Consequential Validity of the Chinese Proficiency Test (HSK) from Macro and Micro Perspectives

Consequential Validity of the Chinese Proficiency Test (HSK) from Macro and Micro Perspectives

By

Shujiao Wang

Cambridge
Scholars
Publishing

Consequential Validity of the Chinese Proficiency Test (HSK)
from Macro and Micro Perspectives

By Shujiao Wang

This book first published 2021

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-7248-X
ISBN (13): 978-1-5275-7248-5

TABLE OF CONTENTS

ABSTRACT viii

CHAPTER 1 1

SETTING THE SCENE

1.1 THE PROBLEM.. 1
1.2 PURPOSE OF THE STUDY 3
1.3 THE CONTEXT... 4
1.3.1 CSL Education and the "Promoting Chinese Internationally" Policy.. 4
1.3.2 The Development and Implementation of the HSK................ 5
1.4 SIGNIFICANCE OF THE STUDY 11
1.5 ORGANIZATION OF THE STUDY 13
1.6 KEY TERMS USED IN THIS RESEARCH 13

CHAPTER 2 16

ESTABLISHING FOUNDATIONS

2.1 INTRODUCTION 16
2.2 EXPLORING THE RESEARCH CONCEPT 16
2.2.1 The Initial Understanding of Validity.................................. 16
2.2.2 The Emerging Concept of Washback 17
2.2.3 Debates in Operationalizing Validity and Washback........... 18
2.2.4 The Nature of Washback and Factors Mediating its Process 19
2.2.5 Reconsidering Washback: From the View of Test Use......... 23
2.3 ARGUMENT-BASED VALIDATION THEORIES 29
2.3.1 Kane's Interpretative Argument... 29
2.3.2 Bachman's Assessment Use Argument................................ 31
2.4 THE VALIDITY AND CONSEQUENCE/IMPACT/WASHBACK STUDIES ON THE HSK 34
2.5 ARTICULATING AN AUA FRAMEWORK WITHIN THE HSK CONTEXT 38
2.6 SUMMARY... 41

CHAPTER 3 42
RESEARCH METHODOLOGY
3.1 OVERVIEW OF THE RESEARCH METHODOLOGIES EMPLOYED IN WASHBACK RESEARCH 42
3.2 THE DESIGN OF THE CURRENT MMR STUDY 47
3.2.1 Rational for this MMR Methodology 47
3.2.2 Mixed Methods Sequential Exploratory Design of the Current Study 49
3.3 SUMMARY 52

CHAPTER 4 53
PHASE 1 - INVESTIGATING THE TEST CONSEQUENTIAL VALIDITY BY USING AN ARGUMENT-BASED FRAMEWORK – A QUALITATIVE CASE STUDY OF THE HSK
4.1 INTRODUCTION 53
4.2 MAIN STUDY 53
4.2.1 Research Design 53
4.2.2 Findings 58
4.3 SUMMARY 68

CHAPTER 5 70
PHASE 2 - IMPROVING LANGUAGE TEACHING AND LEARNING THROUGH LANGUAGE TESTING: A WASHBACK STUDY ON THE HSK
5.1 INTRODUCTION 70
5.2 BACKGROUND 70
5.3 MAIN STUDY OF THE CLASSROOM CONTEXT 71
5.3.1 Research Design 71
5.3.2 Findings 79
5.3.3 Discussion 99
5.4 SUMMARY 102

CHAPTER 6 103
PHASE 3 - EXPLORING THE EDUCATIONAL AND SOCIAL CONSEQUENCES OF THE HSK FROM THE PROSPECTIVE OF TEST USERS – A MIXED METHODS STUDY
6.1 INTRODUCTION 103
6.2 BACKGROUND 103

6.3 MAIN STUDY OF THE EDUCATIONAL AND SOCIAL CONTEXT 104
6.3.1 Research Design *104*
6.3.2 Findings *110*
6.3.3 Discussion and Implications *119*
6.4 SUMMARY 125

CHAPTER 7 127
A COMPREHENSIVE DISCUSSION OF THE CONSEQUENTIAL VALIDITY ARGUMENT FOR THE HSK
7.1 INTRODUCTION 127
7.2 THE CONSEQUENTIAL VALIDITY ARGUMENT FOR THE HSK 127
7.2.1 Intended Consequences of the Test Developers *127*
7.2.2 Revisiting the AUA Framework of the HSK *129*
7.3 SUMMARY 140

