	.114	**.687**	.132
D1	.106	.150	.171	.019	**.661**	.043
D5	.038	-.061	.063	.058	**.647**	.070
D3	.089	.121	.024	.144	**.644**	.149
D2	.126	.139	.182	.136	**.636**	.045
R6	.206	.074	.038	.208	.150	**.660**
R2	.180	.188	.108	.228	.212	**.587**
R4	.219	.220	.179	.247	.157	**.541**

Table 5.6

EFA Factors of the Test-takers' Perceptions of the Test

Factors	Items	Item No	Reliability alpha
F1: Test content and nature	A11, A10, A12, A9, A8	5	.854
F2: Test goals	P6, P4, P7, P5	4	.773
F3: Indicating proficiency	V2, V1, V3	3	.863
F4: Test effects	E5, E6, E2, E4	4	.849
F5: Difficulty levels	D4, D1, D5, D3, D2	5	.797
F6: Requirements	R6, R2, R4	3	.872

5.4.1.2.2 EFA on test-takers' test preparation practices/strategies

To investigate how the strategy items (i.e., 26 items from the "test preparation strategies" section of the questionnaire) clustered, an EFA was performed according to the procedures detailed above. An examination of the Kaiser-Meyer Olkin measure suggested that the sampling was adequate (KMO= .943). The Bartlett's Test of Sphericity reached statistical significance, which indicated that the correlations were sufficiently large for EFA (p= .000). The initial EFA produced a 5-factor solution. After applying the criteria mentioned above, 3 meaningful factors that accounted for 58.672% remained. Table 5.7 presents the loading of each item on each factor. There were 5 items loading on Factor 1 for general cognitive strategies, 4 items loading on Factor 2 for listening strategies, and 4 items loading on Factor 3 for reading and writing strategies. Table 5.8 portrays the names given to the factors, the items in each factor, the number of items, and their reliability (alpha value). Compared with the intended scales, the factoring results corresponded well with the intended scales of the original questionnaire design.

Table 5.7

Factor Loading of Test-takers' Test Preparation Strategies

	Factor		
	1	2	3
TP6	**.614**	.224	.159
TP7	**.581**	.236	.251
TP4	**.564**	.080	.099
TP9	**.547**	.192	.178
TP5	**.542**	.095	.148
TP13	.331	**.663**	.254
TP12	.089	**.575**	.077
TP14	.254	**.568**	.282
TP16	.206	**.533**	.266
TP25	.201	.220	**.659**
TP24	.317	.253	**.614**
TP26	.281	.157	**.524**
TP21	.075	.216	**.501**

Table 5.8

EFA Factors of Test-takers' Test Preparation Strategies

Factors	**Items**	**Item No**	**Reliability alpha**
F1: General strategies	TP6, TP7, TP4, TP9, TP5	5	.799
F2: Listening strategies	TP13, TP12, TP14, TP16	4	.805
F3: Reading & writing strategies	TP25, TP24, TP26, TP21	4	.807

In summary, EFAs were used to determine the factorial structure of the questionnaire items. Six factors of test-takers' perceptions of the test and 3 factors of their test preparation strategies were identified.

5.4.1.3 Structural equation modeling (SEM)

Based on the result of the EFAs and the hypothesized washback model on test-takers in Section 5.2.1, a model of the HSK washback model of test-takers, examining the relationship of perceptions influencing test preparations, was postulated. This model involved 9 latent variables and 37 observed variables, as presented in Table 5.9. Among these, the 6 latent variables (i.e., CN, TG, IP, TE, DL, and TR) were the 6 perception factors generated by EFA, and they were considered as test-takers' perceptions of the test (the left-hand ellipse of the Figure 5.1); the 3 latent variables (i.e., TP1, TP2, and TP3) were the 3 strategy factors generated by EFA, and they were considered as test preparation strategies (the right-hand ellipse of the Figure 5.1). This hypothesized model was tested using SEM.

Table 5.9

Construct of Latent and Observed Variables in the Washback Model of Test-takers

Latent variable	**Observed variable**
Perceptions of test content and nature (CN)	Score report (V1)
	Fairness (V2)
	PCI policy (V3)
	Test specification (V4)
	Content (V5)
Perceptions of test goals (TG)	Job requirement (V6)
	Degree (V7)
	Job seeking (V8)
	Scholarship (V9)
Perceptions of indicating proficiency (IP)	Listening (V10)
	Overall (V11)
	Reading (V12)
Perceptions of test effects (TE)	Study to the test (V13)
	Teach to the test (V14)
	Improve teaching (V15)
	Use in daily life (V16)
Perceptions of difficulty levels (DL)	Writing (V17)
	Overall (V18)
	Speaking (V19)
	Reading (V20)
	Listening (V21)
Perceptions of prerequisite/exit requirement (TR)	Test-taking strategies (V22)
	Energy/money (V23)
	Extra work/pressure (V24)

General strategies (TP1)	Grammar (V25)
	Teacher's advice (V26)
	Memorize vocabulary (V27)
	Prep course (V28)
	Synonym (V29)
Listening strategies (TP2)	Repeat listening (V30)
	Go over options (V31)
	Watch TV (V32)
	Understanding materials (V33)
Reading & writing strategies (TP3)	Avoid mistakes (V34)
	Practice HSK writing (V35)
	Using advanced vocabulary (V36)
	Read questions (V37)

Before conducting the SEM, another round of data cleaning and iteration of the data following the same procedure mentioned in Section 5.2.5 indicated that the statistical assumptions were met and no significant violations were found. SEM was then carried out by using the Maximum Likelihood estimation method. Figure 5.4 presents the structural equation model generated by AMOS.

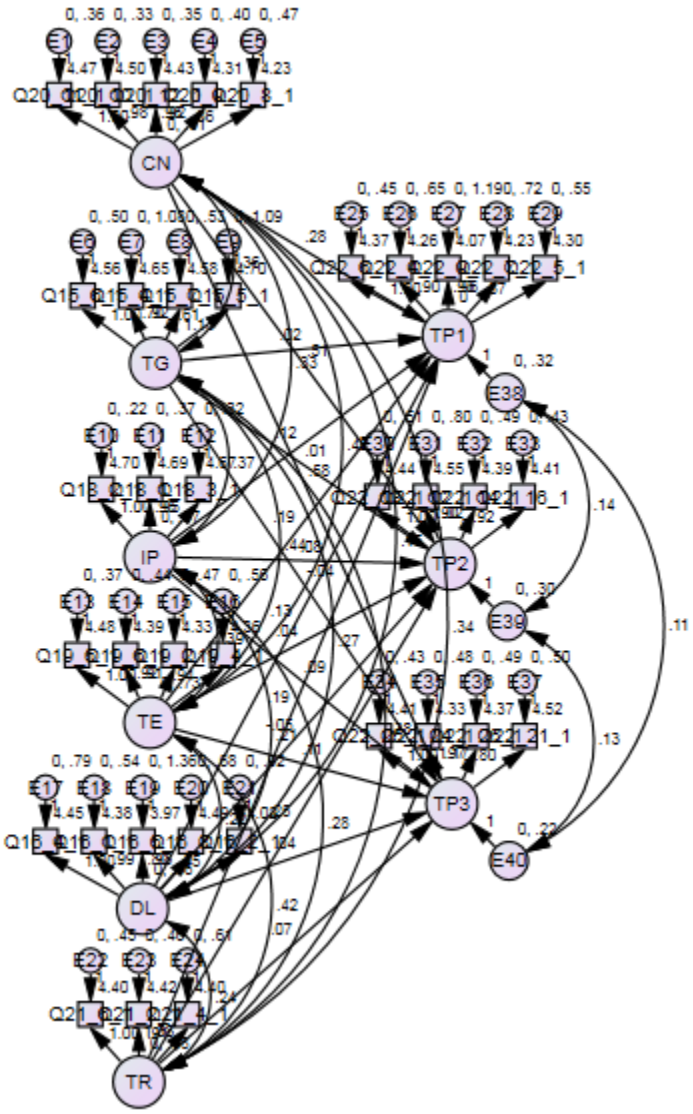


Figure 5.4. Structural equation model of washback on test-takers

The latent variables are enclosed in larger circles, and the items (observed indicators) are enclosed in squares, whereas the measurement errors are enclosed in smaller circles. The inter-related constructs of participants' perceptions of the test (i.e. CN, TG, IP, TE, DL, and TR) were connected to each other with double-headed arrows, which represents a pattern of intercorrelation. The inter-related constructs of test-taking strategies (i.e., TP1, TP2, and TP3) were also connected to each other with double-headed arrows. The single-headed arrows, leading

from the larger circles to the squares demonstrate predicable regression paths, and represent the link between the latent variables and observed variables, while the coefficients represent factor loadings. The single-headed arrows from each of the 6 larger circles on the left-hand side to each of the 3 larger circles on the right side represent regression relationships between perceptions of the test and test-taking strategies. The single-headed arrows from the small circles to the squares represent measurement errors associated with the observed variables. When evaluating model fitness, a Chi-square index of less than 2 and a root-mean-square error of approximation (RMSEA) that is greater than .06 represent a close fit of the model; additionally, the convention of goodness-of-fit index (GFI), comparative fit index (CFI), and Tucker-Lewis Index (TLI) above .90 was adopted as an indication of good model fit (Browne & Cudeck, 1989, 1993; L. Hu & Bentler, 1999). As seen in Table 5.10, the overall goodness of fit suggests that the model is an appropriate representation of the interrelationships of test-takers' perceptions on the test and their test-taking strategies, such that it provided strong evidence for accepting the hypothesized model.

Table 5.10

Goodness of Fit Summary for the Hypothesized Model of Washback on test-takers

Indices	**p**	**χ2**	**df**	**χ2/ df**	**IFI**	**TLI**	**CFI**	**RMSEA**
Values	.000	1134.82	593	1.914	.933	.924	.932	.043

Individual parameter estimates were examined after evaluating the model's fitness. Table 5.11 provides the unstandardized and standardized parameter estimates, standardized error estimates, and the *p* values.

Table 5.11

Parameter Estimates for the Model of Washback on Test-takers

			Estimate	S.E.	C.R.	P
TP1	<---	TG	.021	.042	.502	.615
TP1	<---	IP	.118	.058	2.037	**.042**
TP1	<---	CN	.282	.086	3.268	**.001**
TP1	<---	TE	.188	.072	2.599	**.009**
TP1	<---	DL	.125	.051	2.441	**.015**
TP1	<---	TR	.190	.079	2.396	**.017**
TP2	<---	CN	.514	.088	5.821	***
TP2	<---	TG	.010	.041	.251	.802
TP2	<---	IP	.078	.056	1.386	.166
TP2	<---	TE	.036	.071	.505	.613
TP2	<---	DL	-.050	.050	-1.008	.313
TP2	<---	TR	.253	.079	3.224	**.001**
TP3	<---	CN	.582	.081	7.188	***
TP3	<---	TG	-.036	.036	-.992	.321
TP3	<---	IP	.094	.050	1.869	.062
TP3	<---	TE	.111	.063	1.764	.078
TP3	<---	DL	.036	.044	.803	.422
TP3	<---	TR	.075	.068	1.090	.276

Note: *** represents $p < .001$

As shown in Table 5.11, there are 8 significant paths. According to Lei and Wu (2007), a higher value (between 0-1) of standardized factor loadings in the measurement model suggests a better indicator for the latent variable. This table indicates that the observed variables were satisfactory for their loading latent variables. For example, for the path from CN to TP3, the standardized coefficient value of .58 indicates that as the perceptions of the test content and nature increases by one standard deviation, their reading and writing test preparation strategies are expected to increase by .58 of a standard deviation; thus, TP3 could be predicted by CN. Overall, the SEM results indicate that the proposed model of washback on test-takers could be accepted to present relationships between their perceptions on the tests and their test-taking strategies. Figure 5.5 presents the simplified SEM model, which shows the significant paths among variables. The results also suggest that the test content and nature, and the HSK as a

prerequisite/exit requirement had a strong impact on their strategy used, whereas the perceptions of difficulty level, test effects, and proficiency had less influence.

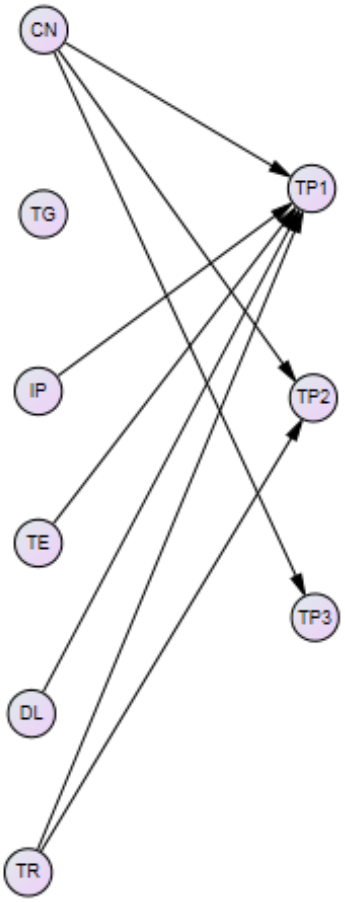


Figure 5.5. Simplified Structural equation model of washback on test-takers

5.4.2 Results from teachers' questionnaire

5.4.2.1 Descriptive Statistics

Among the 133 cases, 92.1% identified Chinese (including Cantonese, Hakka, Min) as their first language, while the remaining participants' first languages were Thai, Korean, Russian, and English. Furthermore, 78.7% (N=100) held a Master's degree or higher. Most of these teachers (N=105) indicated that their pedagogical goal was to help students enhance their communication skills (e.g., writing and speaking), and less than half of all participants indicated

that their goal was to help students 1) succeed on tests or 2) accumulate grammatical and lexical knowledge. The descriptive statistics results of the perception items are presented in Table 5.12.

Table 5.12

Descriptive Statistics of the Teachers' Perceptions on the HSK

Sub-scales	Minimum	Maximum	No. of items	Mean	SD
Test content and format	1	6	7	4.60	1.14
Test nature	1	6	5	4.67	1.05
Test goals	1	6	6	4.76	1.08
Test functions	1	6	4	4.76	1.09
Test effects	1	6	6	4.41	1.18

The results of the perceptions towards the HSK's content and format indicated that the CSL teachers showed varying levels of familiarity with the test content and tasks. For example, 19.5% teachers were not familiar with whether the tasks included content that biases against or favors test-takers. The results also demonstrated that teachers doubted the content and chose to focus more on communicative functions of the language over linguistic knowledge. In addition, they did not find the HSK's difficulty level appropriate. In terms of the HSKK, they strongly believed that if the speaking test was compulsory, they would have spent more time and efforts developing students' speaking ability. They also strongly insisted that the development of the HSK was related to the PCI policy and movement.

Regarding test use and impact, these CSL teachers were in line with HSK's values in general, with the mean of their perceptions ranging from 3.97 to 5.07. In particular, the CSL teachers were confident that the HSK provided useful feedback to students' Chinese language learning and teachers' teaching, and provided a reference for decision-making concerning student recruitment. Similar to the test-takers' perception, they also agreed that obtaining a HSK certificate would enhance their competitiveness in academic and professional development. However, they denied that the HSK forced teachers to teach to the test.

Table 5.13

Descriptive Statistics of Teachers' Teaching methods and practices

Sub-scales	Minimum	Maximum	No. of items	Mean	SD
Teaching methods	1	6	7	4.64	1.12
Teaching practices	1	6	5	4.18	0.99

The results of the CSL teachers' teaching methods and potential influence of the HSK on their practices (see Table 5.13) show that 1) they emphasized fostering students' language use ability with a combined approach of Communicative Language Teaching and the traditional structural method (e.g., grammar-translation method) in their instruction; 2) they involved HSK test questions in homework, exams, and used HSK-related textbooks; 3) they encouraged their students to participate in the HSK and to have students practice mock tests to prepare for the HSK; and 4) they did not find that their teaching methods met the students' expectations for test preparation, thus it was not an appropriate method for helping students pass the HSK.

5.4.2.2 Exploratory factor analysis (EFA)

The EFA of 37 items of teachers' perceptions of the test extracted 6 meaningful factors and accounted for 62.031% of the total variance. The same criteria were used as for the test-takers' questionnaire analysis. An examination of the Kaiser-Meyer Olkin measure suggested that the sampling was adequate (KMO= .838). The Bartlett's Test of Sphericity also reached statistical significance, which indicated that the correlations were sufficiently large for EFA (p= .000). Table 5.14 presents the loadings of the items on each factor. The 5 items loading on Factor 1 were for test effects, the 3 items loading on Factor 2 were for test goals, the 3 items loading on Factor 3 were for the nature of the test, the 3 items loading on Factor 4 and the 3 items loading on Factor 5 were both for perceptions of the HSK as a prerequisite/exit requirement, and finally, the 3 items loading on Factor 6 were for test content.

Table 5.14

Factor Loading of Teachers' Perceptions on the Test

	Factor					
	1	**2**	**3**	**4**	**5**	**6**
TE7	**.717**	.243	.083	.199	.033	.196
TE4	**.710**	.159	.232	.097	.020	.121
TE1	**.708**	.068	.215	-.009	.119	.162
TE2	**.588**	.146	-.020	.087	-.012	.099
TE3	**.517**	.093	.191	.270	.014	.009
TH1	.130	**.667**	.108	.031	-.117	.086
TH2	.150	**.649**	.104	.008	-.046	.121
TH3	.044	**.606**	.158	.300	.014	-.049
TA7	.163	-.036	**.622**	.038	.067	.338
TA9	.028	.175	**.594**	.157	.049	.226
TA10	.085	.168	**.591**	.245	.019	.094
TR1	.112	.018	.187	**.753**	.057	.154
TR3	.098	.171	.387	**.640**	-.013	.019
TR2	.125	.068	.097	**.591**	.207	.123
TR6	-.041	-.075	.150	-.122	**.798**	-.126
TR5	-.003	.040	.094	.069	**.683**	-.073
TR4	.076	-.081	-.009	.110	**.567**	.134
TA1	.381	.259	.251	.226	-.019	**.691**
TA2	.336	.280	.239	.142	.023	**.638**
TA8	.240	.062	.454	.256	.076	**.542**

EFA was conducted for the 14 items on teaching methods and practices, and 3 meaningful factors were identified after applying the same criteria detailed above, which accounted for 68.376% of the total variance. As shown in Table 5.15, the 4 items loading on Factor 1 were about teaching to the test, the 4 items loading on Factor 2 referred to teaching practices, and the 3 items loading on Factor 3 accounted for teaching methods.

Table 5.15

Factor Loading of Teachers' Teaching Practices

	Factor		
	1	2	3
TM7	**.830**	.076	.291
TM6	**.765**	.050	.160
TM8	**.742**	.100	.041
TM9	**.622**	.160	.137
TP2	.026	**.819**	.249
TP4	.463	**.697**	-.163
TP1	.034	**.693**	.241
TP3	.123	**.668**	.230
TM4	.322	.251	**.630**
TM3	.170	.025	**.542**
TM1	.062	.195	**.507**

Table 5.16 portrays the names given to the factors, the items in each factor, the number of items, and their reliability (alpha value). These factors were named based on the commonalities shared by the item loading as well as the intended sub-scales they belonged to. Compared with the intended scales, the factoring results generally corresponded well with the intended scales of the original questionnaire design.

Table 5.16

Items of 9 EFA Factors and Their Relationships to the Intended Scales

Intended scales	**Factors**	**Items**	**Item No**	**Standardized alpha**
Perceptions of test use and impact	F1: Test effects	TE7, TE4, TE1, TE2, TE3	5	.849
	F2: Test goals	TH1, TH2, TH3	3	.742
	F4: Requirements	TR1, TR3, TR2	3	.766
	F5: Requirements	TR6, TR5, TR4	3	.732
Perceptions of test content and nature	F3: Test nature	TA7, TA9, TA10	3	.720
	F6: Test content	TA1, TA2, TA8	3	.872
Classroom teaching	F1: Teach to the test	TM7, TM6, TM8, TM9	4	.853
	F2: Teaching practices	TP2, TP4, TP1, TP3	4	.816
	F3: Teaching methods	TM4, TM3, TM1	3	.768

Considering the sample size, the composite factors were used in the subsequent analysis as observable indicators in modeling the latent measurement model. The skewness, kurtosis, and histograms of all the observed variables were examined and it suggested that the normal distribution assumption had been met. Another round of data cleaning and iteration was performed on the 9 composite factors following the same procedure mentioned in Section 5.2.5 before conducting SEM.

5.4.2.3 Structural equation modeling (SEM)

Based on the EFAs' result and the hypothesized washback model on teachers in Section 5.2.1, a model of the washback effects the HSK has on teachers' teaching practices was postulated. This model involving 3 latent variables and 9 observed variables is presented in Table 5.17. Among these, the 2 latent variables (i.e., UI and CN) were composite factors of 6 perception factors generated by EFA, and they were considered as teachers' perceptions of the test (the left-hand ellipse of the Figure 5.2); the remaining latent variable (i.e., TP) was the composite factor of teaching methods and practices generated by EFA, and it was considered as teaching practices (the right-hand ellipse of the Figure 5.2). This hypothesized model was tested using SEM.

