A Washback Investigation of the National Matriculation English Test in China from the Perspectives of Test-takers' Attitudes and Behaviors

BY

Jun Wang

Thesis submitted to the

Department of Linguistics

of the Faculty of Human Sciences

of the Macquarie University

in partial fulfilment of the requirements for the degree of

Master of Research

in Linguistics

October 2014

Statement of Candidate

I hereby declare that this thesis has not been submitted for any degree to any other university or institution. The sources of information used and the extent to which the work of others has been utilized has been indicated in this thesis in the manner conventionally approved in the research field in which the thesis fits. The approval form Ethics Committee has been obtained.

(Reference number: 5201400385)

Signature of the candidate:

Date: 10 Oct. 2014

Acknowledgement

I would like to acknowledge and thank my supervisor, Dr. Jill Murray, for her professional guidance, invaluable advices and constructive feedback, from whom I have learnt a great deal. I would also like to thank Prof. Peter Petocz for his expert advices in statistical data analysis.

I would also like to thank Principal Huang, who kindly granted me the permission of conducting this research in his school. Particularly, I would like to thank my mother, Ms. Liqing Song, who have always motivated and encouraged me. Without her endless love and support, I could never finish this thesis.

Abstract

In the field of *Language Testing*, 'washback' generally refers to the influences of language tests on teaching and learning. The increasingly significant role that language tests play in the modern world makes the washback effects a high-interest phenomenon nowadays. However, despite such popularity, there were still many washback contexts remaining unexplored and inadequate attention on how candidates-related factors affect the washback effects. Motived by these gaps, this study aims to conduct a washback investigation of the National Matriculation English Test in China from the perspectives of test-takers' attitudes and behaviors in *Jiangxi* province. This test is one of the most important language tests in China, while only a few studies have done in revealing the related washback effects on English education. Survey questionnaires were collected from 285 high-school students who were also the candidates preparing for the test. The results showed positive presence of the negative washback effects on students' attitudes toward the test, test preparation and English learning in general, but, only a moderate degree of washback intensity was found, which was inconsistent with the high-stakes test factor. Thus, this study suggests a further reconsideration on the influences of '*test stakes*' factor on the extent of washback effects.

Table of Contents

List of Tables and Bars

List of Figures

List of Abbreviations:

- CA: *Curricular Alignment*
- CET Band 4: The College English Test Band 4 of China
- FV: *Face Validity*
- *Jiangxi*: an administrative province located in the southeast of China
- JPEEA: The *Jiangxi* Provincial Education Examination Authority
- MDI: *Measurement-driven-instruction*.
- MOE: The Ministry of Education of China.
- NEEA: The National Education Examinations Authority
- NMET: The National Matriculation English Test
- NMTAR: The 2014 National Matriculation Test Annual Report
- PCA: The Principal Component Analysis
- SV： *Systemic Validity*
- NCEE: The National College Entrance Examinations, or the National Matriculation Tests (NMT).
- TPA: The '*test preparation activity*' (same as the '*behaviors*' in this study)
- WV: *Washback Validity*

Chapter 1 Introduction

1.1 Aims of this Study

Recently, the notion of '*washback*' has attracted considerable research interest in the fields of *Language Testing* as well as general *Educational Assessment*. Traditionally testing research has been focussed on the issues or problems related to validity or reliability. Nowadays, the extensive use of test scores for various educational or social purposes and the increasing emphasis on humanism within the testing discipline make the topics of *test impact* and *washback* topical. Unlike the traditional testing research, washback research covers a broader range of issues including not only the test itself, but also the related areas of teaching, learning, and contextual situations. In the exploration of washback phenomena, attitudinal and behavioural evidence from stakeholders, especially *teachers* and *students/test-takers*, is the most common type of data (Alderson & Wall 1993; Messick, 1996). However, comparing with the amount of empirical attention on *teachers', students/test-takers*' perceptions regarding the issues of washback have been much less explored (Cheng & Deluca, 2011; Green, 2013). Thus, this study particularly aims to hear from the test-takers gathering their perceptions on how the test has affected their learning. The NMET, shortened for the National Matriculation English Test, is one of the most important language tests in China, which is used as a language proficiency test as well as a gatekeeping test for

university admission (Cheng & Qi, 2006; Qi, 2007). The history of such usage could trace back to decades ago, but up to now, only have a few empirical investigations been conducted exploring the washback effects in a limited number of regional contexts of China. As an administrative division for the NMET, *Jiangxi* province is one of the unexplored contexts in China. Motived by both the gap of inadequate empirical attention on hearing test-takers voices and the purpose of improving understanding on how the NMET affects students' English learning, this study seeks to (a) identify the occurrence of washback effects from the perspectives of test-takers' attitudes, (b) examine how the NMET affects test-takers' language learning behaviors in a particular context and (c) provide useful suggestions and implications for future test improvement and washback research.