CHAPTER 8 142
CONCLUSION
8.1 SUMMARY OF THE FINDINGS AND DISCUSSIONS 142
8.2 CONTRIBUTIONS OF THE MMR STUDY 144
8.3 IMPLICATIONS OF THE MMR STUDY 145
8.4 LESSONS LEARNED FROM THE MMR STUDY AND SUGGESTIONS FOR FUTURE RESEARCH 147

REFERENCES 149

APPENDICES 162

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/foreign language (CSL/CFL) teaching and learning. In Phase 1 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 analyzed 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. Additionally, the results 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.

This book draws on research supported by the Social Sciences and Humanities Research Council of Canada.

Keywords: Validity, Washback, Teaching Chinese as a second/foreign language, Chinese Proficiency Test (HSK), AUA framework

CHAPTER 1

SETTING THE SCENE

1.1 The Problem

In the field of language testing (LT), validity theory has received increasing attention over the past decades. 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. 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. 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. 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 are 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 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, 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. 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 validation and evaluation. Additionally, HSK may be used for professional purposes, thus evidence from users/stakeholders at the macro level are also considered legitimate and necessary from the social perspective.

[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 on the new HSK will be provided in Section 1.3.2.

1.2 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 2.3.2 for a detailed explanation of AUA). Articulating an AUA within the HSK Level 4 context (refer to Table 1.1 for the description of each HSK level) 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 mixed methods research (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 The Context

1.3.1 CSL Education and the "Promoting Chinese Internationally" Policy

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, 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).

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. The teaching and learning of Chinese also has a long history. As early as the seventh century, there have been records of teaching Chinese to foreigners in China. With regard to teaching CFL 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, where the students were mainly missionaries and sinologists (Tsung & Cruickshank, 2011).

More recently, since the reform and the opening up of China to the world in 1978, the Chinese economy has been rapidly growing, and the country has made much progress in building a highly cultural and ideological civilization. China's role as a world economic power in recent years has led to a growing interest in CSL and CFL. China's current PCI policy is yet another critical aspect of its support of CSL/CFL teaching and learning (TCSL). In the early 1990s, Hanban[2] 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,

[2] Hanban (Chinese: 汉办) is the colloquial abbreviation for the Office of Chinese Language Council International (Chinese: 国家汉语国际推广领导小组办公室). It was established in 1987. According to Hanban's official website, Hanban is "a public institution affiliated with the Chinese Ministry of Education" and is committed to "providing Chinese language and cultural teaching resources and services worldwide". Hanban's goals include "making Chinese language and culture teaching resources and services available to the world", "meeting the demands of overseas Chinese learners", and "contributing to the formation of a world of cultural diversity and harmony".

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). In 2020, Hanban changed its name to "中国教育部中外语言交流合作中心 (the Centre for Language Education and Cooperation)", in a bid to deepen the language exchanges and cooperation between China and other countries. 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.

Comparing to the long history of CSL/CFL teaching and learning, research in TCSL is fairly recent and has been conducted in various contexts, such as CSL/CFL in China and other countries, as a second language to ethnic minority groups in China and post-colonial contexts (e.g., Singapore and Hong Kong), 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 curricula, teaching approaches, as well as in learners and their identities.

1.3.2 The Development and Implementation 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) the 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., Test of English as a Foreign Language (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-taker's choice and what the specific test center offers. In the internet-based test, the writing task (for HSK Level 3 onwards) can be completed by using the Chinese input system[3]. 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 HSKK test stands for Hanyu Shuiping Kouyu Kaoshi, literally "Chinese Proficiency Oral Test". It is a separate test, however, the three HSKK levels correspond with the six HSK levels. The current test structure is presented in Table 1.1, and additional detail on the tests' questions/items/tasks are provided in Appendix 1.

[3] 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.

Table 1.1 The New HSK Test Structure

Leve l	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.	Intermedi ate (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	45 questions	10 items	Designed for learners who can read Chinese newspapers and magazines, watch Chinese films, and are capable of writing and delivering a lengthy speech in Chinese.	Advanced (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)

The new HSK (also called the 2.0 version of 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 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 CEFR, and the Scales, is presented in Table 1.2. According to the test agency, in 2019 alone, over 800,000 test-takers took the HSK test from 1229 testing centers worldwide.