Table 5.17

Construct of Latent and Observed Variables in the Washback Model of Teachers

Latent variable	**Observed variable**
Perceptions of test use and impact (UI)	Test effects (V1)
	Test goals (V2)
	Requirements (V3)
	Requirements (V4)
Perceptions of test content and nature (CN)	Test nature (V5)
	Test content (V6)
Classroom teaching practices (TP)	Teach to the test (V7)
	Teaching practices (V8)
	Teaching methods (V9)

Figure 5.6 presents the structural equation model generated by AMOS. The inter-related constructs of perceptions of the test (i.e. CN and UI) are connected to each other using double-headed arrows. The single-headed arrows from CN and UI to TP represent predicable relationships between perceptions on the test and teaching practices. By adopting the same evaluation criteria mentioned in Section 5.3.1.3, the overall goodness of fit (see Table 5.18) suggested that it was an acceptable model for representing the interrelationships of test-takers' perceptions on the test and teaching practices.

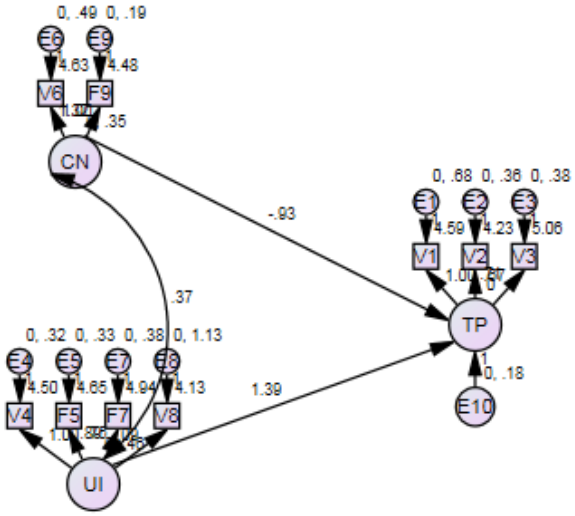


Figure 5.6. Structural equation model of washback on teachers

Table 5.18

Goodness of Fit Summary for the Hypothesized Model of Washback on teachers

Indices	**P**	**χ2**	**df**	**χ2/ df**	**IFI**	**TLI**	**CFI**	**RMSEA**
Values	.063	36.595	25	1.464	.955	.913	.951	.059

Individual parameter estimates were examined after evaluating the model's fitness. Table 5.19 provides the unstandardized and standardized parameter estimates, standardized error estimates, and the *p* values.

Table 5.19

Parameter Estimates for the Model of Washback on Teachers

			Estimate	S.E.	C.R.	P
TP	<---	CN	-.927	.981	-.945	.345
TP	<---	UI	1.387	.898	1.545	.122
V6	<---	CN	1.000			
F9	<---	CN	1.315	.214	6.136	***
V1	<---	TP	1.000			
V2	<---	TP	.713	.160	4.444	***
V3	<---	TP	.675	.155	4.356	***
V4	<---	UI	1.000			
F5	<---	UI	.892	.118	7.588	***
F7	<---	UI	.747	.117	6.365	***
V8	<---	UI	.085	.156	.548	.584

Note: *** represents $p < .001$

As shown in Table 5.19, the paths from CN to TP and UI to TP were found to be not significant. These remaining significant paths indicated that the observed variables (except for V8) were satisfactory for their loading latent variables. The correlation between CN and UI suggests that there is a strong correlation between them. Overall, the SEM results indicated that the proposed model of washback on teachers is acceptable for presenting the relationships between their perceptions on the tests and their teaching practices. The results also suggested that the test content and nature, as well as the use and impact of the HSK, had little or no influence on their teaching practices.

5.4.3 Findings of classroom observation

As explained earlier, the observation was conducted in an undergraduate level credit HSK preparation course in a Chinese International Education program. This university is well known as a key university in China and the International Chinese Education major is a top-ranking program. All of the students (n=28) from this class were international students (most of them are from Asian countries) and were in their senior year when the study was conducted.

When they entered the program, some of them had passed HSK 3 or 4, while others had passed the program's internal language test; some also did not provide any language test scores. The teacher of this class had a Master's degree in TCSL and over 5 years of teaching experience in both CSL and CFL contexts. However, this was her first time to teach a HSK preparation credit course in three parallel classes[20]. She claimed she did not know enough about the HSK before she was assigned to teach these classes. However, after becoming familiar with it, she realized that some teaching materials and workbook assignments in regular CSL classes were related to the HSK to some extent (e.g., vocabulary items, topics, question types, format). In the first two observations, most of the activities she conducted were focused on the content of the text[21] rather than the HSK, and she utilized many pair/group work activities. Only a few test-related activities were found in her class and assignments. However, the observations after the midterm exam reflected an obvious change, such that the test-related activities became more common and more apparent and the lessons became more test-oriented. For example, there was an increase in the amount of class time devoted to "teaching to the test", such as practicing the mock test, providing and discussing test-taking strategies by analyzing test questions, and replacing the textbook with test-like worksheets. When asked about her motivation, the teacher explained,

"尽管这是一门 HSK 课，可是我并没有打算教学生如何应付考试，只是按照课本像其他综合课一样上课。因为我觉得教学是让学生达到交际的目的，只通过考试不能达到

[20] There are 83 senior year international students majoring in Chinese language Studies (officially called International Chinese Education); the students were assigned to 3 classes. The observed class is one of 3 classes. The first two classroom observations were conducted before their mid-term exam, while the other four were conducted afterward. The program's mandatory class evaluation was carried out after the mid-term.

[21] The textbook is *HSK Standard Course [HSK 标准教程]*.

这样的目的。可是从期中考试之后的教学评估反馈来看， 随着考试时间的临近，学生们想更多地学习跟考试内容相关的，所以我需要对教法做出了调整。”

[Translation: Although it is a test-prep course, I didn’t plan to teach to the test. I wanted to teach the course according to the textbook, just like other regular comprehensive CSL classes. I believe the goal of my instruction is to help foster students’ communicative language ability, but passing the HSK doesn’t mean they can understand this goal. From the feedback of the teaching evaluation received after the mid-term exam, my students complained about the communicative instruction by saying that they wanted to learn more about the test since the test was right around the corner. As a result, I had to dedicate more of my class time to teaching test-related content.]

Students generally felt that the HSK had a positive effect on their learning. Below are some of the comments they made: “因为 HSK 与学位挂钩，而且通过六级考试对申请研究生奖学金有很大帮助，所以这激励我更好地学习（汉语）” [Translation: Since the HSK is linked to degrees/diplomas and passing HSK6 will help me apply for a graduate study scholarship, I am motivated to learn (Chinese)]; “通过考试可以证明我的语言水平，而且对我这一阶段的学习有一个客观的评估” [Translation: I can get an objective evaluation of my proficiency through taking the test.]; “平时我和中国人交流没有什么问题，所以不知道如何进一步提高，HSK 让我学习有了目标” [Translation: I don't have many problems when communicating with Chinese-speaking people, so I don't know how to further enhance my proficiency level, but HSK gives me a direction.] However, the students demonstrated different opinions about the test content, format, and difficulty levels. For example, some of them thought

that the writing part was quite challenging, while others believed that the HSKK was unrealistic because their speaking ability was accessed through voice recording.

5.5 Discussion

The findings of the study suggest that the HSK has achieved its intended purpose to a certain extent, particularly in terms of: 1) personal achievement (i.e., test results as proof of language proficiency and individual achievement); 2) learning-oriented characteristics (i.e., the test provides effective feedback to learners and teachers); and 3) public accountability (i.e., the test provides useful evidence of meeting prerequisite/exit requirements). This section focuses on discussing the uses and washback effects of the HSK by comparing the test-takers' and teachers' perceptions.

5.5.1 Perceptions on test content and nature and its uses

Based on the descriptive statistics from the two questionnaires, test-takers and teachers indicated that they were generally satisfied with the values of the test. However, participants were critical of the test content, format, and difficulty. For example, since the HSKK (the oral test) is an optional test separate from the HSK, only 30.4% of the test-taker participants took the HSKK. As indicated by the descriptive statistics and the EFA result on the test preparation strategies, speaking practices were neglected as compared to the other three skills. The teacher participants in this study also did not indicate putting much effort into promoting students' oral ability, which was at odds with their self-reported communicative teaching practices. Not only does this make it difficult to achieve HSK's intended consequences of fostering students' communicative proficiency, but it also creates an underrepresentation in HSK's score interpretation because of the absence of an oral test score. Therefore, it would appear necessary to encourage students to take the HSKK or to integrate the HSKK into the HSK, which would

increase the fairness of the test and promote a balanced development in learners' language proficiency. In addition, for large-scale high-stakes tests, indirect tests of writing have been widely used because of their high rater-reliability and practicality (Weigle, 2002). The writing tasks in the HSK levels 3, 4, and 5 have adopted this testing format. On the one hand, these indirect test items may measure a component of test-takers' writing ability quickly and relatively objectively. However, both the teachers and test-takers questioned its validity and argued that the writing task may not reflect one's real writing ability. This is in line with other research results (e.g., Sato & Ikeda, 2015), which showed that indirect tests are less likely to be perceived as a measure of productive performance ability and thus result in poor face validity.

In terms of HSK's difficulty, the test-takers expressed different opinions from the teachers. They thought taking the HSK encouraged their learning and the relatively low difficulty of the exam boosted their confidence after they passed a certain level. The test results could highlight their shortcomings and thus offered them future learning directions. However, the teachers believed that it was difficult to distinguish the students' proficiency levels if they relied only on their HSK levels/scores. Teachers also thought the new HSK reduced CSL/CFL standards and was too easy, rendering it unsuitable for students who wanted to pursue advanced degrees in Chinese universities. As mentioned in Chapter 4 (Study 1), although both opinions reflected HSK's development in the PCI context, it is difficult to say whether the consequences are positive or negative.

5.5.2 Washback effects on teaching and learning

Based on the findings of the test-takers and teachers' questionnaires, the HSK's washback on learning was generally consistent with the intended test consequences claimed by the test developer - "以考促学" [Translation: informing learning through testing]. On the other

hand, the other intended consequences - “考教结合, 以考促教” [integrating teaching and testing, and informing teaching through testing], which would be HSK’s washback effects on teachers, were not adequately achieved.

A comparison of the two SEM models (see Figure 5.4 and Figure 5.5) illustrates how test-takers’ perceptions of the test had a direct, positive impact on the way they learn. Their views on the test content and nature were significantly intercorrelated with all the other aspects of their perceptions of the test (i.e., CN, TG, IP, DL, TR), which meant that it was a strong variable that influenced their general test-taking preparation strategies. Subsequently, test-takers’ views play a more critical role in their learning practices than anticipated, particularly in how they perceive the uses and effects of the test and how they structure test-taking practices. The findings of their test-taking practices illustrate that not only were test-oriented skills/strategies adopted, but they also acquired strategies that were beneficial for language learning in general and can help them maintain a healthy interest towards language acquisition (Xie, 2010). Another interesting finding was that the motivational affective factors (e.g., test goals and test expectations) did not interact with strategic behaviors, which is inconsistent with the previous SEM studies on test uses (e.g., Sun, 2016; Xie & Andrew, 2012). This may be due to the HSK’s context and how it is a much different exam than the CET. More specifically, the CET test examines the English proficiency of Chinese undergraduate students and aims to ensure that the students have the necessary English ability as specified in the National College English Teaching Syllabus.

In terms of the teachers’ model, results indicate no direct causal relationship between the perceptions of the test and teaching practices. The classroom observation results also demonstrated that even in test preparation courses, the classroom was affected by the test to a

limited degree (e.g., the observed teacher only changed her pedagogical practices due to the pressure from outside). That is to say, indirectly, the HSK seemed to have influenced teaching. Although "there is a general consensus that high-stakes tests produce strong washback" (Qi, 2005, p.3), "it can be seen that the general pattern of teaching approaches had not changed much" (Cheng, 1999, p.268) which also can be found in other empirical studies (e.g., Turner, 2009; Wang, J., 2010). However, according to Andrews et al. (2002), various factors related to the exam itself can influence the degree and type of washback, including when the exam was introduced and how familiar it was to teachers. This could also explain why regular teaching was interrupted when the HSK test date was approaching. In addition, the teacher was teaching this course for the first time. The limited experience of teaching a test preparation course could also affect the way she planned and taught the course. Moreover, this example illustrates how students' attitudes can also play an important role in affecting how teachers teach.

5.6 Conclusion

The findings of Study 2 provided evidence that the intended goals[22] were somewhat observed in terms of promoting CSL/CFL learning, but its ability to inform teaching was limited. From both the test-takers' and teachers' perspectives, the HSK was limited in terms of the task type and language use, and therefore, might induce negative washback on teaching and learning. In addition, this study's findings demonstrated the complexity of the HSK's washback effects on teaching and learning. It supported Xie's (2010) research in that the relationship between washback and test validity was not straightforward, such that tests do not linearly affect teaching, then learning; instead tests can affect various mediating factors. Positive washback might or

[22] As summarized in Chapter 4, the test developers intended goals for the HSK included: 1) making the exam into a reference for educational and social decisions based on individuals' Chinese language proficiency; and 2) promoting CSL/CFL teaching and learning.

might not be produced by valid test design and appropriate test use; similarly, bad test design and test (mis)uses might also cause negative washback. Due to the scale and statue of the HSK (e.g., its powerful role in personal achievement and public accountability, as a prerequisite/exit requirement of a degree program), it is necessary to examine and understand HSK's consequential validity in both the classroom as well as within the educational and socio-political context of the PCI, which will be the focus of Study 3 in Chapter 6 (as shown in Figure 3.1).

Chapter 6 Study 3 (Exploring the educational and social consequences of the HSK from the prospective of test users – A mixed methods study)

6.1 Introduction

As will be recalled from the visual diagram of the current MMR study (see Figure 3.1), Study 3 was the third phase of this research. Using the argumentative conceptual framework for the HSK established in Section 2.8 and the instruments designed from Study 1, Study 3 explored the educational and social consequences of the HSK from the perspective of the test users.

6.2 Background

Some researchers have argued that tests are rarely able to fulfill multiple purposes (Perie, Marion, & Gong, 2007). However, in the real world, an educational assessment/test seldom serves a single objective. The HSK, for instance, is intended to address three needs, namely that it needs to be "1) a reference for an educational institution's decision-making concerning recruiting students, assigning students to different classes, and granting students their academic credits; 2) a reference for employers' decision-making concerning the recruitment, training, and promotion of test takers; and 3) a method for Chinese language learners to assess and improve their proficiency in Chinese" (Hanban, 2014, p. 2).

In order to justify the assessments' intended purposes, there has been growing interest in studying the impact/washback of large scale tests through argument-based approaches in the LT field at both the classroom and societal levels (e.g., Cheng & Sun, 2015; Llosa, 2008; Sun, 2016; Wang et al., 2012; Xi, 2010). Existing empirical studies on washback also revealed that impact/washback is not a monolithic process, and is instead a highly complex phenomenon involving multiple stakeholders and other mediating factors (e.g., contextual factors). Moreover, research that links test validity with the social impact of test use has important implications in

China considering how highly exams are valued and considering the Promoting Chinese Internationally (PCI) policy. Thus, to understand the effects outside the classroom, the link between test validity and the consequences of test use must be established involving the perspectives of other test users. Accordingly, by employing Bachman and Palmer's (2010) AUA framework (the utilization argument, in particular), this study, which is the third phase of the larger MMR study, seeks to identify the HSK score users'[23] perspectives on the test as well as the test consequences and uses (e.g., decisions made based on HSK levels/scores, level/score interpretation).

6.2 Methodology

The research questions of Study 3 are:

In both academic settings (i.e., higher education in China) and non-academic settings (e.g., workplace),

RQ 1: To what extent is the test users' decision-making influenced by the HSK level/score? What specific decisions need to be made to promote the intended consequences of the HSK?

RQ 2: Who will be responsible for making the decisions? What are their viewpoints on the HSK? How do the HSK score users interpret the decisions and react to the consequences of HSK use?

6.2.1 Design of the study

The study adopted a sequential explanatory mixed methods design (Creswell & Plano Clark, 2011). Figure 6.1 presents the research design of Study 3. In Phase 1, the researcher

[23] According to Bachman and Palmer (2010), test users are "those making decisions based on the assessment" (p.135); in the current study, these are administrative staff who use HSK as a criterion to make admission/employment decisions, in both academic and non-academic settings.

initially collected and analyzed the data from two exploratory questionnaires (one for institutional test users and one for organization/company test users), which were designed based on the AUA framework and were also informed by Study 1's findings. In the second phase (semi-structured interviews), qualitative data were collected through interviews to help explain and build on the statistical results by exploring participants' views in more depth (Creswell, 2015; Creswell & Plano Clark, 2011; Teddlie & Tashakkori, 2009).

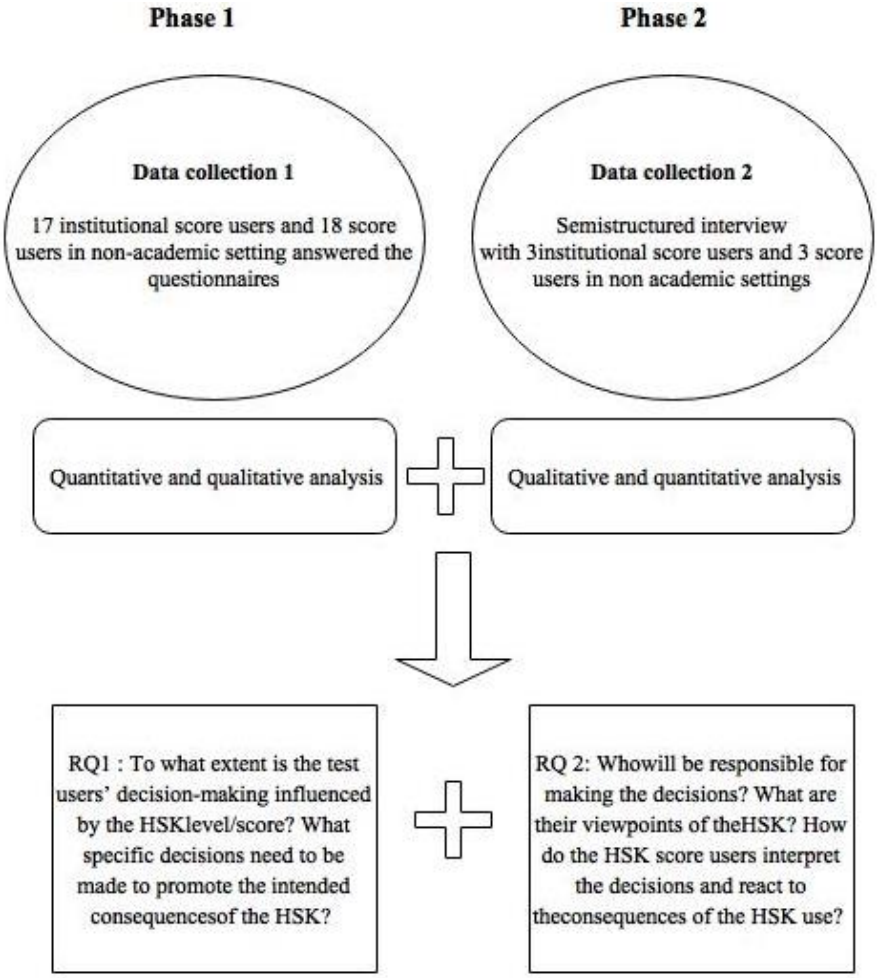


Figure 6.1. Research design of Study 3

6.2.1.1 Participants

This study used a snowball sampling method, which is a network-based technique and is convenient for selecting samples and locating target survey respondents (Noy, 2008). First of all,

a recruitment announcement e-mail was sent through the researcher's personal contacts, who then forwarded the message to the score users in both academic and non-academic settings, such as program coordinators, university officers of academic affairs involved in making international students admission and graduation decisions in Chinese universities (including all six participating universities in Study 2, potential employers of university graduates with HSK certificates, and employers of non-Chinese employees who use Chinese in their work inside and outside China[24]. These contacts referred those they knew and these individuals in turn referred other relevant personnel they knew. In March 2017, this inquiry resulted in a total list of 37 people identified as being directly involved in admission/employment policy decisions and in the use of HSK and/or other Chinese proficiency test scores. All of these individuals were asked to give consent to take part in the survey research. In the end, 35 of them completed the questionnaires in Phase 1, including 17 institutional score users and 18 score users in non-academic settings. The participants for Phase 2 (the qualitative phase) were selected from the questionnaire respondents, who indicated on their questionnaire that they were interested in being interviewed and had representative responses. Interviews with 3 institutional score users and 3 score users in non-academic settings were subsequently conducted. The recruitment of the interviewees was based on the individuals' questionnaire results, which indicated that they were the most appropriate candidates for explaining the issues of interest identified in the quantitative phase (Creswell & Plano Clark, 2011). In addition, to obtain a diverse set of representatives, candidates from different backgrounds (e.g., CSL, CFL, nationalities, nature of the

[24] The HSK has become an important criterion for employment, pay increase and promotion in more and more governments and multinationals. For example, it has become the new standard for Thailand's immigration police officer recruits (Li, May 26, 2015). As well, as a global conglomerate, Hyundai Motor Group encourages its staff to learn Chinese, stating that a good HSK score is an advantage for employee promotion.

organization/company) were considered and given priority. Their information is described in Table 6.1.