1.2 The Organization of this Study

This study is presented through **seven** integrated chapters. The first chapter is this chapter including the illustration of research purposes and the organization of this study. Chapter two is the *literature review* reviewing both the theoretical and empirical developments of washback in order to build a rationale for the current study. Before the demonstration of research method, the third chapter represents the analyses of *context factors* including the test factor, the social factor and the school factor to establish underlying linkages between those factors and washback effects. Then, the next two chapters are namely the chapters of *research method*

and *data analysis* respectively explaining the selection of research methodology and the procedures of data analysis. Finally, the results and major findings are provided in the last two chapters： *discussions* and *conclusions*. *Implications* for both future research and educational practices are presented in the last chapter conveying suggestions and future directions to researchers, educationalists and testing practitioners who are interested in the topic of washback.

Chapter 2 Literature Review

2.1 Introduction

Through reviewing the literature, this chapter aims to (a) explore the definitions of washback, (b) review the empirical achievements of washback studies heretofore and then (c) build a theoretical framework for the current study.

This chapter begins with the illustration of how the concept of 'washback' was defined in the literature and how scholars used these definitions to explain its nature. In the second section, it represents the prominent research models that scholars have developed throughout the years to systematically guide washback investigations. After that, in response to the theories displayed in the previous two sections, the third section examines the empirical applications of washback theories for the purposes of pointing out the gap in our present understanding and identifying the leading issues in washback research. The last section is the

conclusion of this chapter. It summarizes all the arguments presented in preceding sections, demonstrates the background framework and provide a theoretical rationale for the research questions.

2.2 Exploring the Concept of Washback

2.2.1 Defining Washback

The notion of *washback* emerged from the ongoing discussion of the impact of testing. The history of testing can go back to thousands of years ago when imperials in Han Dynasty, China, used tests to select the most talented government officials (Cheng, 2008). Over the years, not only in China, the use of tests to select for educational or social purposes occurred worldwide. Examinations were "exploited as a method of control and power---as a way to select, to motivate, to punish" (Spolsky, 1995, p.1). Such control and power was often labelled as inhuman and harmful, but despite that, testing is continuously used in modern societies (Shohamy, 2001, 2007). In the field of *Language Testing*, tests are used to make inferences about learners' language ability or make decisions based on test-takers' language performance. Tests are classified according to their functions, such as 'achievement tests', 'placement tests', 'selection tests', and 'gatekeeping tests' (Hughes, 2013). The controlling power of testing was found to have an enormous degree of influence on test-takers, teachers,

schools, educational systems or even society as a whole (Li, 1990; Bachman & Palmer, 1996; Shohamy, 1997, 1999 and 2001; McNamara, 2000).

For the purpose of bringing to light how test impact works, theorists have presented numbers of different concepts focusing on various aspects of the complex relationship between testing and education. *Measurement-driven-instruction* (MDI) was one of the early notions attempting to illustrate this relationship. Popham (1987) explained the concept of MDI by describing assessments as the 'curricular magnets' with the power of *driving* teachers' instructions, and also claimed that *teach-to-test* was a positive activity when tests were properly designed. The idea of MDI basically supported the undisputable power of testing over education. But later on, opponents to MDI argued that MDI was 'nothing more than psychometric imperialism' (Madaus, 1988, p.84), which would lead to inevitable negativities in teaching and learning. 'Testing programs should…be seen as an ancillary tool of curriculum and instruction…' (Madaus, 1988, p.84). *Curricular Alignment* (CA) was the concept encouraging this type of relationship (Shepard, 1990, 1992). CA believed that "tests should be for monitoring but not driving instructions' (Shepard, 1992, p.5). However, CA was criticized for being 'unethical' by threatening to the test validity. Then, the concepts of *Systemic Validity* (Frederiksen & Collins, 1989) and *Consequential Validity* (Messick, 1989) appeared attempting to clarify the connection between test validity and test impact (further details see section 2.2.2). CA was also

criticized for containing only negative connotations, while the impacts of testing on education also had positive aspects (Hamp-Lyons, 1997). Thus, the term '*washback*' occurred in the 1980s and was applied as the concept encompassing both positive and negative influence of testing on teaching and learning (Alderson and Wall, 1993).