Table 1.2 The Estimated Equivalence among the New HSK Tests, the CEFR, and the *Scales*

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)

In 2021, the “Chinese proficiency grading standards for international Chinese language education” (《国际中文教育中文水平等级标准》（GF0025-2021））has been released that guides all aspects of international Chinese learning, teaching, testing, and evaluation. Accordingly, the HSK will be reformed to reflect the new standards, (3.0 version of the HSK). The new standards has abandoned the former five-level scale and incorporated a brand new 3-stage-9-level scale, aiming to describe learners’ Chinese abilities more accurately. As of 2022, HSK levels 1-3 will be categorized into the elementary stage, levels 4-6 will be categorized into the intermediate stage and levels 7-9 will be categorized into the advanced stage. This study is centred on the new HSK (2.0 version of the HSK).

Although the development of HSK has come a long way (as shown in Figure 1.1), evolving and witnessing profound changes, the research was mainly focused on the old HSK, the HSK technique aspects (e.g., internal reliability) published by the test developers, and the application of the HSK in CSL curriculum. Very few empirical studies of the new HSK have been carried out to reflect the reform and the context.

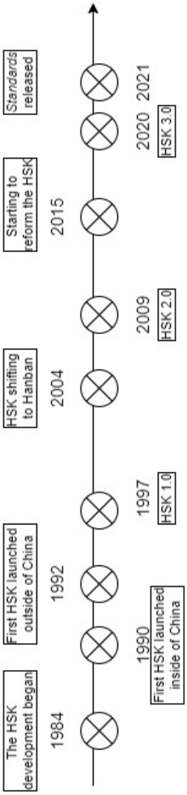
The HSK development began
1984
1990
First HSK launched inside of China
First HSK launched outside of China
1992
1997
HSK 1.0
HSK shifting to Hanban
2004
2009
HSK 2.0
Starting to reform the HSK
2015
2020
HSK 3.0
2021
Standards released

Figure 1.1. The development of the HSK

1.4 Significance of the Study

This 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-making related issues) in a broader social context.

Second, the 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 Study

This book consists of eight chapters and its organization is as follows.

Chapter 1 offers a preview and introduction to the study including the rationale, purpose, context, significance, and organization of the study.

Chapter 2 provides a foundation overview of the validity theory with the focus on consequential validity and washback, and studies on HSK.

Chapter 3 starts with the rational of adopting the mixed-method research design and the overall design of this study is described.

Chapters 4, 5, and 6 respectively present the data collection, data analysis, findings, and discussions for three phases of the study.

Building on the earlier chapters, Chapter 7 includes an overall discussion of the major findings 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 Key Terms Used in This Book

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

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 stakeholders (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.

Hanyu Shuiping Kaoshi (HSK): 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) 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.

CHAPTER 2

ESTABLISHING FOUNDATIONS

2.1 Introduction

This chapter establishes the research foundations by reviewing literature on 1) the evolution and current issues with validity and washback, that is, 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 Exploring the Research Concept

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 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.

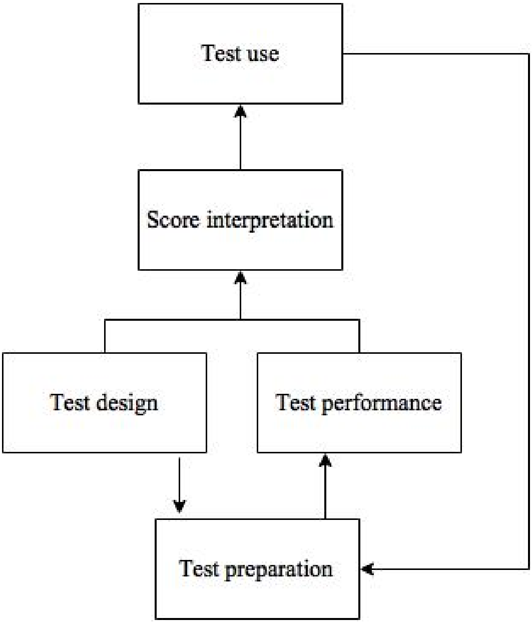

Figure 2.1. Conceptual model: Relationships between washback and test validity (p.68, Xie, 2010)

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).

2.2.4 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 is still 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 English as a second language (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.