Table 6.1

Profile of the Interviewees

Name	Nationality	Affiliation	Location	Position
AS1	Chinese	A key university[25] in China	Beijing, China	Administration officer
AS2	Chinese	A non-key university in China	Nanjing, China	Director, School of International Culture Exchange
AS3	Chinese	A key university in China	Guangzhou, China	Administration staff
NAS1	Chinese	A multinational enterprise	Shanghai, China	Human resource manager
NAS2	Korean	A multinational enterprise	Guangzhou, China	General manager
NAS3	Thai	A non-profit organization	Bangkok, Thailand	Program coordinator

6.2.1.2 Instruments

In Phase 1, the questionnaire for academic settings (AS) (refer to Appendix 5 and 6 for the survey instruments) included two parts: 1) Personal and institutional information (including biographical questions, multiple choice questions, and short answer questions); and 2) perception questions relating to the HSK design, score interpretations, uses and consequences (including 6-point Likert-scale items). The following table (Table 6.2) illustrates the questionnaire's construct, the links to the research questions, and the 3 scales for decisions, decision makers, and

[25] National Key Universities (Chinese: 国家重点大学) refers to prestigious universities that receive a high level of support from the Chinese government.

test interpretation. Items that elicit participants' opinions regarding the use of the HSK as a graduation requirement are the same items used for the test-takers' and teachers' questionnaire seen in Study 2.

Table 6.2

The Construct of Questionnaire for Academic Setting

RQs	Scales	Item numbers in the questionnaire	Items
1	Decisions	5, 6, 8, Q2 (1-6)	9
2	Decision makers	1, 2, 3, 4, 7, 9, 10	7
2	Interpretation	Q1 (1-9), Q2 (1-6)	15

The questionnaire for non-academic setting (NAS) (see APPENDIX 6) was developed based on Study 1's findings as well as on Pan and Roever (2016). The constructs were similar to the questionnaire used for AS, which also utilized two sections: 1) Personal and organization/company information (Question 1, 2, and 3); and 2) perception questions of the test use in their recruitment/promoting process (Question 4, 5, 6, 7, 8, 9, and 10).

The survey questions in both questionnaires were employed to gain information about the consequences of using the HSK for different purposes (e.g., admission in higher education, employment), and to provide evidence needed to inform test design, use, admission/employment decision-making, and score interpretation.

6.2.3 Data collection procedure

To ensure the validity and reliability of the questionnaire measurement, the following procedures were followed before the data collection: For AS setting questionnaire, three admission officers in Chinese universities who were responsible for international students' admission were invited as a focus group to discuss their beliefs about the HSK and other

language tests used in the contexts of their work and what the tests were measuring, as well as their experience in using test information for decision making. The questionnaire was then pilot tested by this group. They read and responded to the questions, and then commented on their clarity and practicality. Revisions were made based on their feedback; for example, in order to ensure the response rate, some open-ended questions were changed to multiple-choice questions (e.g., Question 8). Regarding the questionnaire used in NAS, a procedure similar to the AS one was used. The focus group discussion with 2 managers of multinational companies confirmed the findings from Study 1, namely that employment and promotion decisions were not noticeably affected by individuals' HSK levels/scores and that employers do not have sufficient knowledge about the HSK. Therefore, the researcher decided to change those questions into more practical and suitable open-ended questions related to the language requirements in the employment/promotion process (i.e., not only on the HSK, but also related to other Chinese proficiency test certificates' use). In the further iterative pilot process, minor changes were made based on the suggestions of participants from a CFL context. The finalized questionnaires were then sent to the participants.

Afterwards, open in-depth individual interviews were undertaken with the participants' consent. This gave them an opportunity to give their opinion and the story behind their experiences. The general interview guide was informed by the questionnaire findings, focusing on the important results, themes, and issues that emerged from the quantitative analysis. Each of the 6 interviews lasted between 30 and 45 minutes and was audio-recorded and then transcribed to facilitate data analysis. All of them were conducted by the researcher in Chinese and/or English, depending on each participant's preference. The purpose of this phase was that qualitative interviews are presumed to be an appropriate and effective approach to inquire about

specific social processes or particular individuals' perspectives through direct contact with those involved in natural contexts (Locke, Spirduso, & Silverman, 2000); all of this can help contribute to a more comprehensive and nuanced understanding of the consequences of HSK score use.

6.2.4 Data Analysis

To answer the research questions, a quantitative analysis of the multiple-choice questions and scales questions in the questionnaires was performed by SPSS 22.0. Descriptive statistics (e.g., frequency counts of responses to each question) were then generated. For the open-ended questions and interview data, both quantitative and qualitative analyses were carried out, making use of grounded theory techniques (Strauss & Corbin, 1998) and the aforementioned descriptive statistics analysis. Open coding was conducted over multiple iterative readings, with a focus on identifying general categories, namely decision-making, test use, and selection criteria(s). Following this, a second level of coding for all the data was conducted using an iterative process. Finally, different data sources were synthesized and integrated.

6.3 Findings

In this section, the findings for each research question are described in AS and NAS respectively, corresponding to the test developers' intended test functions.

6.3.1 RQ 1: In academic settings (AS)

The findings indicated that the 17 survey respondents were from various types of institutions (for example, located in first tier and second tier cities[26], key universities, and non-key universities). The answers showed that HSK was the most widely used method to verify

[26] First-tier cities (Chinese:一线城市) refers to the major cities that play an important role in a country's economic and political spheres. These cities are generally ahead of others in terms of infrastructure, revenue, consumption, attractiveness to talents, etc. Those widely recognized as the first-tier cities in China are Beijing, Shanghai, Guangzhou, and Shenzhen. Accordingly, the second tier (Chinese: 二线城市) generally refers to other major and less famous cities.

Chinese proficiency in the admission procedure. It was recognized by all of the surveyed institutions for undergraduate or graduate international students (degree program) admission purposes. Besides the HSK test, some of the other language tests were also accepted, such as the institutional placement test. According to the findings from their admission websites, most of these institutions required HSK 4 or 5 (with a cut-off score of 180, overall score 300, and the minimum score of 60 in each subset); only one institution listed their requirement as HSK 3, and one institution required HSK 6. Some institutions had different requirements for applicants from different departments. For instance, arts/ humanities departments usually have higher requirements than in science/ technology departments. However, none of them mentioned HSKK, the oral test, in their recruiting policy. When formulating these requirements, both AS1 and AS2 noted that administrative staffs (and even some academic staff) were not very familiar with the test and the meaning of the levels/scores; they often refer to the MOE's documents and the cut-off scores of other universities, as well as the actual situation of their institutions.

In terms of other uses of the HSK during the admission process, international students who also applied for Confucius Institute Scholarships [27] and Chinese Government Scholarships needed to provide qualified HSK and HSKK levels and scores. Generally, these scholarships' requirements were higher than those for university entrance; they also needed a higher level and score in order to be more competitive and to receive the scholarships. However, there were

[27] Confucius Institute Scholarship: In order to support development of Confucius Institutes, facilitate Chinese language promotion and Chinese cultural transmission in the world, cultivate qualified Chinese language teachers and talented students of Chinese language, Confucius Institute Headquarters (Hanban) launched the "Confucius Institute Scholarships" program for providing sponsorship to students, scholars and Chinese language teachers of other countries for pursuing study in relevant universities in China.

exceptions; AS3 noted that "sometimes it is not easier for students with a HSK 6 to enroll than those who have HSK 5". She elaborated on this by stating that

> "HSK6 级很难，至少比 5 级难得多。6 级高分的学生通常是华裔，母语是汉语或者长到十几岁才出国的这种。这些人跟真正的留学生比拼的话不公平，奖学金是给那些人准备的，他们更需要这种机会。"
>
> [Translation: HSK 6 is very difficult, much more difficult than HSK 5. An applicant with a high HSK 6 score usually has a strong Chinese background, such as Chinese as their mother tongue, or perhaps they were Chinese but had moved abroad as teenagers. This means that these individuals should not compete with other international applicants with lower levels and scores. We created these scholarships in order to encourage international students to apply. They need more chance to learn and polish their Chinese in our programs.]

Besides the uses of the HSK in admission procedures, a few institutions mentioned that passing specific HSK levels (usually HSK 5 or 6, which is higher than general admission requirement) is an additional prerequisite or graduation requirement for specific majors (e.g., International Chinese Education).

Due to the different administrative setups in the institutions surveyed, the responses showed that different administrative staff and faculty members are involved in making admission decisions. This can be loosely divided into four types: 1) Dean/director of the relevant departments, for example the Institute for International students[28], the Admission Office for

[28] The Institute for International Students is both an administrative institution responsible for international students' enrollment at the university and an academic institution that conducts teaching and researches mainly in Chinese language and culture. There are different names of this kind of departments/institutions, such as the School of Teaching Chinese as a Second Language, Institute of International Cultural Exchange, International Education Exchange Center, and so on.

International Students, the Foreign Affairs Office; 2) administrative staff; 3) CSL program coordinators and teachers; 4) others (e.g., in terms of graduate-level admission, a graduate students' academic committee in some institutions is also involved in international students' admission).

Table 6.3 shows how strictly decision-makers apply the cut-off scores (including levels and scores) for the HSK and/or other language tests (see Appendix 5, item 8) when they admit a student.

Table 6.3

Application of the HSK Cut-off Scores in Admission

Items	Number	Percentage
The cut-off scores are not applied. We make the admissions decisions on other criteria.	0	0%
The cut-off scores are not always applied strictly because other admissions criteria are sometimes judged to be more important than scores on the language tests.	1	5.88%
We usually respect cut-off scores, but we make occasional exceptions when the rest of the student's application is very strong.	13	76.47%
We always apply cut-off scores. We never accept students into our programs if their language test scores are below the cut-off.	3	17.65%
I am not sure what happens.	0	0%

The table shows that cut-off scores were applied to differing extents among the institutions surveyed. However, the majority (76.47%) of them said they usually respected cut-off scores/levels, but made occasional exceptions for some cases. The interviewees with the test users in AS also confirmed this finding, as both AS1 and AS3 mentioned that they usually applied the cut-off scores strictly because 1) it could help ensure the students' language ability during their studies, and 2) it could help them make the admission selection more efficient.

However, it did not mean they did not make occasional exceptions. For example, although some graduate students majoring in science and technology do not have any language test scores, their supervisors can have the final say in the admission process and still approve these students' applications. These students can even write their thesis in English, even though the program itself is in Chinese.

Regarding the test users' beliefs toward the HSK test (see Table 6.4), the findings revealed that in the PCI context, the HSK was a relatively trustworthy test and was widely recognized by universities. The test users generally agreed with the statement that an applicant with a high level/score HSK certificate would have more opportunities for scholarships (m=4.13). However, they doubted the test's fairness (m=3.44) and accuracy in reflecting one's proficiency (m=2.88), and believed that the score-based interpretation did not provide sufficient information for them to make decisions (m=3.50). During the interview with AS2, he indicated that his institution had not undertaken formal tracking of international students' academic performance (e.g., GPA) in their program, but he believed that HSK 4 (their official cut-off score) was not adequate to complete their academic programs (degree education). It appeared easier for students with higher HSK levels/scores to complete their program. The inference was also supported by the responses to Item 10 (m= 4.00).

Table 6.4

The Test Users' Beliefs towards the HSK Test

Item	Minimum	Maximum	Mean	SD
1. The HSK provides an accurate measure of test-takers' overall Chinese proficiency.	1	6	3.94	1.24
2. The HSK is fair for all test-takers during the whole procedure of the test.	1	6	3.44	1.63
3. The score-based interpretation provides relevant and sufficient information to make decisions.	1	6	3.50	1.46
4. The development of the new HSK is related to the PCI movement.	1	6	4.06	1.53
5. The HSK is trustworthy in terms of its validity and reliability.	1	5	4.06	1.12
6. The HSK is widely recognized by universities, companies, and organizations in the recruitment process.	1	6	4.13	1.59
7. An applicant with a high level/score HSK certificate would use Chinese more proficiently than those who have a low level.	1	5	2.88	1.15
8. An applicant with a HSK certificate would use Chinese more proficiently than those who do not have a HSK certificate.	1	6	3.75	1.34
9. An applicant with a high level/score HSK certificate would have more opportunities to apply for scholarships at your institution.	1	6	4.13	1.41
10. An applicant with a high level/score HSK certificate would have more opportunities to achieve a high GPA at your institution.	2	5	4.00	.89

Regarding the washback effect of the HSK when it is a prerequisite (or an exit requirement) in their programs, the test users thought the requirement could encourage students to put more effort into learning Chinese and thus enhance their proficiency (m=4.13). However,

such a requirement did not cause learners to set ‘passing the HSK’ as their main goal for learning Chinese (m=3.56), and it also did not force the students’ to place extra work or pressure to pass the test. This was confirmed by one of the representative interviewees – AS3, who stated that:

> Our (HSK) level/score is easy to acquire, so it won’t push the students to work hard on it. But they should know passing HSK 4 is just a start, they need to work harder on learning Chinese to benefit from their future study in the university. As far as I know, most of these students continue to take the higher levels of the HSK after entering our program… Other application documents show that a high learning ability is important.

In terms of the effects on teachers, they also believed that this kind of admission policy did not affect teachers’ instructions (m=3.38). As pointed out by AS1, she did not find that the teachers in her institution taught to the test unless requested by the students, as their program does not have any teaching evaluation criteria related to the students’ HSK score.

6.3.2 RQ 2: In non-academic settings (NAS)

According to media reports, an increasing number of foreign governments and multinational companies have recognized HSK, HSKK, BCT, and other Chinese tests. They have subsequently used them as a criterion for recruitment, performance evaluation, and promotion. The findings of this study also indicated that possessing an acceptable level of Chinese proficiency was an undeniable advantage for people when seeking employment or promotion in many fields.

Among the 18 survey respondents, executive decision-making level accounted for 16.67%, middle management level took up 44.44%, operational level occupied 27.78%, and the remaining 11.12% of the respondents were from human resources departments. Their

companies/organizations were from various industries (see details in Figure 6.2). Around a third of the respondents were from manufacturing industries, a third from education, culture, art, cinema, and television industries, and another third from other industries. Among all these companies/organizations, 27.78% of them had been in operation for less than 10 years, 27.78% for 11-50 years, and 44.44% had over 50 years of history.

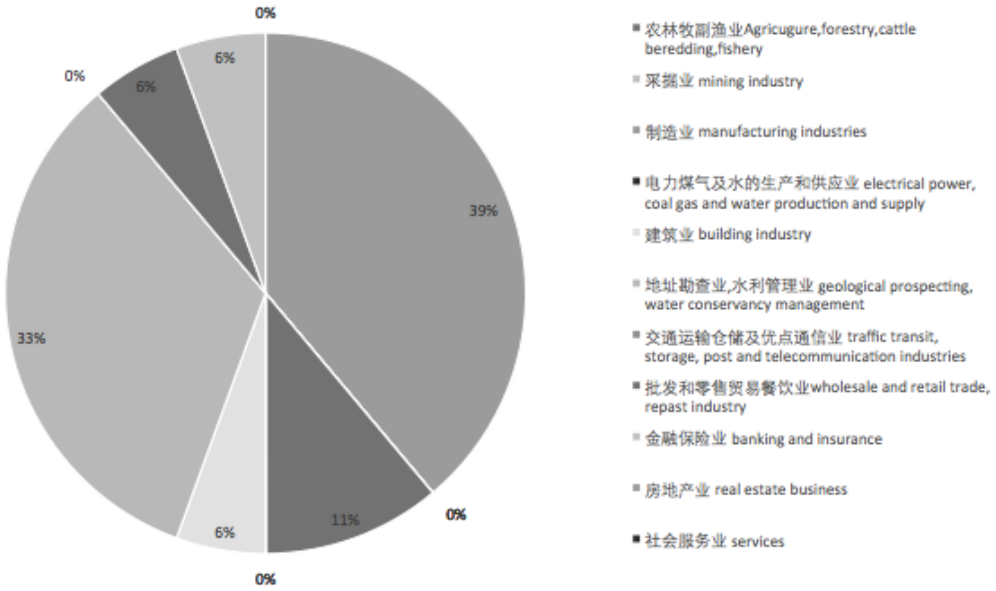


Figure 6.2. Attribution of industries

Regarding Chinese test certificates in the Chinese-related job recruitment/promotion process in these organizations, many of them required the applicants to have Chinese proficiency certificates; however, over 75% of them said the certificates are not necessary. When asked about their familiarity with the Chinese proficiency tests, the respondents reported that HSK accounted for 83.33% and outclassed other tests, such as the BCT (16.67%) and the TOCFL (5.56%). Only 4 of the 18 respondents noted that they required candidates to have a specific HSK level to indicate potential employees' working knowledge of Chinese. Most of the survey respondents (88.89%) said they assessed the Chinese proficiency of new employees (or potential employees) in person or through their internal language tests (e.g., translation tasks), as a part of

the interview process. Furthermore, 72.22% of them recommended their employees or prospective employees to take the HSK test. AS3, an interviewee from Thailand, said the following:

> In recent years (in our country), there are many Chinese companies or local companies, such as Bangkok Bank, who need to hire Chinese-speaking employees. However, the HSK test and other tests (e.g., BCT) are not generally required in these companies' (recruiting process). Only some Taiwanese companies will ask for TOCFL certificates (as a criterion). Very differently, almost all the Japanese companies ask their employees or potential employees to pass the Japanese Language Proficiency Test (JLPT) … Although obtaining a (Chinese proficiency) certificate is good, we put more weight on one's communicative ability, especially in our professional expertise domain. Another reason is that these employees or potential employees include a large number of overseas Chinese or ethnic Chinese. They don't need to demonstrate their Chinese proficiency because they are native speakers of Chinese.

Regarding the important considerations of recruiting or promoting employees, the analysis of the qualitative data indicated that six codes were found through frequency counts (see Table 6.5). It showed that Chinese proficiency was in the third place, after professional working competence and comprehensive competence. Thus, although the work required one to be proficient in Chinese, it was not the most vital concern. In order to make a more dependable recruiting/promoting decision, language requirements, working competence, and other individual factors were evaluated holistically.

Table 6.5

Codes of Important Criteria in Making Recruiting/Promoting Decision

Codes	Frequency
Professional working competence	13
Comprehensive competence	11
Chinese proficiency	10
Personality	8
Previous working experiences	7
Other aspects	3

NAS2 pointed out the different criteria requirements for various positions when hiring/promoting:

> When we select the Chinese area manager, the headquarter considers more on one's comprehensive ability and their contribution (and/or potential contribution) to the company based on our overall planning. (Chinese) language proficiency is not necessary, because it can be improved after training. Language is just a tool; as long as they can understand each other, it's good enough. More importantly, understanding Chinese culture, especially the business culture, is even more vital. However, when recruiting new employees whose positions need to use Chinese, we assess their language ability. If they were awarded a Chinese test certificate, they will have an advantage over those who don't have one.

Among the 4 companies/organizations in this sample (3 are Korean) using the HSK as a language proficiency requirement in employment/promotion procedures, all of them pointed out that they thought the HSK is trustworthy in terms of its validity and reliability, thus the HSK could provide an accurate measure of one's overall Chinese proficiency. However, some of them

did not have minimum HSK levels/scores. When they made their language requirement policy, they wanted to make sure the employees (or potential employees) have 1) sufficient Chinese proficiency to complete the job, 2) a strong learning ability, and 3) good communicative skills. Similar to the situation in the AS, none of the 18 survey respondents mentioned the HSKK test in their language requirement. When the researcher asked why the HSKK was not included, some of them expressed that they did not know the oral test was a separate test, as they thought the HSK tested all four skills (listening, speaking, reading, and writing). However, they all recognized that Chinese oral communication capability was very important.