Since the 1980s and 1990s, scholars defined 'washback' in various ways. Pearson (1988) explained that "it is generally accepted that public examinations influence the attitudes, behaviors, and motivations of teachers, learners and parents, and, because examinations come at the end of a course, this influence is seen as working in a backward direction--- hence the term 'washback'' (p.98). Alderson and Wall (1993) defined 'washback' as "teachers and learners to do things they would not necessarily otherwise do because of the test" (p.1). Messick (1996) restricted the concept of 'washback' to mean "the extent to which the introduction and use of a test influence language teachers and learner to do things they would not otherwise do that promote or inhibit language learning" (p.241). Shohamy et al. (1996) used 'washback' to express the overall relationship between testing and learning. Summing up ideas from the available literature, Bailey (1996) stated that washback is "generally defined as the influence of testing on teaching and learning" (p.259). Nowadays, Bailey's (1996) definition is most widely acknowledged and frequently applied in studies (Alderson, 2004, Cheng & Curtis,

2004; Fulcher & Davidson, 2007; Hughes, 2013). Hence, this definition will be used in the present study.

Some of the scholars distinguish the influences of testing on teaching and learning in the classroom and the influences on school, educational system and society as a whole by referring them as the 'micro-level' and the 'macro-level' of test impact respectively (Bachman & Palmer, 1996; Bailey, 1996; McNamara, 2000). In recent studies, scholars commonly use the term of '*washback*' to refer to the 'micro-level' and the term of '*test impact*' to the 'macro-level' (Wall, 1997; McNamara, 2000; Shohamy, 2001; Hughes, 2013). Following this distinction, therefore, the concept of 'washback' used in the present study particularly refers to the micro-level of test impact. Moreover, the words, 'backwash' and 'washback', in literature are seen as interchangeable. There is neither pragmatic nor semantic difference between them, the choice purely by virtue of authors' preference (Alderson, 2004). In literature, the term 'washback' appears with higher frequency. Hence, 'washback' will be used throughout the present study, except in quotations.

2.2.2 Exploring the Nature of Washback

Despite there being a commonly accepted definition for washback, there is no commonly accepted explanation for the nature of washback. There are diverse

interpretations of washback's nature appearing in the literature since the beginning of washback studies. Three of them are the most influential: *bipolar nature, relationship with test validity* and *complex nature,* and the last nature: *complex nature* is the most commonly used one in modern studies.

Firstly, we will consider the *bipolar nature* or the *bidirectional nature* or the *neutral nature*. "The term 'washback' is itself a neutral one, and can be related to 'influence'. If the test is '*poor*', then the washback may be felt to be *negative*…then *good* tests should have *good* effects…" (Alderson & Wall, 1993, p.117). This nature of washback pertains to the existence of two directional effects: *positive* and *negative. Positive washback* generally refers to the beneficial aspects of test impact, whereas *negative washback* refers to the detrimental effects. Reviewing the literature, Pan (2009) summarizes two useful lists of positive and negative phenomena from the perspectives of *classroom settings* and *educational system* (see Table 2.1 and Table 2.2).

Positive Washback	
Classroom Settings	1, Tests induce teacher to cover their subjects more thoroughly, making them complete their syllabi within the prescribed time limits. 2, Tests motivate students to work harder to have a sense of accomplishment and thus enhance learning. 3, Good tests can be utilized and designed as beneficial teaching-learning activities so as to encourage positive teaching-learning processes.
Educational /social system	Decision makers use the authority power of high-stakes testing to achieve the goals of teaching and learning, such as the introduction of new textbooks and new curricula.