Although most of the companies/organizations surveyed did not list having a Chinese proficiency certificate(s) as a recruiting/promoting criterion, they still thought an applicant/employee holding a Chinese proficiency certificate might have better language learning ability and might be more capable of working in Chinese. However, they did not feel that they were more proficient than those who do not have these certificates. As NAS1 mentioned, holding a certificate meant applicants had the ability to pass the test, but their actual proficiency might not be better than those who did not have one: "In general, I thought those who have certificates have more standardized language, and it was more concrete than just saying they could speak Chinese." His opinion was representative of most participating companies/organizations' positive attitude towards the language proficiency certificates. In addition, 72.22% of them noted that employees whose Chinese was better and who had been awarded Chinese test certificates would be given more opportunities for promotion.

Over half of the survey participants indicated that their companies/organizations provided incentives (benefits) to learn the Chinese language; some even gave opportunities to learn Chinese at their companies/organizations. For example, NAS2's company provided free

language courses to middle (and/or above) level employees three times a week in the early morning (before work starts). NAS1's company had a policy where they reimburse up to 2,000 RMB annually for employees to take Chinese language courses. These policies had highly increased foreign employees' motivation to study Chinese.

6.4 Discussion and implications

6.4.1 In academic setting

Although the Chinese MOE required foreign students who enroll in Chinese degree program at universities to obtain the HSK level 4 or above, different institutions have adjusted this policy according to their actual situation. The HSK certificate has undoubtedly become international students' gateway to studying in China and has become a priority for scholarship applications. The use of the HSK has thus been beneficial to these programs.

The following table (Table 6.6) shows the number of words mastered, the equivalence of the International Chinese Proficiency Standard (ICPS) Level and the Common European Framework (CEF), and the description of each level's proficiency for HSK levels 3-6.

Table 6.6

The Descriptions of the HSK 3-6

Level	Vocabulary	ICPS	CEF	Description
3	600	III	B1	Test takers who are able to pass the HSK 3 can communicate in Chinese at a basic level in their daily, academic and professional lives. They can manage most communication in Chinese when travelling in China.
4	1200	IV	B2	Test takers who are able to pass the HSK 4 can converse in Chinese on a wide range of topics and are able to communicate fluently with native Chinese speakers.
5	2500	V	C1	Test takers who are able to pass the HSK 5 can read Chinese newspapers and magazines, enjoy Chinese films and plays, and give a full-length speech in Chinese.
6	5000	V	C2	Test takers who are able to pass the HSK 6 can easily comprehend written and spoken information in Chinese and can effectively express themselves in Chinese, both orally and on paper.

Note: adapted from the HSK specification document (http://www.chinesetest.cn/userfiles/file/dagang/HSK3.pdf) p. 2.

As indicated above, although the HSK provides test users with a general idea of candidates' Chinese proficiency at each level, it does not suggest any recommendations about which level students in higher education (degree education in particular) should obtain to ensure they have adequate proficiency for pursuing advanced degrees. Take the HSK 4 for example, which is requested by most of the universities: the vocabulary of 1200 words is only one third of a native elementary school speaker's literacy level[29]. In terms of writing, there are two task types in HSK 4's writing section: 1) 完成句子 (rearrange the order of given words to make a meaningful sentence); and 2) 看图用词造句 (make a sentence based on the given picture and

[29] According to *the nine-year compulsory education fulltime primary school Chinese teaching syllabus* (Chinese:《九年义务教育全日制小学语文教学大纲》), primary school literacy is 2400 words, secondary school (junior high school) literacy is 3800 words, and the complete secondary school literacy is 6600 words.

word). A review of sample tasks in test papers revealed that most of these contents were not related to academic writing and were too easy when compared with other commonly used foreign language proficiency tests for university admission purposes (e.g., TOEFL and IELTS Academic). Consequently, the adequacy of the HSK 4 cut-off score was not sufficient. It confirmed the findings from this researcher's previous study (Wang, 2013), such that even when students had a qualified level/score, they might still feel that they were not proficient enough for university-level work, which might result in poor understanding of lectures, heavy work load when finishing assignments, and low academic achievement. A comparison of AS learners' (test-takers), teachers', and the score users' beliefs towards using the HSK as a prerequisite or graduation requirement (see Table 6.7) showed that score users held significantly different opinions than the learners and teachers. For example, learners and teachers thought the policy could ensure learners' proficiency level and could motivate learners; however, the score users did not believe so. The inappropriate cut-off scores could be a key reason for this difference. As indicated in the above results section, they thought that the cut-off score was relatively low, thus applicants could pass it easily; the low difficulty level of the test could not demonstrate whether candidates had adequate proficiency for completing their university study. That is to say, decisions made using the HSK scores did not take into account students' actual proficiency and its impact on their academic future in the university.

Table 6.7

A Comparison of Test-takers', Teachers', and the Score Users' Beliefs

Items	Test-takers		Teachers		Score users in AS	
	m	**sd**	**m**	**sd**	**m**	**sd**
R1. Can enhance learners' Chinese proficiency.	4.61	.98	4.94	.95	4.13	1.50
R2. Can increase the amount of energy/money learners allocated to Chinese learning.	4.42	1.07	4.85	.98	4.06	1.44
R3. Can ensure the proficiency level of learners.	4.58	.95	4.80	1.01	4.00	1.41
R4. Can place motivate learners to work harder and pass the test.	4.40	1.05	4.31	1.30	3.38	1.15
R5. Can make 'passing the test' the main learning objective for students	4.39	1.03	4.08	1.27	3.56	1.09
R6. Can encourage learners to learn test-taking strategies but not really learn how to communicate in Chinese.	4.40	1.02	4.13	1.36	3.81	.98

However, from another perspective, the relatively low requirement could attract more international applicants, which was in line with the PCI policy. The increasing number of HSK test-takers and the large number of applicants to Chinese universities reflected the positive impact of the PCI policy, which also prompted the development of the new HSK in 2010 (as discussed in Chapter 1 and 3). This brought new considerations to the admission staff, such that more attention should be paid to other admission criteria besides providing evidence of language proficiency. For example, the writing component of the HSK writing section might not accurately reflect candidate's capabilities. Therefore, other documents such as writing samples, cover letters, and personal statements could be integrated and used to examine their writing ability. As Chalhoub-Deville and Turner (2000) suggested, considering "how language ability, individual factors, and academic requirements fit together [can] ensure more dependable admission decisions" (pp. 537–538). Meanwhile, the admission staff and decision makers needed increased language assessment literacy so they could more efficiently and ethically

interpret HSK scores, such as becoming more aware of the meaning and interpretation of the HSK score profiles through training, communicating with academic staff, and following up with incoming students to learn whether they have any issues due to their insufficient language proficiency. This finding has been supported in other studies exploring test users' literacy of language proficiency tests, such as Baker, Tsushima, and Wang (2014), Banerjee (2003), and O'Loughlin (2011, 2013). Furthermore, the survey respondents demonstrated awareness that academic language proficiency was a constantly evolving construct. In order to ensure that students' study at the university goes smoothly and students' proficiency continues to increase after admission, the HSK should not only be used as a threshold for entering the university, but should be used to guide their future learning. As the HSK's philosophy states "以考促学" [*improve learning through testing*], a learning-oriented test can not only tell you where you are, but can also point out where you should go and how to get there. With this in mind, institutions could provide courses such as academic writing and communication skills to assist students' learning. From another aspect, they could even raise the overall cut-off level/score, increase the minimum sub-score of each section (i.e., listening, reading, writing), or add a new policy where students must pass a certain level/score HSK (higher the admission requirement) to meet the graduate requirement. These policies could not only motivate students to become more proficient, but also increase their competitiveness in the job market.

6.4.2 In non-academic settings

As the findings indicated, the interviewed companies/organizations that required HSK certificates as an employment prerequisite or promoting requirement were in the minority (only 4 out of 18). Although the use of the HSK did not significantly affect the selection procedures, they noted that holding a HSK certificate (and/or other Chinese proficiency certificate) was an

advantage. The HSK certificate, just like other certificates commonly required during employment (e.g., College English Test, National Computer Rank Examination) was beneficial, but not essential, for every organization.

The HSK's relatively weak consequence on the workplace was mainly due to the fact that candidates' underlying language abilities were not the primary consideration for employment/promotion; instead, their specialization, professional skills, and personalities were considered more valuable to employers. In addition, there was a discrepancy between employers' perceptions of the HSK (or other Chinese certificates) and employees' actual Chinese ability. For example, from the perspective of some business representatives in this study, those with a good level/score could not necessarily use Chinese in real-life situations fluently and professionally; this was highlighted when NAS3 said that candidates with high scores might "just have high test-taking skills." In addition, the HSK's weak consequence was also because of the nature and the content of the HSK. In McAloon's (2008) Ph.D. dissertation investigating advanced language use in Chinese-related careers, he argued that the HSK was a test of general linguistics skills and in relatively culture-free environments; as a standardized test, it measured test-takers' ability to take the test but could not accurately attest learners' ability to perform in Chinese at a professional level. The present study also supported this finding, as employers indicated that they wanted real language ability that allows candidates to function at work. Assessments of language abilities, especially at the advanced levels, should also reflect employers' needs. As such, some of the interviewees stated that although they asked for the HSK certificate as a prerequisite, they still had to use their internal language test to evaluate potential employees' professional Chinese proficiency. This was because they believed that the HSK does not accurately reflect candidates' ability to use the language at a professional level.

Compared with the score users in AS, these business representatives were much less familiar with the HSK and other Chinese language tests. However, this was not the only challenge, as respondents have stated that the test supplier has not provided a meaningful way to interpret the scores so the HSK can be used in business contexts. It would be more helpful if the test suppliers could train these users to "have a good professional understanding of the test's theoretical or conceptual basis, technical documentation, and guidance on the use and interpretation of the scale scores" (International Test Commission, 2011, p.106).

A holistic comparison of the companies/organizations surveyed indicated that those who had set clear criteria including a Chinese proficiency test certificate in their employment procedure paid much more attention to improving their employees' Chinese proficiency than those who did not, both before and after being hired. For example, NS2 noted that because China was his company's largest market, it has used the HSK certificate as an important criterion during employment/promotion to ensure their employees' Chinese proficiency. He also noted that Chinese courses were provided by his company and that they used the HSK test results as a periodical evaluation method. In addition, they also had a policy where candidates could get a bonus if they passed a higher level of the HSK. As such, his company had formed a very good Chinese learning atmosphere, and their employees' Chinese proficiency level improved immensely. In contrast, for representative companies/organizations who did not use the language certificate requirement in their employment/promotion procedure, most of them did not provide incentives to learn the Chinese language.

Among the HSK users in companies/organizations, it was interesting to find that South Korean companies were the ones who required the HSK most (3 out of 4) and that they were most supportive about using the HSK score/level in the employment/promotion procedure. The

close trade relation between China and South Korea was an important cause of this phenomenon. For instance, China has become the biggest overseas market of Samsung Electronics since 2005, and now Samsung is the biggest multinational corporation in China. It was reported that in 2014, Samsung China clearly stipulated that all Korean employees should learn Chinese and take the HSK test, and they also stated that their employees' Chinese proficiency and the HSK level/score were an important reference factor for promotions and pay increases. The HSK's test-taker distribution around the world also indicated that Asia is the best-represented continent and Koreans accounted for 54.37% of all the participants (Luo et al., 2012). Among the Korean test-takers, students who planned to study in China contributed the most, while business people were in second place. NS2 indicated that there was pressure for his acquisition of Chinese by his company's policy, Korean colleges, friends, and family, as well as the test-driven culture of Korea. He viewed the HSK test-taking experience as a motivator and a way to set new learning goals.

6.5 Conclusion

In general, the HSK levels/scores and other related information (e.g., score report, level interpretation) provided users with relevant, useful, and meaningful data for making decisions about candidate selection. In AS, the HSK was widely recognized by Chinese universities for international students' degree education admissions. The HSK 4, as suggested by the MOE, was considered the most common cut-off score. However, in order to academically succeed in the university after admission, students must continue to improve their proficiency in the language. In NAS, the findings implied that holding a HSK certificate gave candidates an advantage when seeking employment/promotion, but more importantly, they needed to have satisfactory Chinese proficiency to pass employers' internal tests or to fulfill work responsibilities that require

Chinese. In addition, the HSK provider must continue to actively build and promote the assessment literacy of all test users to increase their understanding of test score interpretation. Study 3 thus concluded the three phases of the larger MMR study. In the next chapter, a discussion based on the primary findings of all three studies is presented using the AUA framework.

Chapter 7 Discussion

7.1 Introduction

In Chapters 4, 5, and 6, the perceptions of CSL teachers and test-takers towards the content, use, and impact of the HSK were explored. Potential washback effects in CSL instruction/learning and possible relationships between perceptions of the test and teaching/learning practices have also been examined. Furthermore, the perceptions of other score users in both AS and NAS concerning score interpretation, as well as the decisions made based on HSK levels/scores, were investigated. This chapter expands on the primary findings of all three studies as well as the AUA for the HSK (as discussed in Chapter 2); the chapter will also include a merged interpretation (see Figure 3.1) of the extent to which the results can support or refute the warrants underlying the three claims in the AUA framework (i.e., consequences, decisions, and interpretations).

7.2 The consequential validity argument for the HSK

7.2.1 Intended consequences of the test developers

Before moving to the argument of the claims, the test developers' intended consequences are presented. In the case of the HSK, there are three main stakeholder groups: learners (i.e., test-takers), teachers (e.g., CSL/CFL teachers), and the score users in AS and NAS (e.g., university administrative officers, program coordinators, employers). The intended consequences for these stakeholders are listed in Table 7.1, which is synthesized from the test-developers' documents and the results of the MMR study.

Table 7.1

Intended Consequences for Stakeholders

Stakeholders	**Intended beneficial consequences**	
	Of using the HSK	**Of the decisions that are made**
Test-takers	Test-takers in the CSL/CFL programs will realize that the test is integrated into the teaching, their learning will benefit by taking the test, and thus it is relevant to their target language use needs. They will know their language proficiency level, strengths, and weaknesses through the effective feedback provided by taking the HSK. In addition, by getting a qualified HSK level/score, they will be able to obtain an advantage and/or qualify for an educational program, graduate with a degree, compete for academic scholarships, and fulfill job requirements.	The score interpretation will provide relevant and sufficient information to make decisions. Test-takers are classified only using the cut-off levels/scores and decision rules, and not according to any other considerations. They will also be fully informed about how the decisions are made and whether decisions are actually made in the way described to them. The decisions will be fair to all test takers, realizable to the Chinese language use domain in which the decision is made.
Teachers	Teachers will benefit from using a test that promotes desirable instructional practice and effective learning, and help make placement decisions. They will also understand the strengths and weaknesses of their instruction if their program uses the test performance-based class evaluation.	CSL/CFL teachers will benefit from being able to focus their instruction on a group of students who are relatively homogeneous in their language abilities. The CSL/CFL program will be able to evaluate teachers' instruction through their students' HSK performance. (For the teachers in introductory level university academic classes, they will benefit from having students who are able to have sufficient proficiency in their classes.)
Score users (e.g., University administrative officers,, employers)	The score users will benefit from using a test whose scoring criteria are consistent with the performance objectives for the course/program/work they supervise.	The score users can gain access to potential employees who, based on their HSK level/score, are able to fulfill their position's linguistic requirements.

As illustrated in the table, HSK's intended consequences are for different stakeholders to gain specific benefits based on how the test is leveraged in various areas (e.g., teaching, learning, evaluation, and selection). For example, in terms of university admission, by adopting the HSK 4 as the cut-off score, the administrative staff can ensure that the candidates' proficiency is sufficient for the university study.

7.2.2 Revisiting the AUA framework of the HSK

Following the AUA framework for the HSK, a summary table (Table 7.2) of the validity arguments for the HSK concerning the three claims (i.e., consequences, decisions, and interpretations) is presented, which illustrates warrants, backing/rebuttal evidence, and judgments of each claim based on the data from this dissertation study. A discussion of the claims is provided afterward, with a focus on the information not widely covered by the discussion sections for Studies 1, 2, and 3; some of them may intertwine in more than one claim.

Table 7.2

The Validity Argument for the HSK Based on the AUA Framework

Claims	Warrants	Backing (and/or rebuttal evidences)	Judgments
Consequences	W1-1: The consequences of using the HSK that are specific to immediate stakeholder groups (students, teachers, programs) will be beneficial. W1-2: In language instructional settings, the HSK promotes desirable instructional practice and effective learning; the use of the HSK is thus beneficial to students, teachers, the programs, and so on.	- Test-taker interviewees felt that the HSK played a motivational function in their learning, but unintended or negative consequence was also generated. - Teacher interviewees did not feel that the test affected their teaching practices and beliefs. - Acceptable model fit with significant factor loading and strong correlations were found in SEM of washback on test-takers. - Acceptable model fit with no causation relationship between test's content, use, and teaching practices was found in the SEM of washback on teachers.	Warrants partially supported
Decisions	W2-1: Decisions made on the HSK scores take into account the existing educational and societal values and relevant legal requirements in both academic and non-academic settings. W2-2: Test takers are classified only according to the cut-off scores and decision rules, and not according to any other considerations; test takers and other affected stakeholders are fully informed about how the decisions are made and whether	- Interviews with multiple stakeholders suggested that the PCI policy affected decisions made on the HSK levels/scores. - Descriptive statistics results from the AS test users' decisions made on the HSK levels/scores does not account for students' actual proficiency and its possible negative impact on their future learning in the university. - Both quantitative and qualitative results of the test users in NAS showed that potential employees need to have satisfactory Chinese	Warrants partially supported

	decisions are actually made in the way described to them.	proficiency, but the HSK is not widely required.	
Interpretations	W3-1: The HSK is meaningful and generalizable for its content representativeness and relevance in accordance with the Scales and the curriculum objectives. W3-2: The assessment tasks do not include content that offend or favor test takers, and individuals are treated impartially during the whole assessment procedure. W3-3: The assessment-based interpretation provides relevant and sufficient information to make decisions.	- Descriptive statistics results of the test-takers' and teachers' questionnaires demonstrated an overall positive attitude toward the test content, uses, and impacts. The teachers believed that the HSK reflects the goals and objectives of the Scales and the Standards. - The analysis of the questionnaires indicated that they deemed the test level and score to be a generally appropriate indicator of their overall Chinese ability, except for the HSKK (which cannot fully reflect their speaking ability). - The interviews with the test-takers and teachers also showed there are still limitations with the test design. - The analysis of test users' questionnaire in AS illustrated that they doubted the test's fairness and accuracy in reflecting one's proficiency. It also showed that they did not believe score-based interpretations could provide sufficient information for them to make decisions.	Warrants partially supported

Claim 1: Consequences

Claim 1 of the AUA framework asserts that the consequences of HSK use in decision-making are positive. Potential effects of the HSK on the direct stakeholders (i.e., students,

teachers, and programs) are summarized in Table 7.3. The table indicates that both positive and negative effects were found for all groups of stakeholders, which confirmed the complexity of washback reported in other research (e.g., Andrews, Fullilove & Wong, 2002; Cheng, 1998, 2005; Hamp-Lyons, 1997; Qi, 2005; Shih, 2007) and highlighted the motivational role the test can play in the educational context. As Harlen (2007) noted, "the impact that assessment can have on students can be either positive or negative. What happens depends on how the teacher mediates the impact of assessment on students." (p.181). In the current study, although the quantitative data showed that the test had a minor influence on teaching, the qualitative data revealed the existence of washback effects in classroom instruction. For example, the teachers utilized HSK test questions in homework and exams (both formative and summative purposes), in addition to using HSK-related textbooks.

Table 7.3

Potential Effects of the HSK on Teachers, Learners, and Their Programs

	Positive effects	**Negative effects**
Effects on teachers	- Encourages teachers to improve instruction and helps them align instruction with standards; - Motivates teachers to adjust their teaching methods; - Supports better diagnosis of individual student's needs and redirect instruction;	- Persuades teachers focus more on specific test content than on curriculum standards; - Leads teachers to engage in inappropriate test preparation;
Effects on learners	- Encourages learners to improve their learning methods; - Motivates students to work hard; - Helps set clearer short-/long-term goals; - Provides learners with diagnosis information about their own knowledge and skills; - Helps learners succeed in scholarship/university application and/or job employment/promotion;	- Puts too much pressure on students; - Persuades learners to focus more on test-specific content rather than on acquiring comprehensive ability; - Discourages learners in their future academic study due to the insufficient language proficiency; - Causes learners to devalue classroom-based assessments;
Effects on programs	- Provides a reliable reference for admission/graduation decision-making; - Provides an effective threshold to screen candidates;	- Applies inappropriate cut-off score, which might raise the question of whether the students have sufficient language proficiency in study in the university;

The motivational function of an assessment was highly related to its stake status, such that any high-stakes test may motivate learning and teaching (e.g., Anderson et al, 1990; Burger & Kroeger, 2003; Harlen, 2007; Sun, 2016). The descriptive statistics analysis of why participants took the HSK indicated that most of them completed the exam because they wanted

to identify their language proficiency level and to pursue higher education, while a small number of them wanted to increase their career opportunities in China. Their motivation was demonstrated at two interrelated categories: 1) personal motivators - individual interest in Chinese language and culture and sustained efforts in spending time on Chinese learning and resorting to different learning resources (including test-preparation resources) were important motivators; 2) instrumentally oriental motivators. These were strongly related to the high stakes of the test and its instrumental use for learners (e.g., university application, degree conferment, and employment). Researchers have noted that instrumental goals are associated with the utilitarian values of learning a new language (Gardner, 2006).

Research (e.g., Wang, 2016; Xie & Andrew, 2012) also demonstrated that motivation was highly correlated with outcome, and higher effort/behaviors would yield better performance. These goals could persuade learners to focus more on specific test content rather than on acquiring comprehensive language abilities, which might result in negative washback. Thus, to some degree, the unintended consequences distorted the intended purposes of the HSK.

In sum, although there is much evidence demonstrating HSK's positive impact, the unintended negative aspects were still notable. As such, the consequence claim is only partially supported.

Claim 2: Decisions

Assessments are developed for use in a particular educational system or social segment along with the corresponding values of that context (Bachman & Palmer, 2010). In order to investigate the test's use and the decisions made by the assessment, it is crucial to identify the stakeholders involved in this procedure. Table 4.1 presents the decisions made based on the HSK levels/scores, the stakeholders affected by said decisions, and the individuals responsible for

making these decisions. These level/score-based decisions generally fell into two categories: 1) institutional decisions made by programs, such that "the HSK could provide a reference for educational institutions' decision-making concerning recruiting students, assigning students to different classes, allowing students to skip certain courses, and granting academic credits to students" (Office of Chinese Language Council International, 2010, p.2); and 2) decisions made by employers in the social dimension, such that "the HSK could provide a reference for employers' decision-making concerning the recruitment, training, and promotion of test takers" (Office of Chinese Language Council International, 2010, p.2). The official HSK handbook suggested that in order to obtain a HSK certificate of a certain level, the passing score of each HSK level was set as 180 out of 300 (or 120 out of 200), and the subsets (i.e., listening, reading, and writing) were set at 60 out of 100. The following table (Table 7.4) shows the cut-off scores of the HSK. These cut-off scores were widely applied in international student admission/employment decision-making procedures. For example, Chinese MOE's regulation suggested that foreign students who enrol in Chinese degree program at universities needed to pass the HSK level 4 or above. However, a recent HSK technical report (Zhang & Zhang, 2014) noted that in order to better utilize the HSK's incentive role, there was no suggested passing score of HSK 5 and 6 after 2013 and the passing score of other levels would also be removed in the future.

Table 7.4

Cut-off Scores of HSK

	Level	**Listening**	**Reading**	**Writing**	**Speaking**	**Total**	**Cut-off**
HSK	Level 1	100	100	n/a	n/a	200	120
	Level 2	100	100	n/a	n/a	200	120
	Level 3	100	100	100	n/a	300	180
	Level 4	100	100	100	n/a	300	180
	Level 5	100	100	100	n/a	300	180
	Level 6	100	100	100	n/a	300	180
HSKK	Elementary	n/a	n/a	n/a	100	100	60
	Intermediate	n/a	n/a	n/a	100	100	60
	Advanced	n/a	n/a	n/a	100	100	60

The test developers explained that a passing score was set because 1) it continued the tradition of the old HSK; 2) it met the test-takers' needs; and 3) it further reduced the test's difficulty level. However, according to the feedback from CSL teachers, the difficulty of the test was relatively low. If the cut-off score was set as 180 out of 300, test-takers who reached 180 in HSK 6 would believe that they have achieved the highest proficiency level, while it was inconsistent with their actual linguistic capabilities. Their subsequent complacency was misled by the cut-off score, as it no longer encouraged them to continue learning the language. This consequence was the opposite of the test's original intention. As such, the cut-off scores of HSK 5 and 6 were removed. In addition, the score reporting system was also reformed by involving norm-referenced scores and criterion-referenced scores of the sub-sections (i.e., listening, reading, and/or writing) and the total score. From the test-takers' perspective, they were aware that the inclusion of subset scores could more accurately reflect their proficiency as well as better

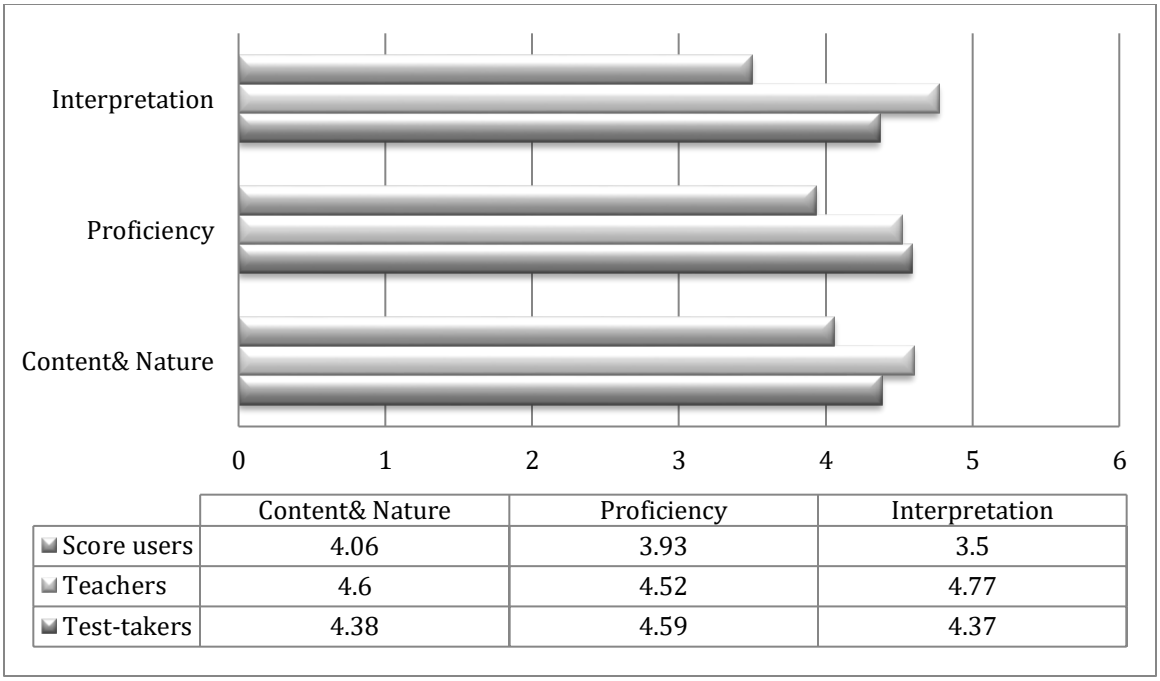
diagnose their strengths and weaknesses on each skill. However, according to the comment from one test-taker, "(due to the new score report), we need to make more effort to improve our scores because all three components are displayed on the report…It discourages students from aiming at a narrow pass or fail, and this causes a lot of pressure." While test-takers acknowledged that the score report was informative and diagnostic, the score users (especially those from the NAS who have relatively less knowledge of the test) were less satisfied. They argued that the test supplier had not provided a meaningful way to interpret the scores so it can be used in business contexts. It would be more helpful to if there was guidance/suggestions about how to use the scores (e.g., setting passing scores, and setting score range scales: 0-180, fail; 181-210, pass; 211-240, good; and 241-300, excellent). Furthermore, training on test literacy was provided to users involved in admissions/employment decision-making. According to the findings in AS, cut-off scores were set to help identify test takers' proficiency and to help make classification or selective decisions. The major types of institutional decisions based in the HSK level/score are 1) admission decisions; 2) graduation decisions; 3) placement decisions; and 4) teacher evaluation decision (in part). Although the participants benefitted from applying the cut-off level/score when making decisions, they also involved other admission criteria besides providing evidence of language proficiency. This was because the decision did not take into account test-takers' actual proficiency and its possible negative impact on their future learning in the university. Thus, such policy (i.e., the reform of the score reporting system and setting/removing cut-off scores) was considered more from its motivational function aspect, while the test developers insufficiently took into account the opinions from the other score users. As Bachman and Palmer (2010) proposed, stakeholders in the decision-making process should be widely consulted. This was because the test developer needed to conduct research to explore all groups of stakeholders'

perceptions related to a HSK decision. In this way, not only can the motivational power of such policy be strengthened, but other groups of score users can also benefit from the policy when they make admission/employment decisions.

In sum, the findings of the study indicated that different organizations adjusted the cut-off level according to their actual needs in candidate selection. In order to inform applicants, they disseminated information related to the language proficiency test and other requirements on their website or their official recruitment notice, so that the organization's recruitment or hiring policy was as transparent as possible. That is to say, to a great extent, test-takers and other affected stakeholders were fully informed about how the decisions were made. However, it was not guaranteed that test-takers were classified only according to the cut-off scores and decision rules. Thus, Claim 2 is also only partially supported.

Claim 3: Interpretation

In order to have a direct visual comparison of test-takers', teachers', and score users' perceptions of the test interpretation, a bar chart (Figure 7.1) was produced. As illustrated, there is a discrepancy between score users' attitude and test-takers' and teachers' attitudes. More specifically, teachers and test-takers hold positive perceptions of the test content, construct, and nature, and they believed that the HSK could provide a relative accurate measure of test-takers' overall Chinese proficiency. Furthermore, they thought the score-based interpretation provided relevant and sufficient information for decision-making. While holding less positive attitudes on the test content and nature (m=4.06), the score users were also unsure as to whether the HSK reflects one's actual proficiency (m=3.93); they even denied HSK's interpretation function (m=3.50).



	Content& Nature	Proficiency	Interpretation
Score users	4.06	3.93	3.5
Teachers	4.6	4.52	4.77
Test-takers	4.38	4.59	4.37

(Notes: Content & nature refers to the content, format, and nature of the HSK; proficiency refers to whether the HSK provides an accurate measure of test-takers' overall Chinese proficiency; and interpretation refers to whether the score-based interpretation provides relevant and sufficient information to make decisions.)

Figure 7.1. A comparison chart of test-takers', teachers', and score users' perceptions

All three groups believed that cultivating communicative skill should be the ultimate goal of learning Chinese and taking the HSK. It was consistent with the core focus of the *Scales* and *Standards*, the official guidelines for CSL/CFL teaching and learning. In order to understand the theoretical basis of the new HSK's construct and the reform of the HSK, it is necessary to investigate how the *Scales* and *Standards* defines the construct based on language ability. When the HSK was launched in 1980s under the psychometric structuralist approach, the old HSK was designed as a test that focused on measuring test-takers' linguistic knowledge. On the other hand, due to Bachman's proposal of communicative language ability in the 1990s, well-established large-scale language tests like TOEFL and IELTS were designed as communicative tests in response to the worldwide demands for learners' communicative language proficiency. The HSK

reform kept pace with this trend by including more communicative-oriented tasks and constructive response items. The new HSK increased the weight of such tasks, developed the HSKK (the oral test), and reduced the tasks that focused on linguistic aspects. In this sense, it was a positive and appropriate reform to exert positive washback on CSL/CFL learning and teaching.

From a top-down perspective, the *Scales* and *Standards* set detailed language skills requirements at different levels; the HSK specification also provided a description of test-takers' ability to use their knowledge and skills of the Chinese language for communication according to their levels. Table 7.5 presents the relationship between the New HSK Tests and *Scales*, and demonstrates that the development of the new HSK was greatly affected by the *Scales*. These official guideline documents served as references for drawing up CSL/CFL teaching syllabus, for compiling Chinese textbooks, and for assessing the language proficiency of CSL learners. In this sense, the HSK was meaningful and generalizable for its content representativeness and content relevance in accordance with the *Scales* and the curriculum objectives.

On the other hand, from a bottom-up perspective, the teachers participating in this study acknowledged the importance of communicative language ability but asserted that linguistic competence should serve as its basis. Some of them pointed out that the old HSK was more advanced than the new HSK in terms of its validity and reliability in reflecting learners' actual linguistic knowledge and ability. In addition, they also indicated that their teaching was not closely related to the *Scales* and the HSK. In terms of the test-takers in this study, they spent time memorizing the vocabulary and grammar points listed in test specifications. It was because that they thought the vocabulary served as a foundation for understanding the test content. Such test-specific learning and test-taking skills could help them eliminate distractors (e.g., use advanced

vocabulary and structures and avoid mistakes in writing). They acknowledged the importance of gaining communicative language (particularly oral communication), but they did not spend much effort on speaking during their test preparation. This might be because the HSKK was a separate test or that they doubted the HSKK could fully reflect their speaking ability. According to the findings in CSL programs, opening a HSK preparation course was a common practice in some universities. The regular CSL class covered test-related content to some extent, mainly according to the program or students' needs. On the other hand, most regular CFL courses seldom covered HSK-related content in teaching. CFL representatives in this study expressed that they did not want their teachers to teach test-related content in their limited class time. This meant that the test-takers did not receive equal opportunities to learn the test in class, but they could prepare for the test by themselves and/or from any other resources. Consequently, Warrant 3-2 "the individuals are treated impartially during the whole procedure of the assessment" was achieved to an extent. Ultimately, Claim 3 is only partially supported.

Table 7.5

Relationship among the New HSK Tests and Scales

New HSK	**Vocabulary**	***Scales***
HSK Level 6	5000	Band 5
HSK Level 5	2500	
HSK Level 4	1200	Band 4
HSK Level 3	600	Band 3
HSK Level 2	300	Band 2
HSK Level 1	150	Band 1

(The Office of Chinese Language Council International, 2010, p. 1)

7.3 Conclusion of the chapter

According to the test developers, the HSK is intended to achieve two interrelated goals, namely: 1) to act as a reference for educational and social decisions centered on individuals' Chinese language proficiency; and 2) to promote CSL/CFL teaching and learning. In order to justify the consequential validity of the HSK by employing the AUA framework within the PCI context, this MMR study's findings provide evidence that Claim 1 (Consequences), Claim 2 (Decisions), and Claim 3 (Interpretations) are all partially supported, such that the intended goals have only been achieved to some degree. For the students in this study, the HSK seemed to play a motivational role. For teachers, the test appeared to affect and influence their teaching practices and beliefs in different ways. Furthermore, for test users, the HSK achieved its intended consequence to a great extent in terms of providing test users with useful information for making decisions in the educational context; this was not true in the social context. However, there still are limitations of the test (e.g., HSK test design, test difficulty level, score reporting system). These limitations may contribute to unintended negative consequences for teachers, students, and

other score users. They may also give rise to validity and ethical concerns about the test. In the next and final chapter, implications for the HSK test developers and other stakeholders, limitations of this MMR study, and directions for future research are discussed.

Chapter 8 Conclusion

8.1 Summary of the findings and discussions

The rise of China has brought progress and opportunities to many people and countries around the world. China is not imposing its language on other people or nations; rather, other people and nations are becoming increasingly motivated to learn the Chinese language. Against this backdrop, teaching and learning of CSL worldwide has created a breeding ground for HSK development and reform. This context also provides opportunities to examine the HSK under the PCI policy using both top-down and bottom-up approaches from the perspective of consequential validity. Adapting Bachman and Palmer's (2010) argumentative approach, an AUA conceptual framework for the HSK was established, which provided a methodological guideline for this MMR study. Moreover, by reviewing and learning from existing models (e.g., the hybrid model of English language teaching innovation in Japan, Henrichsen, 1989), the current MMR study offers unique insights into ways to improve the CSL and to implement the new HSK.

This dissertation study employed a mixed-methods sequential exploratory (MMSE) design, whereby a qualitative study (i.e., Study 1) was first conducted to identify theoretical issues and to develop the measurement instruments and hypotheses for the subsequent quantitative study. Quantitative studies (i.e., Study 2 and 3) were then carried out to identify whether concepts/issues established from a comparable small number of cases could be described and explained in a greater domain (Creswell, 2015; Kelle, 2006). The summaries of each study are provided below.

Study 1: In this qualitative research, the data were obtained from Hanban's official reports, technical reports, and official documents that reflect the test developers' intentions. Interview data collected was from twelve HSK stakeholders (i.e., an officer from the Education

Office of the Consulate General of the PRC, four CSL teachers, four test-takers, an administrative officer from a Chinese university, a human resources manager from a multinational enterprise, and a HSK test center director) to elicit a multifaceted understanding of the HSK's consequential validity. By adopting a two-cycle analysis approach (Saldaña, 2009), the results from the Nvivo analysis revealed that 1) the intended consequence of promoting CSL and CFL teaching and learning has only been achieved to a limited extent; 2) the HSK achieved its intended consequences in terms of providing test users with information for making decisions only in the educational context; and 3) the HSK reform demonstrated how implementing the PCL policy helped enhance the test's quality. The findings of this study and previous washback/impact/consequence research were then used to develop four questionnaires.

Study 2: This study adopted a sequential explanatory mixed-methods design to establish the relationships between CSL/CFL teachers' and test-takers' actual classroom practices and their perceptions towards the washback effect and use of the HSK. The questionnaire participants consisted of 136 CSL/CFL teachers and 512 HSK test-takers, and data were analyzed quantitatively (i.e., EFA, SEM) using SPSS 24.0 and Amos 24.0. In order to understand the quantitative results, six classroom observations were conducted in a HSK preparation class from a CSL program. The findings of this study demonstrated that the HSK was somewhat successful in its goal of promoting CSL/CFL learning, but it did not really inform teaching. This study's findings demonstrated the complexity of the HSK's washback effects. Both the test-takers and teachers believed that there were limitations to the HSK (e.g., the task type), which may subsequently induce negative washback on teaching and learning.

Study 3: By analyzing data from two exploratory questionnaires and the semi-structured interviews, the findings show that HSK scores and other related information (e.g., score report,

level interpretation) generally provided users with relevant, useful, and meaningful data for candidate selection. In AS, Chinese universities widely recognize the HSK and use it as a major requirement for the admission of international students. In NAS, holding a HSK certificate gave candidates an advantage when seeking employment/promotion, but their ability to work in a company was seen as more important than their Chinese language proficiency.

Overall, based on the AUA conceptual framework of the HSK, the findings provided evidence that Claim 1 (Consequences), Claim 2 (Decisions), and Claim 3 (Interpretations) were partially supported, such that the test developers' intended goals for the HSK were only achieved to a certain degree. For example, at the classroom level, the HSK seemed to play a motivational role for CSL learners and influenced their test-taking strategies; however, it had no significant effect on CSL teachers' teaching practices and beliefs. At the macro level, the HSK greatly achieved its intended consequence in terms of providing test users with useful information for making decisions. However, this finding was limited to the educational context and was not found in the social context. Moreover, the findings of this study also highlighted the unintended consequences of the test. For instance, the test not only created validity and ethical concerns, but also had unintended negative consequences for the stakeholders (e.g., teaching/learning to the test, inadequate language proficiency in future study).

8.2 Contributions of the MMR study

By adopting an argument-based approach to verify HSK's consequential validity in the CSL context, the findings of this study contribute to a deeper understanding of the consequences and uses of the HSK. As mentioned in the previous chapters, argument-based approaches in LT have become increasingly popular and have provided a new perspective for conducting validation research (Chapelle et al., 2008; Chapelle & Voss, 2013; Knoch & Chapelle, 2017).

This study was among the first few attempts to investigate a large-scale high-stakes Chinese proficiency test and to apply such a framework to it in a mixed methods study.

First of all, employing the MMSE design in an argumentative validation research study not only can enable researchers to explore a specific educational/societal context at the macro level, but also allows them to examine and gain detailed insights on specific cases at a micro level. More specifically, the MMSE design was appropriate and useful for this study in terms of its overall design. Firstly, Study 1 (a qualitative study) was conducted to explore the HSK's actual consequences and to develop measurement instruments and hypotheses for the subsequent quantitative studies. Study 2 and 3 were then carried out to test the instruments and to apply the findings more broadly (e.g., both at micro and macro levels). Triangulation of the data and findings at any phase of the design allowed the current author to obtain multiple perspectives on the consequence/impact/washback effects to support or refute claims in an AUA framework, to explore the research questions more deeply, to provide more convincing findings than monomethod studies, and to give a more comprehensive analysis that can enrich the existing research methods in the washback literature.

Second, the sub-studies (i.e., Study 2 and 3) also adopted MMR methods, which was another important methodological feature of this dissertation study. For example, Study 2 was a quantitative oriented study that utilized an MM sequential Explanatory design. The follow-up classroom observation (qualitative phase) helped explain the quantitative findings concerning the HSK's washback. In Study 3, the interviews with AS and NAS test users after the questionnaire phase provided deeper meaning to the quantitative results, especially when unexpected results occurred. In all, the MMR could overcome the weaknesses of, while leveraging the strengths of, a monomethod design.

Moreover, as Mathew (2004) argued, the incorporation of different stakeholders' perspectives is very important to test development and validation. This study collected evidence from multiple stakeholders, and effectively linked test validity to test consequence and use. It also highlighted the importance of understanding a phenomenon through its context; in other words, the contextual factors also contributed to the test consequences in the social dimension.

Last but not least, this study provided insights for investigating the interactive relationship between language policy and language testing. On the one hand, since the implementation of the PCI policy, Chinese has become an increasing popular language to acquire and the HSK has been revised systemically and comprehensively. On the other hand, the test has also played a significant role in implementing the policy in CSL/CFL contexts. The test consequence boost interest in learning Chinese. For example, by reducing the difficulty level of the test, more test-takers are more motivated to take the exam and the passing rate has increased. In order to continue to increase the exam's quality, the test developers should collaborate with other stakeholders. By doing so, the test will become increasingly credible and the intended consequences of improving CSL education can be achieved.

8.3 Implications of the MMR Study

The findings of this MMR study provide several implications for HSK's future developments. First of all, since the goal of the HSK is to support the relationship between teaching and testing, and to facilitate teaching and learning through testing [考教结合，以考促学、以考促教], the test needs to more closely reflect CSL/CFL curriculums and to provide an appropriate and expansive interpretation of the *Scales*. Although the test developer asserted the connection and equivalence of the HSK to the *Scales*, no convincing empirical evidence has been released. The separation between the test and the curriculum can lead to certain issues, such as

unintended negative consequences of the test and construct underrepresentation. To address this issue, the HSK test should be redefined as a learning-oriented test that is based on the curriculum and the *Scales*. For example, by providing accompanying materials (e.g., textbooks and workbooks) for CSL/CFL teachers and students, more positive washback effects could occur. This may also generate a more positive cycle of learning, teaching, and testing, with assessments functioning as the pivotal mediator in the cycle. In addition, developers should consider varying the type and format of the tasks and broadening the range of topics and text types used in the tasks. Test quality can also be improved by incorporating interactive testing methods (e.g., integrated writing task in TOEFL test).

Second, the HSK developers should clarify how the test can be used and continue to actively build and promote the assessment literacy of all test users, as this will increase their test score interpretation abilities. For example, they can give a recommended HSK level/score for admitting students to higher education institutes; they can also provide a meaningful way to interpret the levels/scores of candidates' language proficiency in the NAS context. As suggested by researchers (e.g., Nicholas & Williams, 2009), test developers should collect information on test score use. HSK test developers should also investigate the consequences associated with the test's uses and how they can more effectively communicate with test stakeholders. In this way, the intended impact of the EPT can be achieved. This process can also inform the development of other large-scale high-stakes exams.

Since 2017, following the example of other renowned testing agencies in the world (e.g., ETS and Cambridge assessment), Hanban has started to sponsor external researchers to conduct studies on its tests. In light of this, a series of research projects[30] were undertaken to verify the

[30] See http://www.chinesetest.cn/gonewcontent.do?id=41477821 for more information.

test's quality and to inform its development. Hopefully, by collaborating with a large body of researchers, the HSK will become more credible, objective, and accepted by test users.

8.4 Limitations of the MMR Study and Suggestions for Future Research

This study attempted to advance the current understanding of the consequential validity of a particular test; however, it is not without its limitations.

First, in Study 2, the researcher relied on the questionnaires (i.e., test-taker and teacher questionnaires) as the main instruments for answering the research questions. However, the quality of self-reported data, such as the responses to statements on Likert scales, can be affected by numerous factors (e.g., personality of respondents, and respondent fatigue or boredom).

A second limitation of the study concerns the data collection from CSL/CFL teacher participants. Due to practical restrictions, only one class was observed in Study 2. It would be difficult to know whether other test preparation classes were similar to it. In addition, considering logistical restrictions, teacher participants in this study were mainly sampled from the CSL context rather than CFL and CHL. Given the restrictions in participants and contexts, the generalization of the current study's findings to other contexts should be undertaken with caution. Moreover, since the number of teacher participants was relatively small, only a few factors were considered in the teacher's washback model. This could also partially explain why the proposed factor structure contributed to a less satisfactory model fit for the data. Moreover, in order to develop a comprehensive SEM model of washback involving teaching and learning, future studies should recruit test-takers and teachers who have a more direct relationship with each other (e.g., from the same course).

Another limitation would be the lack of genuine test-takers' test performance data. Despite efforts to obtain the performance data from the HSK test centers, the researcher was not

granted access to the database. The test scores of the test-takers were subsequently self-reported by category. The three reportable ranges were 0-179 (or 0-119), 180-239 (or 120-159), and 240-300 (or 160-200)[31]. Due to the HSK's high pass rate and the large range of each category, the accuracy of the scores was poor and was not used in data analysis. In future research, if raw scores can be obtained instead of categorical data, the relationships among the perspectives of the test, test-taking preparation strategies, and test performance can be more thoroughly examined. Moreover, in an argument-based validity framework, score evidence is also needed to show that the score data are consistent and reliable, and that they accurately represent the measured test constructs.

In sum, although this dissertation research is not without its limitations, the MMR study can be seen as the first attempt to investigate the consequential validity of the new HSK by using Bachman and Palmer's AUA framework in the TCSL context. It not only provides findings that can inform CSL teaching/learning and test development, but also complements current efforts in the field of LT to broaden our understanding of the consequences of language assessments from educational and social perspectives.

[31] The HSK has a maximum score of 200 (with 120 required to pass) at elementary level, and 300 (with 180 required to pass) at intermediate and advanced levels.

REFERENCE

Alderson, J.C. & Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing, 13*(3), 280-297.

Alderson, J.C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(3), 115-129.

Anderson, J. O., Muir, W., Bateson, D. J., Blackmore, D., & Rogers, W. T. (1990). *The impact of provincial examinations on education in British Columbia*. British Columbia, Canada: Ministry of Education.

Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 37-50). Mahwah, NJ: Lawrence Erlbaum Associates.

Andrews, S. & Fullilove, J. (1994). Assessing spoken English in public examinations- why and how? In J. Boyle & P. Falvey (Eds.), *English language testing in Hong Kong* (pp. 57-86). Hong Kong: Chinese University Press.

Andrews, S. J., Fullilove, J., & Wong, Y. (2002). Targeting washback: A case study. *System*, *30*, 207–33.

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

Bachman, L. F. (2004). Linking observations to interpretations and uses in TESOL research. *TESOL Quarterly*, *38*(4), 723-728.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

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

Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Baker, B., Tsushima, R., & Wang, S. (2014). Investigating language assessment literacy: Collaboration between assessment specialists and Canadian university admissions officers. *Language Learning in Higher Education, 4*(1), 137-157.

Banerjee, J. & Luoma, S. (1997). Qualitative approaches to test validation. *Encyclopedia of language and education, 7*, 275-287.

Bellassen, J. (2011). "Is Chinese Europcompatible? Is the Common European Framework Common? The Common European Framework of References for Languages Facing Distant Language", *New Prospect for Foreign Language Teaching in Higher Education —Exploring the Possibilities of Application of CECR*. Tokyo: World Language and Society Education Center.

Bogdan, R. & Biklen, S. K. (1998). *Qualitative Research for Education: An Introduction to Theory and Methods.* London: Allyn & Bacon.

Brown, J. D. (1997). Designing surveys for language programs. In D. Griffee & D. Nunan (Eds.), *Classroom teachers and classroom research* (pp. 109-122). Tokyo: The Japan Association for Language Teaching (JALT).

Brown, H. (2004). *Language Assessment: Principles and Classroom Practices.* New York: Pearson Education.

Browne, M. & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

Burger, J. M., & Krueger, M. (2003). A balanced approach to high-stakes achievement testing: An analysis of the literature with policy implications. *International Electronic Journal for Leadership in Learning, 7*(4).

Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian adult migrant English program. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.113-128). Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chalhoub-Deville, M. & Turner, C.E. (2000). What to look for in ESL admission tests: Cambridge Certificate exams, IELTS, and TOEFL. *System, 28*, 523-539.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369-383

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment*, pp.1-13. Oxford: Wiley.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). London: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice.*

Chen, H. (2006). 结构效度与汉语能力测验—概念和理论[Construct validity and Chinese language proficiency tests—concept and theory]. In K. Zhang, (Ed.), 汉语测试理论及汉语测试研究 [Test theory for Chinese and Chinese assessment research] (pp.200-225). Beijing, PRC: The Commercial Press.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Educatio*n, *11*(1), 38-54.

Cheng, L. (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. (2001). Washback studies: Methodological considerations. *Curriculum Forum*, *10*(2), 17-32.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp.147-170). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: A Washback Study.* Cambridge: Cambridge University Press.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 2479-2494). New York, NY: Springer Science + Business Media LLC.

Cheng, L. (2014). Consequences, impact and washback. In A.J. Kunnan (Ed.), The companion to language assessment (pp.1130-1146). John Wiley & Sons. Doi:10.1002/9781118411360.wbcla071.

Cheng, L. & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within an argument-based validation framework. *Language Assessment Quarterly, 12*(1), 50–66.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing*, *28*, 221-249.

Cheng, L., Klinger, D., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing*, *24*(2), 185–208.

Creswell, J., & Plano Clark, V.L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: SAGE.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. (2015). *A concise introduction to mixed methods research.* Thousand Oaks, CA: Sage.

Custer, C. (2010, December 13). How many people are learning Chinese? *The World of Chinese.* Retrieved from http://www.theworldofchinese.com/2010/12/how-many-people-are-learning-chinese/

Davies, A. (1997). Special issue: Ethics in language testing. *Language Testing, 14*.

Davison, C. (2006). Collaboration between ESL and content teachers: How do we know when we are doing we are doing it right? *The International Journal of Bilingual Education and Bilingualism, 9*(4), 454 – 475.

Davison, C. (2008). *Using summative assessments for formative purposes: The ultimate justification for learners and teachers.* Presented at 30th Annual Language Testing Colloquium, (LTRC 08). Hangzhou.

Doe, C. (2015) Student Interpretations of Diagnostic Feedback. *Language Assessment Quarterly, 12*(1), 110-135.

Ferman, I. (2004). The washback of an EFL national oral matriculation. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 191-210). Mahwah, NJ: Lawrence Erlbaum Associates.

Gardner, J. (Ed.) (2006). *Assessment and learning.* London, UK: Sage.

Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175-199). Mahwah, NJ: Lawrence Erlbaum.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24,* 355 - 392.

Gipps, C. (2002). Sociocultural perspectives on assessment. In G. Wells and G. Claxton (Eds.), *Learning for life in the 21st century* (pp. 73-83). Oxford, UK: Blackwell.

Glover, P. (2014). Do language examinations influence how teachers teach? *International Online Journal of Education and Teaching, 1*(3), 17-26.

Gosa, C. M. (2009). *Investigating washback: A case study using student diaries.* Saarbrücken, Germany: VDM VerlaG.

Green, A. (2006). Watching for washback: Observing the influence of the International English Language Testing System Academic Writing Test. *Language Assessment Quarterly 3*(4), 333-368.

Green, A. (2007). *Studies in language testing: Vol. 25. IELTS washback in context – Preparation for academic writing in higher education*. Cambridge: Cambridge University Press.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Greene, J. C., Caracelli, V. J., & Graham W. F. (1989). Toward a Conceptual Framework for Mixed-method Evaluation Designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Gu, X. (2005). *Positive or Negative? An Empirical Study of CET Washback on College English Teaching and Learning in China.* Unpublished doctoral dissertation, Shanghai Jiao Tong University, China.

Guion, R. (1980). On Trinitratian doctrines of validity. *Professional Psychology, 11,* 385-398.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17–27.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, *14*, 295–303.

Hawkey, R. A. H. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000. Studies in language testing*, *24*. Cambridge, UK: Cambridge University Press / Cambridge ESOL.

Henrichsen, L.E. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan, 1956-1968.* New York, NY: Greenwood Press.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, *28*(4), 579-591.

Hanban (2014). *HSK [Chinese Proficiency Examination]*, Beijing: Hanban. Retrieved from http://www.hanban.edu.cn/tests/node_7486.htm.

Harlen, W. (2007). *Assessment of learning.* London, UK: Sage.

Herman, J. & Golan, S. (1991). *Effects of Standardized Testing on Teachers and Learning – Another Look. CSE Technical Report #334.* Los Angeles: Center for the Study of Evaluation.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Huang, C. (2013). HSK 对汉语作为第二语言教学中学习行为的反拨效应 [Washback effect of the HSK on learning behaviorin teaching Chinese as a second language]. *云南师范大学学报* [*Journal of Yunan Normal University (Teaching and Research on Chinese as A Foreign Language Edition)*], *2013*(1), 10-17.

Huang, C., & Li, G. (2010). HSK 对汉语作为第二语言教学的反驳效应 [Washback effect of HSK on teaching Chinese as a second language]. *中国考试* [*Chinese Examination*], *2010*(1), 26-32.

Huang, Y. (2014). 新 HSK 反拨效应研究：辅导课及语言课 [The washback effect of the new HSK: test preparatory class and regular language class]. 北京地区对外汉语教学研究生论文集[The proceeding of TCSL graduate students in Beijing].

Hughes, A. (2003). Introducing a needs-based test of English language proficiency into an English-medium university in Turkey. In A. Hughes (Ed.), *Testing English for university*

study (pp. 134-146). London, UK: Modern English Publications in association with the British Council.

Hughes, A. (1993). *Backwash and TOEFL 2000.* Unpublished manuscript, University of Reading.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practices*, *21*(1), 31–41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective, 2(3),* 1351–1370.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.17–64). Westport, CT: American Council on Education.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement 50*(1), *1–73.*

Kellaghan, T. & Greaney, V. (1992). *Using Examinations to Improve Education: A Study in Fourteen African Countries*. Washington, DC: World Bank.

Kellaghan, T. & Greaney, V. (2001) *Using Assessment to Improve the Quality of Education.* Paris, France: International Institute for Educational Planning.

Kelle, U. (2006). Combining qualitative and quantitative methods in research practice: Purposes and advantages. *Qualitative Research in Psychology, 3*(4), 293-311. doi: 10.1177/1478088706070839.

Kelley, T. L. (1927). *Interpretation of Educational Measurement*. New York: Macmillan.

Knoch, U. & Chapelle, C. (2017). Validation of rating process within an argument-based framework. *Language Testing.*

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment.* Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004) Test fairness. In M. Milanovic & C. Weir (Eds.), *European Year of Languages Conference Papers, Barcelona* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Lei, P., & Wu, Q. (2007). An NCME instructional module on introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice, 26*, 33–43.

Li, C. (2012). 汉语国际推广战略瓶颈及对策研究 [Bottleneck and conuntermeasure on international promotion of Chinese language]. *南阳理工学院学报 [Journal of Nanyang Institute of Technology], 2012*(4), 1-4.

Li, J. & Zhang, C. (2011). 论新汉语水平考试对对外汉语教材编写与出版的反拨效应 [The washback effects of the new HSK on CSL textbook compilation]. *中国出版杂志 [China Publishing Journal], 2011*(21), 36.38.

Liu, Y. (1994) 汉语水平考试(HSK)研究(续集) [HSK research (continued volume)]. Beijing, PRC: Modern Press.

Liu, X., Huang, Z., Fang, L., Sun, J., & Guo, S. (1986) 2006. 汉语水平考试的设计与测试**.** In K. Zhang (Ed.), 汉语水平考试研究 [HSK research] (pp.9-21). Beijing, China: The Commercial Press.

Liu, Y. (2013). Meritocracy and the Gaokao: A survey study of higher education selection and socio-economic participation in East China. *British Journal of Sociology of Education, 34*, 868–887.

Llosa, L. (2008). Building and Supporting a Validity Argument for a Standards-Based Classroom Assessment of English Proficiency Based on Teacher Judgments. *Educational Measurement Issues and Practice, 27(*3), 32-42.

Locke, L., Spirduso, W. W., & Silverman, S. J. (2000). *Proposals that work* (4th ed.). Thousand Oaks, CA: Sage.

Lu, J. (2014). 汉语国际教育专业的定位问题 [The orientation of teaching Chinese as a second language program], 语言教学与研究 Language teaching and Linguistics studies, *2014(2).*

Lumley, T. & Brown, A. (2005). Research methods in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (2nd ed., pp.833-857). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Luo, M., Zhang, J., Xie, O. & Huang, H. (2011). Report on Overseas Enforcement of New Chinese Proficiency Test (New HSK). *China Examinations, 4*, 17-21.

Madaus, G. F., & O'Dwyer, L.,M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695. Retrieved from https://proxy.library.mcgill.ca/login?url=https://search.proquest.com/docview/218467914?accountid=12339

Manjarrés, N.B. (2005). Washback of the foreign language test of the state examination in Colombia: A case study. *Arizona Working Papers in SLAT, 12*. Retrieved from http://w3.coh.arizona.edu/awp/AWP12/AWP12%5BManjarres%5D.pdf

Mathew, R. (2004). Stakeholder involvement in language assessment: Does it improve ethicality? *Language Assessment Quarterly, 1*, 123–135.

Maxwell, J.A. (1996). *Qualitative Research Design - An Integrative Approach*. Thousand Oaks, London: Sage.

McAloon, P. (2008). *Chinese at Work: Evaluating Advanced Language Use in China-related Careers*. Unpublished PhD dissertation, Ohio State University.

McLeod, J. (2009). *An introduction to counselling*. Maidenhead, UK: Open University Press.

McNamara, T. (2000). *Language testing*. Oxford, England: Oxford University Press.

McNamara, T. & Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA and Oxford, UK: Blackwell.

McNamara, T. (2008) The social-political and power dimensions of tests. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education Vol. 7: Language testing and assessment* (2nd ed., pp. 415–27). Dordrecht, The Netherlands: Springer.

McNamara, T.F. (2010). The use of language tests in the service of policy: Issues of validity. *Rev. franç. de linguistique appliquée*, *15*(1), 7–23.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational researcher, 10*(9), 9-20.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 1*(23), 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Meyer, F. K. (2014). *Language Proficiency Testing for Chinese as a Foreign Language: An Argument Based Approach for Validating the Hanyu Shuiping Kaoshi (HSK)*. Frankfurt am Mein: Peter Lang.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–62.

Muñoz, A. P., & Álvarez, M, E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing, 27*, 33-49.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3–9.

Noy, C. (2008). Sampling knowledge: The hermeneutics of snowball sampling in qualitative research. *International Journal of Social Research Methodology: Theory & Practice, 11*(4), 327-344.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

O'Loughlin, K. (2011). The interpretation and use of proficiency test scores in university selection: How valid and ethical are they? *Language Assessment Quarterly, 8*(2), 146–160.

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing, 30*(3), 363–380.

Odinye, S. & Odinye, I. (2012). Rise of China and Spread of Mandarin in 21st Century. *Quarterly Journal of Chinese Studies, 1*(4), 136-144.

Office of Chinese Language Council International. (2009). *Chinese Language Proficiency Scales For Speakers of Other Languages.* Beijing: Foreign language teaching and research press.

Office of Chinese Language Council International. (2010). *Chinese Language Proficiency Test Levels 1 to 6*. Beijing: The Commercial Press.

Pan, Y. & Roever, C. (2016). Consequences of test use: A case study of employers' voice on the social impact of English certification exit requirements in Taiwan. *Language Testing in Asia, 6*(6). DOI 10.1186/s40468-016-0029-5

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

Pellegrino, J.W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología_Educativa, 20*(2), 65–77.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments.* National Center for the Improvement of Educational Assessment, Dover, NH.

Phelps, R. (2012). The effect of testing on achievement: meta-analysis and research summary, 1910-2010. *International Journal of Testing*, *12*, 21–43.

Porter, A. C., & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. *Yearbook-National Society for the Study of Education, 2*, 60–80.

Prodromou, L. (1995). The backwash effect: From testing to teaching. *ELT Journal*, 49(1), 13-25.

Purpura, J. E. (2008). *Assessing grammar.* Cambridge: Cambridge University Press.

Qi, L. (2004). Has a high-stakes test produced the intended changes? In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 171-190). Mahwah, NJ: Lawrence Erlbaum Associates.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.

Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China, *Assessment in Education: Principles, Policy & Practice, 14*(1), 51-74.

Read, J, and Hayes, B. (2003) The impact of IELTS on preparation for academic study in New Zealand in *IELTS Research Reports, Volume 4*, IELTS Australia Pty Ltd, Canberra, pp 153-205.

Sadeghi, S. (2014). High-stake Test Preparation Courses: Washback in Accountability Contexts. *Journal of Education & Human Development, 3*(1), 17-26.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. Los Angeles, CA: SAGE.

Shih, C. (2007). A new washback model of students' learning. *Canadian Modern Language Review, 65*(1), 135 -162.

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests.* Harlow, UK: Longman.

Shohamy, E. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton and K. Toohey (Eds.), *Critical pedagogies and language learning* (pp. 72-93), Cambridge University Press.

Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches.* London, UK: Routledge.

Shohamy, E. (2007). Washback from language tests on teaching, learning and policy: evidence from diverse settings. *Assessment in Education, 14*(1), 1–7.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298-317.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*(4), 513-521.

Strauss, A. & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.

Sun, D. (2009). 汉语水平考试发展问题略论[Brief discussion on development issues of the HSK]. *Chinese Examinations, 6*, 18-22.

Sun, Y. (2016). *Context, construct, and consequences: Washback of the College English Test in China.* Unpublished PhD dissertation, Queens University.

Tan, M. (2009). *Changing the language of instruction for Mathematics and Science in Malaysia: The PPSMI policy and the washback effect of bilingual high-stakes secondary school exit exams.* Unpublished PhD dissertation, McGill University.

Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research, 15*(3), 325-342.

Tashakkori, A. &Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social & behavioral research.* Thousand Oaks, CA: Sage.

Teddlie, C. & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative techniques in the social and behavioral sciences.* Thousand Oaks, CA: Sage.

Toulmin, S.E. (2003). *The uses of argument.* Cambridge, UK: Cambridge

Toulmin, S.E. (2003). *The uses of argument (2nd ed.).* Cambridge, UK: Cambridge University Press.

Tsung, L & Cruickshank, K (2011). *Teaching and Learning Chinese in Global Context*. London, UK: Continuum.

Turner, C.E. (2001). The need for impact studies of L 2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle. In A. Brown et al. (Eds.), *Experimenting with uncertainty: Language testing essays in honour of Alan Davies* (pp.138-149). Cambridge, UK: Cambridge University Press.

Turner, C.E. (2006). Professionalism and high-stakes tests: Teacher perspectives when dealing with educational change introduced through provincial exams. *TESL Canada Journal*, *23*(2), 54-76.

Turner, C.E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal of Pedagogies and Learning*, *5*(1), 103-123.

Turner, C.E. (2012). Classroom assessment. In G. Fulcher & F. Davidson (Eds.), *Routledge Handbook of Language Testing* (pp. 65-78). New York: Routledge, Taylor & Francis Group.

Turner, C. E. (2013). Mixed Methods Research. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1403-1417). Chichester, UK: John Wiley & Sons Ltd

Turner, C.E. & Purpura, J.E. (2015). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp.255-272). Berlin, Germany/Boston, MA: DeGruyter Mouton.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, *13*(3), 334-354.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson (Eds.), *Language Testing and Assessment, Encyclopedia of Language and Education* (Vol. 7, pp. 291-302). Dordrecht, UK: Kluwer.

Wall, D. (1999). *The impact of high-stakes examinations on classroom teaching: A case using insights from testing and innovation theory.* Unpublished doctoral dissertation, University of Lancaster, England.

Wall, D. & Alderson, J.C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41-69.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. (2012). Review of Pearson test of English academic: Building an assessment use argument. *Language Testing, 29*, 603-619.

Wang, S. (2016). Exploring the Chinese proficiency test, Hanyu Shuiping Kaoshi and its washback effects: The test-takers' perspective. In Docherty & Barker. (Eds) *Studies in Language Testing: Language Assessment for Multilingualism*, pp.433-453. Cambridge: UCLES/Cambridge University Press.

Wang, H. (2010). *Investigating the justifiability of an additional test use: An application of assessment use argument to an English as a foreign language test*. Unpublished doctoral dissertation. University of California, Los Angeles.

Wang, J. (2010). *A Study of the role of the "teacher factor" in washback.* Unpublished PhD dissertation, McGill University.

Wang, S. (2013). *Exploring the Washback of a Large-scale high-stakes Chinese Test, the HanyuShuipingKaoshi, on Learner Factors*. Unpublished Master Thesis, McGill University.

Watanabe, Y. (1996). Investigating washback in Japanese EFL classrooms: Problems of methodology. In G. Wigglesworth & C. Elder (Eds.), The language testing circle: From inception to washback, pp. 208-239. Melbourne, AU: Applied Linguistics Association of Australia.

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods*, pp.19-36. Mahwah, NJ: Lawrence Erlbaum.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170.

Xie, Q. (2010). *Test design and use, preparation, and performance: A structural equation modeling study of consequential validity*. Unpublished doctoral dissertation, the University of Hong Kong, Hong Kong.

Xie, Q. & S. Andrews. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing, 30*(1), 49-70.

Xie, X. (2011). 为什么要开发新 HSK 考试? [Why should a new HSK be launched?]. *China Examinations, 3*, 10-13.

Yang, C. & Liao, F. (2000). HSK 的反拨效应对泰汉教学中的个案研究[A case study: the washback effect of the HSK on teaching Chinese to Thai speakers], *Capital Education Journal, 11*(4).

Young, J. (2008). Ensuring Valid Content Tests for English Language Learners. R& D Connections. Princeton, NJ: Educational Testing Service. Available online: https://www.ets.org/Media/Research/pdf/RD_Connections8.pdf

Zhang, J., Xie, N., Wang, S., Li, Y., & Li, T. (2010). The Report of Researching and Producing the New HSK. *China Examinations, 9*, 42-48.

Zhang, Q., & Zhang, J. (2014). 新汉语水平考试为什么取消合格线？[Why did the new HSK cancel passing score?] *China Examinations, 9*, 14-17.

Appendix

Appendix 1 The new HSK test structure and tasks

Level	Section	Part	Item	Question Type	Detailed Description
1	**Listening**	1	5	True or False	每题都是一个短语,试卷上提供一张图片, 考生根据听到的内容判断对错。 Each question contains an audio clip of a short phrase. Students must use the listening to determine whether the image presented in the examination booklet is true or false.
		2	5	Multiple choice	每题都是一个句子,试卷上提供 3 张图片, 考生根据听到的内容选出对应的图片。 Each question contains an audio clip of a sentence. Students must use the listening to select the corresponding image from a set of three found in the examination booklet.
		3	5	Multiple choice	每题都是一个对话,试卷上提供几张图片, 考生根据听到的内容选出对应的图片。 Each question contains an audio clip of a conversation. Based on the conversation they hear, students must select the corresponding image.

		4	5	Multiple choice	每题都是一个人说一句话,第二个人根据这句话问一个问题并说出 3 个选项,试卷上每题都有 3 个选项,考生根据听到的内容选出答案。 In each listening, Speaker A will say a sentence, while Speaker B will ask a question based on Speaker A's utterance. Speaker B will also provide three possible responses to the question, which are also indicated on the exam booklet. Students must then select the correct answer based on the information given by Speaker A.
	Reading	1	5	True or False	每题提供一张图片和一个词语,考生要判断是否一致。 Students must determine whether the image and the phrase in each question correspond to each other.
		2	5	Multiple choice	试卷上有几张图片,每题提供一个句子,考生根据句子内容,选出对应的图片。 Each question contains a set of images and a sentence. Students must select the image that corresponds to the sentence.
		3	5	Matching	提供 5 个问句和 5 个回答,考生要找出对应关系。 Students must correctly match each of the five listed questions to one of the five possible answers.
		4	5	Multiple choice	每题提供一个句子,句子中有一个空格,考生要从提供的选项中选词填空。 Each question contains a sentence with a blank. Students must then fill in the blank with one of the listed options.

2	**Listening**	1	10	True or False	每题都是一个句子,试卷上提供一张图片,考生根据听到的内容判断对错。 Based on the sentence they hear, students must determine whether the image provided in the question is true or false.
		2	10	Multiple choice	每题都是一个对话,试卷上提供几张图片,考生根据听到的内容选出对应的图片。 Based on the conversation they hear, students must select the corresponding image from the provided selection.
		3	10	Multiple choice	每题都是两个人的两句对话,第三个人根据对话问一个问题,试卷上提供 3 个选项,考生根据听到的内容选出答案。 In each question, students will hear a 2-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the three possible responses. Each recording is played twice.
		4	5	Multiple choice	每题都是两个人的 4 到 5 句对话,第三个人根据对话问一个问题,试卷上提供 3 个选项,考生根据听到的内容选出答案。 In each question, students will hear a 4- to 5-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the three possible responses. Each recording is played twice.

	Reading	1	5	Multiple choice	试卷上有几张图片,每题提供一个句子,考生根据句子内容,选出对应的图片。 Each question contains a set of images and a sentence. Students must select the image that corresponds to the sentence.
		2	5	Multiple choice	每题提供一到两个句子,句子中有一个空格,考生要从提供的选项中选词填空。 Each question contains one or two sentences with a blank. Students must then fill in the blank with one of the listed options.
		3	5	True or False	每题提供两个句子,考生要判断第二句内容与第一句是否一致。 Each question contains two sentences. Students must determine whether the information in the second sentence is the same as that in the first sentence.
		4	10	Matching	提供 20 个句子,考生要找出对应关系。 Students must determine the logical (matching) relationship between a list of 20 sentences.
3	**Listening**	1	10	Multiple choice	每题都是一个对话,试卷上提供几张图片,考生根据听到的内容选出对应的图片。 Each question contains an audio clip of a conversation. Students must then select the corresponding images from the selection provided on the examination booklet.
		2	10	True or False	每题都是一个人先说一小段话,另一人根据这段话说一个句子,试卷上也提供这个句子,要求考生判断对错。

					Each question contains a recording of two speakers, where Speaker A will first speak briefly on a topic and Speaker B will respond with a sentence. This sentence is also written in the exam booklet, so that students can determine if it is true or false.
		3	10	Multiple choice	每题都是两个人的两句对话,第三个人根据对话问一个问题,试卷上提供 3 个选项,考生根据听到的内容选出答案。 In each question, students will hear a 2-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the three possible responses.
		4	10	Multiple choice	每题都是两个人的 4 到 5 句对话,第三个人根据对话问一个问题,试卷上提供 3 个选项,考生根据听到的内容选出答案。 In each question, students will hear a 4- to 5-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the three possible responses.
	Reading	1	10	Matching	提供 20 个句子,考生要找出对应关系。 Students must determine the logical relationship between a list of 20 sentences.
		2	10	Multiple choice	每题提供一到两个句子,句子中有一个空格,考生要从提供的选项中选词填空。

					Each question contains one or two sentences with a blank. Students must then fill in the blank with one of the listed options.
		3	10	Multiple choice	提供 10 小段文字,每段文字带一个问题,考生要从 3 个选项中选出答案。 Students must read a short text and then respond to the question found at the end of the text by using one of the three possible choices. There is a total of 10 texts.
	Writing	1	5	Completing sentences	每题提供几个词语,要求考生用这几个词语写一个句子。 Each question contains several words, which students must use to write a sentence.
		2	5	Filling in the blanks	每题提供一个带空格的句子,要求考生在空格上写正确的汉字。 Each question contains a phrase with a blank. Students must fill in the blank by writing down the correct character.
4	**Listening**	1	10	True or False	每题都是一个人先说一小段话,另一人 根据这段话说一个句子,试卷上也提供这个句子,要求考生判断对错。 Each question contains a recording of two speakers, where Speaker A will first speak briefly on a topic and Speaker B will respond with a sentence. This sentence is also written in the exam booklet, so that students can determine if it is true or false.
		2	15	Multiple choice	每题都是两个人的两句对话,第三个人 根据对话问一个问题,试卷上提供 4 个选项,考生根据听到的内容选出答案。

					In each question, students will hear a 2-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the four possible responses.
		3	20	Multiple choice	这部分试题都是 4 到 5 句对话或一小段 话,根据对话或语段问一到两个问题,试卷上每题提供 4 个选项,考生根据听到 的内容选出答案。 In each question, students will hear a 4- to 5-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the four possible responses.
	Reading	1	10	Multiple choice	每题提供一到两个句子,句子中有一个空格,考生要从提供的选项中选词填空 Each question contains one or two sentences with a blank. Students must then fill in the blank with one of the listed options.
		2	10	Rearrange the orders	每题提供 3 个句子,考生要把这 3 个句子按顺序排列起来。 Each question contains three sentences, which the students must reorder into the correct sequence.
		3	20	Multiple choice	每段文字带一到两个问题,考生要从 4 个选项中选出答案。

					At each section of the text, students are given one or two sentences. Students must answer these questions by using one of the four choices.
	Writing	1	10	Completing sentences	每题提供几个词语,要求考生用这几个词语写一个句子。 Each question contains several words, which students must use to write a sentence.
		2	5	Making sentences	每题提供一张图片和一个词语,要求考生结合图片用这个词语写一个句子。 Each question contains an image and a phrase. Students must use the image to help them write a sentence containing the phrase.
5	**Listening**	1	20	Multiple choice	每题都是两个人的两句对话,第三个人 根据对话问一个问题,试卷上提供 4 个选项,考生根据听到的内容选出答案。 In each question, students will hear a 2-line conversation between Speaker A and Speaker B. Speaker C will then ask a question based on this short conversation. By using their understanding of the conversation, students must answer this question with one of the four possible responses.
		2	25	Multiple choice	一段对话或一段话, 根据对话或语段问一个或几个问题,试卷上每题提供 4 个选项,考生根据听到的 内容选出答案。 After listening to either a conversation or a monologue, students must answer the listed question(s) with one of the four listed options.

	Reading	1	15	Multiple choice	每篇文字中有几个空格,空格中应填入一个词语或一个句子,每个空格有 4 个选项,考生要从中选出答案。 Each text contains several blanks. Students must fill each blank with either a word or a phrase from a selection of four possible choices.
		2	10	Multiple choice	每题提供一段文字和 4 个选项,考生要选出与这段文字内容一致的一项。 Each question contains a text and four choices. Students must select the option that is in agreement with the text.
		3	20	Multiple choice	每篇文字带几个问题,考生要从 4 个选项中选出答案。 After each text, students must answer the listed questions. For each question, students can select their answers from a list of four possible choices.
	Writing	1	8	Rearranging order of given words	每题提供几个词语,要求考生用这几个词语写一个句子。 Each question contains several words, which students must use to write a sentence.
		2	2	Short essay	第一题提供几个词语,要求考生用这几个词语写一篇 80 字左右的短文;第二题提供一张图片,要求考生结合图片写一篇 80 字左右的短 文。 Question 1: Students are given several phrases, which they must use to create a short text of approximately 80 words.

					Question 2: Students must write a short text of approximately 80 words based on the image presented.
quw	**Listening**	1	15	Multiple choice	每题播放一小段话,试卷上提供 4 个选 项,考生根据听到的内容选出与其一致的一项。 Each question contains an audio clip of a short monologue. From a set of four possible choices, students must select the one that is in agreement with the listening.
		2	15	Multiple choice	播放三段采访,每段采访后带 5 个试题, 试卷上每题提供 4 个选项,考生根据听到的内容选出答案。 Students will listen to three interviews. After each interview, students must answer five questions. Each question will have four possible choices; students must select the one that best reflects the information they heard in the recording.
		3	20	Multiple choice	播放若干段话,每段话后带几个问题, 试卷上每题提供 4 个选项,考生根据听到的内容选出答案。 After listening to a longer monologue, students must respond to a series of questions based on the information they just heard. For each question, students must select one of the four possible choices.
	Reading	1	10	Multiple choice	每题提供 4 个句子,要求考生选出有语病的一句。 Each question contains four sentences. Students must select the sentence with a language error.

		2	10	Multiple choice	每题提供一小段文字,其中有 3 到 5 个空格,考生要结合语境,从 4 个选项中选出最恰当的答案。 Each question contains a short reading with three to five blanks. Students must use the context to select the most appropriate response from a list of four.
		3	10	Multiple choice	提供两篇文字,每篇文字有 5 个空格,考生要结合语 境,从提供的 5 个句子选项中选出答案 This section contains two texts. Each text contains five blanks. To fill in the blank, students must use the context to select the most appropriate sentence from a list of five.
		4	20	Multiple choice	提供若干篇文字,每篇文字带几个问题,考生要从 4 个选项中选出答案。 Several texts are given to students. At the end of each text, students must answer the listed questions by selecting one of four possible choices.
	Writing	1	1	Essay (condensation)	先要阅读一篇 1000 字左右的叙事文章,时间为 10 分钟;然后将这篇文 章缩写为一篇 400 字左右的短文,时间为 35 分钟。标题自拟。只需复述文章内 容,不需加入自己的观点。 After reading a text of approximately 1000 characters in 10 minutes, students must summarize it into a shorter text of approximately 400 characters. Students have 35 minutes to write their summary; they do not need to add new information.

Appendix 2

Questionnaire for HSK Test-taker Participants (English Version)

The purpose of this questionnaire is to investigate your perceptions on the HSK, CSL/CFL/CHL learning, and test preparation. Please fill in this questionnaire based upon your own experience. Any information you provide will be held in the strictest confidence and used solely for research purpose.

I. Your Background Information. Please indicate your answer with a checkmark (✓) where appropriate.

1. Your gender: [ ] female [ ] male
2. Your age [ ]<18 [ ] 18-22 [ ]23-30 [ ]>30
3. Your nationality: _____
4. What is your first language(s)? _____
5. How many years have you been learning Chinese?
   [ ] 0-1 [ ] 1-2 [ ] 2-3 [ ] 3-5 [ ]5+
6. What degree you are pursuing now?
   [ ] Secondary school
   [ ] Bachelor
   [ ] Master
   [ ] PhD.
   Please specify your major: _____
7. What types of Chinese courses have you taken?
   [ ] University credit course (please specify your major: _____)
   [ ] Weekend language school
   [ ] HSK preparation course
   [ ] Confucius Institute course
   [ ] Other (please specify: _____)
8. What is your purpose in learning Chinese:
   [ ] You are interested in learning new languages.
   [ ] You would like to study/work/travel in China or related to China.
   [ ] It is required by my academic program/professional program.
   [ ] It is encouraged by my parents or friends.
   [ ] Other (please specify: _____)
9. Which HSK level have you taken most recently?
   [ ] Level 1 [ ] Level 2 [ ] Level 3 [ ] Level 4 [ ] Level 5 [ ] Level 6
   Please specify your score: [ ] 0-180 [ ]181-210 [ ] 211-240 [ ] 241-270 [ ] 271-300
   Please specify your score at each section: Listening ___ Reading ___ Writing (if applicable) ___
10. Which HSKK level have you taken?
    [ ] Elementary [ ] Intermediate [ ] Advanced [ ] None of above
    Please specify your score: [ ] 0-60 [ ]61-70 [ ] 71-80 [ ] 81-90 [ ] 91-100

II. This section includes statements about your purposes for taking the HSK, test expectancy, values of the HSK, and other aspects of the HSK. Please circle ONE number to indicate

the extent to which you agree with each of the statements on a 6-point scale.
1= strongly disagree 2=disagree 3=somewhat disagree 4= somewhat agree 5= agree 6=strongly agree

Q1. I took the HSK mainly to	
P1. measure my Chinese proficiency.	1 2 3 4 5 6
P2. challenge myself and prove my Chinese proficiency.	1 2 3 4 5 6
P3. get the HSK certificate for entering an educational program.	1 2 3 4 5 6
P4. graduate with a degree.	1 2 3 4 5 6
P5. compete for academic scholarships.	1 2 3 4 5 6
P6. get the HSK certificate for fulfilling job requirements.	1 2 3 4 5 6
P7. obtain advantage in job seeking or promotion.	1 2 3 4 5 6
Other purposes (please specify)	

Q2. Regarding the difficulty level of the HSK, I think	
D1. the overall HSK is difficult (which means the specific HSK level you took is difficult).	1 2 3 4 5 6
D2. the listening subtest is difficult.	1 2 3 4 5 6
D3. the reading subtest is difficult.	1 2 3 4 5 6
D4. the writing subtest is difficult. (if applicable)	1 2 3 4 5 6
D5. the speaking subtest is difficult. (if applicable)	1 2 3 4 5 6
Other comments (please specify)	

Q3. Regarding my expectations of the HSK,	
TE1. considering the difficulty of the HSK and my own ability, I was confident that I would do well on the HSK.	1 2 3 4 5 6
TE2. if I prepare in appropriate ways, I would do well on the HSK.	1 2 3 4 5 6
TE3. if I fail, it must be that I do not work hard enough.	1 2 3 4 5 6
TE4. I set passing the HSK as a staggered goal in Chinese learning.	1 2 3 4 5 6
TE5. taking the HSK motivates me to work harder in Chinese learning.	1 2 3 4 5 6
TE6. passing the HSK is very important to me.	
Other comments (please specify)	

Q4. Regarding the values of the HSK , I think	
V1. the test level and score are an appropriate indicator of my overall Chinese ability.	1 2 3 4 5 6
V2. the test level and score are an appropriate indicator of my listening ability.	1 2 3 4 5 6
V3. the test level and score are an appropriate indicator of my reading ability.	1 2 3 4 5 6
V4. the test level and score are an appropriate indicator of my writing ability. (if applicable)	1 2 3 4 5 6
V5. the HSKK level and score are an appropriate indicator of my speaking ability. (if applicable)	1 2 3 4 5 6
V6. the score report provides useful feedback for my Chinese language learning.	1 2 3 4 5 6
V7. obtaining a HSK certificate will enhance my competitiveness in future studies.	1 2 3 4 5 6
V8. obtaining a HSK certificate will enhance my competitiveness in scholarship applications.	1 2 3 4 5 6

V9. obtaining a HSK certificate will enhance my competitiveness in future job markets.	1 2 3 4 5 6
Other values (please specify)	

Q5. I think the effects of the HSK

E1. motivate students to enhance their proficiency in Chinese.	1 2 3 4 5 6
E2. motivate teachers to improve their teaching.	1 2 3 4 5 6
E3. encourage memorization of vocabulary and language rules.	1 2 3 4 5 6
E4. encourage students to use Chinese in daily life.	1 2 3 4 5 6
E5. force students to study to the test.	1 2 3 4 5 6
E6. force teachers to teach to the test.	1 2 3 4 5 6
Other comments (please specify)	

Q6. Regarding the format, content and other aspects of the HSK, I think

A1. the overall format of the HSK is satisfactory.	1 2 3 4 5 6
A2. the instructions for the test are clear.	1 2 3 4 5 6
A3. the HSKK should be included in the HSK.	1 2 3 4 5 6
A4. if the HSKK were compulsory, I would spend more time and effort cultivating my speaking ability.	1 2 3 4 5 6
A5. the inclusion of Chinese input system weakens the ability of Chinese characters writing.	1 2 3 4 5 6
A6. the overall content of the HSK is satisfactory.	1 2 3 4 5 6
A7. the test content focuses more on communicative functions of the language than linguistic knowledge.	1 2 3 4 5 6
A8. HSK's tasks don't include contents that offend or favor test takers.	1 2 3 4 5 6
A9. the HSK reflects the goals and objectives of the test specifications.	1 2 3 4 5 6
A10. the HSK is fair for all test-takers throughout the whole process.	1 2 3 4 5 6
A11. the score report provides relevant and sufficient information to make decisions.	1 2 3 4 5 6
A12. the development of the new HSK is related to the Promoting Chinese Internationally (PCI) movement.	1 2 3 4 5 6
Other comments (please specify)	

Q7. Regarding the impact of the HSK as a prerequisite/or an exit requirement in educational and social context, I think the requirement

R1. can enhance learners' Chinese proficiency.	1 2 3 4 5 6
R2. can increase the amount of energy/money learners allocate to Chinese learning.	1 2 3 4 5 6
R3. can ensure the proficiency level of learners.	1 2 3 4 5 6
R4. can place extra work or pressure on learners in order to pass the test.	1 2 3 4 5 6
R5. can make the learners' main goal in learning Chinese to pass the test.	1 2 3 4 5 6
R6. can encourage learners to learn test-taking strategies but not really learn the ability to communicate in Chinese.	1 2 3 4 5 6
Other comments (please specify)	

III. This section includes statements about your HSK preparation practice. Please circle ONE number to indicate the extent to which you agree with each of the statements on a 6-point scale.

1= strongly disagree 2=disagree 3=somewhat disagree 4= somewhat agree 5= agree 6=strongly agree

TP1. I spend more time on my weak points.	1 2 3 4 5 6
TP2. I analyze HSK papers to identify the question types.	1 2 3 4 5 6
TP3. I analyze HSK score distribution to judge the relative importance of sections.	1 2 3 4 5 6
TP4. I memorize HSK vocabulary required in the test specification document.	1 2 3 4 5 6
TP5. I pay more attention to the differentiation synonym.	1 2 3 4 5 6
TP6. I review HSK grammar points required in the test specification document.	1 2 3 4 5 6
TP7. I seek teachers' advice on how to improve my HSK performance.	1 2 3 4 5 6
TP8. I prefer teachers to cover more HSK related content in classes.	1 2 3 4 5 6
TP9. I take HSK prep courses or hire HSK private tutors.	1 2 3 4 5 6
TP10. I communicate with Chinese native speakers whenever possible.	1 2 3 4 5 6
TP11. I practice HSKK topics.	1 2 3 4 5 6
TP12. I watch Chinese TV and/or listen to Chinese radio broadcasts.	1 2 3 4 5 6
TP13. I repeatedly listen to the listening section in past HSK test papers and mock test papers.	1 2 3 4 5 6
TP14. During listening, I go over the options beforehand so as to focus my attention accordingly in listening.	1 2 3 4 5 6
TP15. During listening, I try to write down important information.	1 2 3 4 5 6
TP16. I try to fully understand all the HSK listening materials I've practiced.	1 2 3 4 5 6
TP17. I read Chinese textbooks aloud.	1 2 3 4 5 6
TP18. I read Chinese newspapers/magazines/websites.	1 2 3 4 5 6
TP19. I practice reading sections in HSK test papers and mock test papers.	1 2 3 4 5 6
TP20. I focus on understanding difficult and complex sentences in the passages I read.	1 2 3 4 5 6
TP21. I read questions before looking for key words and sentences in the passage while practicing reading.	1 2 3 4 5 6
TP22. I practice selecting answers out of the options by elimination strategies.	1 2 3 4 5 6
TP23. I write emails/diaries/blogs in Chinese.	1 2 3 4 5 6
TP24. I practice writing sections in past HSK test papers and mock test papers.	1 2 3 4 5 6
TP25. I try to avoid grammar and writing mistakes while practicing writing.	1 2 3 4 5 6
TP26. I try to use more advanced vocabulary and structures.	1 2 3 4 5 6
Other comments (please specify)	

Would you be willing to participate in a one-on-one interview? Yes [ ] No [ ]

If YES, please leave your contact information: _____

Thank you very much for your time!

Appendix 3

Questionnaire for Teacher Participants (English Version)

The purpose of this questionnaire is to investigate your perceptions of teaching, learning, and testing. Please fill in this questionnaire based upon your own experience. Any information you provide will be held in the strictest confidence and used solely for research purpose.

I. Your Background Information. Please indicate your answer with a checkmark(✓) where appropriate. ***(Check off all the answers that apply.)***

11. Your gender: [ ] female [ ] male
12. Your age: [ ] 21-29 [ ] 30-39 [ ] 40-49 [ ] 50-59 [ ] over 60
13. Your nationality: _____
14. What is your first language(s)? _____
15. How many years have you been teaching Chinese?
    [ ] 1-5 [ ] 6-10 [ ] 11-15 [ ] 16-20 [ ]21+
16. Your academic background: [ ] Bachelor [ ] M.A./M.Ed. [ ] PhD. Please specify your major: _____
17. What types of Chinese courses have you taught?
    [ ] University credit course
    [ ] University CSL course
    [ ] HSK preparation course
    [ ] Confucius Institute course
    [ ] Other (please specify: _____)
18. The main goal of your teaching is to help students:
    [ ] succeed on the tests (e.g., final class exams and HSK)
    [ ] acquire language proficiency
    [ ] accumulate knowledge of grammar and vocabulary
    [ ] enhance their communication skill
    [ ] Other (please specify: _____)
19. What is your current status as a Chinese language teacher?
    [ ] Pre-service teacher
    [ ] In-service teacher in Chinese as a Second Language context
    [ ] In-service teacher in Chinese as a Foreign Language context
    [ ] In-service teacher in Chinese as a Heritage Language context
    [ ] Other (please specify: _____)

II. This section includes statements about your perceptions of teaching, learning, and testing. Please circle ONE number to indicate the extent to which you agree with each of the statements on a 6-point scale.

1= strongly disagree 2=disagree 3= somewhat disagree 4= somewhat agree 5= agree 6= strongly agree, or 0 = I am not familiar with it.

Q1. When I teach CSL/CFL/CHL classes, I

TM1. use Communicative Language Teaching (CLT) methods in my instruction.	1 2 3 4 5 6
TM2. use the traditional structural approach method (e.g., grammar-translation method) in my instruction.	1 2 3 4 5 6
TM3. use a combined approach of CLT and the structural approach.	1 2 3 4 5 6
TM4. focus on fostering students' language use ability.	1 2 3 4 5 6
TM5. focus on teaching students' linguistic knowledge.	
TM6. have my students practice with mock tests to prepare for the HSK.	1 2 3 4 5 6
TM7. involve HSK test questions in homework, mid-terms, or final exams.	1 2 3 4 5 6
TM8. use HSK-related textbooks.	1 2 3 4 5 6
TM9. always encourage my students to participate in the HSK.	1 2 3 4 5 6
Other methods (please specify)	

Q2. I think my teaching practice

TP1. is an effective foreign language teaching method.	1 2 3 4 5 6
TP2. helps foster student comprehensive skills in Chinese.	1 2 3 4 5 6
TP3. reflects the goals and objectives of the *Scales* and the *Standards.*	1 2 3 4 5 6
TP4. is the most appropriate method for helping students pass the HSK.	1 2 3 4 5 6
TP5. meets the students' expectations for test preparation.	1 2 3 4 5 6
Other comments (please specify)	

Q3. I think the new HSK is intended to

TH1. measure the ability of linguistic knowledge.	1 2 3 4 5 6
TH2. measure the ability of language use.	1 2 3 4 5 6
TH3. provide a reference for decision-making concerning recruiting students	1 2 3 4 5 6

TH4. provide a reference for decision-making concerning assigning students to different classes.	1 2 3 4 5 6
TH5. provide a method for my institution to evaluate our teaching results	1 2 3 4 5 6
TH6. provide a method for students to assess and improve their Chinese proficiency.	1 2 3 4 5 6
Other comments (please specify)	

Q4. Regarding the values of the HSK , I think

TV1. the test level and score are an appropriate indicator of a student's overall Chinese ability.	1 2 3 4 5 6
TV2. it provides useful feedback to students' Chinese language learning.	1 2 3 4 5 6
TV3. it provides useful feedback to teachers' Chinese language teaching.	1 2 3 4 5 6
TV4. obtaining a HSK certificate will enhance a student's competitiveness in future studies.	1 2 3 4 5 6
TV5. obtaining a HSK certificate will enhance a student's competitiveness in scholarship applications.	1 2 3 4 5 6
TV6. obtaining a HSK certificate will enhance a student's competitiveness in future job markets.	1 2 3 4 5 6
Other values (please specify)	

Q5. I think the effects of the HSK

TE1. motivate teachers to improve their methodology in teaching Chinese.	1 2 3 4 5 6
TE2. motivate students to enhance their proficiency in Chinese.	1 2 3 4 5 6
TE3. encourage memorization of vocabulary and language rules.	1 2 3 4 5 6
TE4. encourage the use of advanced teaching methodologies.	1 2 3 4 5 6
TE5. force students to study to the test.	1 2 3 4 5 6
TE6. force teachers to teach to the test.	1 2 3 4 5 6
TE7. changed my instructional focus from linguistic knowledge to language	1 2 3 4 5 6

use.
Other comments (please specify)

Q6. Regarding the content, format, and other aspects of the HSK, I think

TA1. the overall content of the HSK is satisfactory.	1 2 3 4 5 6
TA2. the overall format of the HSK is satisfactory	1 2 3 4 5 6
TA3. the test content focuses more on communicative functions of the language than linguistic knowledge.	1 2 3 4 5 6
TA4. the HSKK should be included in the HSK	1 2 3 4 5 6
TA5. if the HSKK is compulsory, I would spend more time and efforts cultivating students' speaking ability.	1 2 3 4 5 6
TA6. the inclusion of Chinese input system weakens the ability of Chinese characters writing.	1 2 3 4 5 6
TA7. HSK's tasks don't include the content that offends or favors test takers.	1 2 3 4 5 6
TA8. the HSK reflects the goals and objectives of the *Scales*.	1 2 3 4 5 6
TA9. the HSK's difficulty level is appropriate.	1 2 3 4 5 6
TA10. the HSK is fair for all test-takers throughout the whole process.	1 2 3 4 5 6
TA11. the score report provides relevant and sufficient information to make decisions.	1 2 3 4 5 6
TA12. the development of the HSK is related to the Promoting Chinese Internationally (PCI) movement.	1 2 3 4 5 6
Other comments (please specify)	

Q7. Regarding the impact of the HSK as a prerequisite/or an exit requirement in higher education programs, I think it

TR1. can enhance students' Chinese proficiency.	1 2 3 4 5 6
TR2. can increase the amount of energy/money students allocate to Chinese learning.	1 2 3 4 5 6

TR3. can ensure the proficiency level of learners.	1 2 3 4 5 6
TR4. can place extra work or pressure on learners in order to pass the test.	1 2 3 4 5 6
TR5. can make the learners' main goal in learning Chinese to pass the test.	1 2 3 4 5 6
TR6. can encourage learners to learn test-taking strategies but not really learn the ability to communicate in Chinese	1 2 3 4 5 6

Other comments (please specify)

Would you be willing to participate in a one-on-one interview? Yes [ ] No [ ]
If YES, please leave your contact information: _____

Thank you very much for your time

Appendix 4

CLASSROOM OBSERVATION SCHEDULE

Date: _____ Time of Observation: _____ Name of School: _____
Class: _____ Name of Textbook: _____ Unit of Textbook: _____
Observer: _____ Number of students: _____
Description of classroom:-

_____
_____

Event No.	Time	Activity	Topic	Organization Patten	Focus of instruction	Materials	Language	Skill Area	Relevance to the test	Other information

Notes: 1) Classroom Organization Patterns: percentage of class time spent on student-centered activities (e.g., pair-work, group work, individual work, role-play); percentage of class time spent on teacher-centered activities (e.g., teacher lecturing to the whole class without interactions with students– teacher presentations, explanations of sentences, reading aloud, translations, etc.).

2) Focus of Instruction: frequency of explaining language points with a focus on language forms (e.g., explanation of sentence structures, rote practice and mechanical grammar exercises; explanation of vocabulary in a decontextualized manner); frequency of involving students in meaning-based activities (e.g., discussion, role-play, comprehension exercises at the discourse-level, etc.).

3) Relevance to the Test: percentage of class time spent on aural/oral aspects of Chinese (e.g., listening practice, oral practice at the discourse level encouraged by the Scales) as well as on fast reading practice (effected by the HSK); frequency of giving information or advice about the HSK (old/new) or test-taking strategies.

4) Medium of Instruction: English/Chinese/half English/half Chinese

5) Teaching Materials: textbooks, test-related materials (e.g., the old HSK papers or simulated test papers), audio or audio-visual materials, or other supplementary teaching materials.

Appendix 5

Questionnaire for Test User Participants in Universities

The purpose of this questionnaire is to investigate your perceptions of the HSK test in the admission process. Please fill in this questionnaire based upon your own experience. Any information you provide will be held in the strictest confidence and used solely for research purpose.

I. Personal and institution information.

1. What is the best description of your job?

[ ] Admissions committee officer
[ ] Administrator
[ ] Other (please specify)_____

2. How many years have you been working in this position?

[ ] Less than 1 year
[ ] 1-5 years
[ ] 6-10 years
[ ] 11-15 years
[ ] 16+ years

3. What is your present role in the admissions process of your institution?

[ ] I answer questions from applicants and/or potential applicants about admissions issues by email/ by telephone/ in person.
[ ] I compile admissions information (including the language test scores) to aid in admissions decisions.
[ ] I read admissions files in order to make admissions decisions
[ ] I decide on the cutoff (minimum) scores of language tests for admissions
[ ] I inform students about admissions decisions (by writing letters, sending emails, or inputting decisions on an online platform which students can access)
[ ] I do none of these activities. (If checked, then survey cannot be completed).
[ ] Other, please specify:_____

4. Who else at your institution is involved in making admission decisions (e.g., read admission files, decide on cutoff scores)? How many of them?

_____
_____
_____

5. Name all the language tests (e.g., HSK, HSKK, institutional placement test) that you know are accepted for undergraduate or graduate admissions purposes for international students at your institution.

_____
_____
_____

6. What are cutoff scores (minimum scores for admissions) for each of the tests you accept? (If the cutoff scores vary by program, please mention each one separately)

_____
_____

_____

7. Please explain the reason why these particular cutoff scores on the HSK have been chosen at your institution.

_____
_____
_____

8. How strictly do you apply the cutoff scores for language tests in degree education admissions decisions? Check all the boxes which apply.

[ ] The cutoff scores are not applied. We make the admissions decisions on other criteria.
[ ] The cutoff scores are not always applied strictly, because other admissions criteria are sometimes judged to be more important than scores on the language tests.
[ ] We usually respect cutoff scores, but we make occasional exceptions when the rest of a student's application is very strong.
[ ] We always apply cutoff scores. We never accept students into our programs if their language test scores are below the cutoff.
[ ] I am not sure what happens.
[ ] Other, please comment if none of the above situations fits your case.

9. How many international students (degree education) are there in your institution?

[ ] 0-100
[ ] 101-300
[ ] 301-500
[ ] 501-1000
[ ] 1001+
[ ] I don't know

10. What is the acceptance rate?

[ ] 0-20%
[ ] 21%-40%
[ ] 41%-60%
[ ] 61%-80%
[ ] 81%-100%
[ ] I don't know

II. HSK score interpretations, uses and consequences

Please circle ONE number to indicate the extent to which you agree with each of the statements on a 6-point scale in Q1 and Q2.

1= strongly disagree 2=disagree 3=somewhat disagree 4= somewhat agree 5= agree 6=strongly agree

Q1. I think

1. the HSK provides an accurate measure of test-takers' overall Chinese proficiency.	1 2 3 4 5 6
2. the HSK is fair for all test-takers during the whole procedure of the test.	1 2 3 4 5 6

3. the score-based interpretation provides relevant and sufficient information to make decisions.	1 2 3 4 5 6
4. the development of the new HSK is related to the PCI movement.	1 2 3 4 5 6
5. the HSK is trustworthy in terms of its validity and reliability.	1 2 3 4 5 6
6. the HSK is widely recognized by universities, companies, and organizations in the recruitment process.	1 2 3 4 5 6
7. an applicant with a high level/score HSK certificate would use Chinese more proficiently than those who have a low level.	1 2 3 4 5 6
8. an applicant with a HSK certificate would use Chinese more proficiently than those who do not have a HSK certificate.	1 2 3 4 5 6
9. an applicant with a high level/score HSK certificate would have more opportunities to apply for scholarships at your institution.	1 2 3 4 5 6
10. an applicant with a high level/score HSK certificate would have more opportunities to achieve a high GPA at your institution.	1 2 3 4 5 6
Other comments (please specify)	

Q2. I think making the HSK a requirement in educational contexts

1. can enhance learners' Chinese proficiency.	1 2 3 4 5 6
2. can increase the amount of energy/money learners allocate to Chinese learning.	1 2 3 4 5 6
3. can ensure the proficiency level of learners.	1 2 3 4 5 6
4. can place extra work or pressure on learners in order to pass the test.	1 2 3 4 5 6
5. can make the learners' main goal in learning Chinese to pass the test.	1 2 3 4 5 6
6. can encourage learners to learn test-taking strategies but not really learn the ability to communicate in Chinese.	1 2 3 4 5 6
7. can encourage teachers to teach to the test.	1 2 3 4 5 6
Other comments (please specify)	

Would you be willing to participate in a one-on-one interview? Yes [ ] No [ ]
If YES, please leave your contact information: _____

Thank you very much for your time!

Appendix 6

Questionnaire for test user participants (in companies, governments, and organizations)

The purpose of this questionnaire is to investigate your perceptions of the HSK test in the employment process. Any information you provide will be held in the strictest confidence and used solely for research purposes.

Personal information.

1. What does your company do? How long has your company been in operation? How many employees are there in your company?
2. What is your position?
3. What is your role in the recruitment process?

Chinese test certificates in the recruitment/promoting process

4. Are there any sectors in your company that require/encourage employees with a certain level of Chinese proficiency?
5. There are a number of Chinese proficiency tests such as the HSK, BCT, TOCFL. Which ones are you the most familiar with? Which tests would you recommend your employees or prospective employees take?
6. What do you consider to be important attributes when recruiting new employees? For example, their skills, personality, Chinese proficiency, etc.
7. How do you evaluate your employees' Chinese language levels in terms of ability to do work that involves Chinese? Do you require any test certificates? What are you looking for when you require a certain level of the HSK certificate or other test certificates?
8. What do you think about an applicant/employee holding a Chinese proficiency certificate? Is it an advantage? Does it mean his/her Chinese proficiency level is higher than others who don't hold one? Do these certificates demonstrate Chinese abilities at work? Or do these certificates have other meanings?
9. Are employees whose Chinese is better given more opportunities for promotion? Are employees who have been awarded Chinese proficiency test certificates given more opportunities for promotion?
10. Does your company provide incentives (benefits) to learn the Chinese language? Are there opportunities provided to learn Chinese at your company?

Thank you very much for your time!