Table 2.1 A Summary of Positive Washback (Pan, 2009, p. 261)

Negative Washback	
Classroom Settings	1, Tests encourage teachers to narrow the curriculum and lose instructional time, leading to 'teaching to the test.' 2, Tests bring anxiety both to teacher and students and distort their performance. 3, Students may not be able to learn real-life knowledge, but instead learn discrete points of knowledge that are tested. 4, Cramming will lead students to have a negative attitude toward tests and accordingly alter their learning motivation.
Educational /societal system	Decision makers overwhelmingly use tests to promote their political agendas and to seize influence and control of educational systems

Table 2.2 A Summary of Negative Washback (Pan, 2009, p. 261)

These two lists summarize some of the positive and negative washback effects, which may also be utilizable as a set of criteria in identifying positive and negative washback effects from empirical evidence, but they are less complete in capturing washback's bipolar nature because washback effects varies to different cases. Instead of listing an inventory, Green's (2006, 2007) represents a diagrammatic model to link the generating of positive and negative washback with test design (see Figure 2.1). In this diagram, positive washback occurs in the overlapping areas between test focal construct and test characteristics (format, content, etc.), and negative washback on test preparation is produced when the test characteristics appear to be inconsistent with the test construct. It links the occurrence of positive and negative washback effects with test design, and more importantly, it reveals the potential connection between washback and test construct validity.

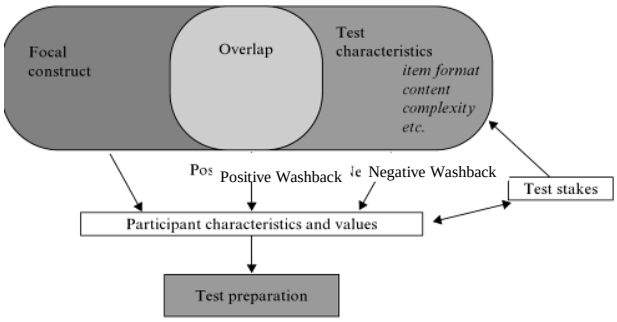


Figure 2.1 A Model of Washback's Direction (Green, 2007, p. 17)

In this model, Green's (2007) employs the concept of test *construct validity* to explain washback phenomena. *Construct validity* regards whether a test measures what it supposes to measure (Bachman, 1990; Bachman & Palmer, 1996; Hughes,

2013). In other words, a test has high degree of construct validity when the test characteristics maximally and accurately reflect the test construct. Applying this definition in Green's (2007) model, it implies that the higher degree of the test validity, the more positive washback the test would produce. It is a comprehensible way to explain the nature of positive washback. As stated in 2.2.1, in the early theoretical development of washback, using test validity concepts to define washback and explore its nature was once a popular approach.

In early exploration of washback theory, some people suggested that the validity of a test should be judged by how much beneficial influence the test had produced. Hence, the concept of *Washback Validity* (WV) occurred, and the key criterion for evaluating WV was whether a test had 'good' washback (Morrow, 1986). *Systemic Validity* (SV) was another concept similar to WV but it had more conceptualized ideas. "A systemically valid test is the one that induces in the education system, curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure" (Frederiksen & Collins, 1989, p. 27). The set of criteria for evaluating SV was "(a) the *directness* of cognitive assessment and (b) the degree of *subjectivity* or judgment required in assigning a score to represent the cognitive skills" (Frederiksen & Collins, 1989, p.28).

However, proponents argued that both WV and SV had far exaggerated the *directness* of the relationship between washback and test validity: a 'good' test could produce negative washback and a 'poor' test could also have positive washback (Alderson & Wall, 1993; Messick, 1996). Alderson and Wall (1993) states that both WV and SV have not been verified by any actual investigations and both theories are more asserted than established. They also point out that the undesirable washback is unlikely to be the direct result of a 'poor' test quality, and the existence of other potential forces could also make contributions to the occurrence of washback. Messick (1996) agrees, and contends that both notions of WV and SV were built upon the ideal perception that learning exercises and test exercises should be seamless to produce the optimal level of positive washback. Such ideal perception is virtually based on circumstantial evidence from insufficient empirical investigations. More importantly, Messick (1996) stresses this ideal rarely exists in reality. "To some degree, *construct under-representation* and *construct-irrelevant variance* are ever with us. The test is never a completely faithful exemplar of criterion behaviors" (Messick, 1996, p.244). Moreover, the activities involved in the test preparation should not be exclusively viewed as test washback. "Washback is only one form of testing consequence that needs to be weighed in evaluating validity, and testing consequences are only one aspect of construct validity needing to be addressed" (Messick, 1996, p.242). To wit, washback effects could only be considered as a type of consequential evidence for test validity when the activities have been

substantially proved to be the production of test influence. This statement clarifies the underlying relationship between washback and test validity, which is that they are not directly connected and the establishment of the latter should not rely on the occurrence of the former. The quality of a test should not be evaluated according to washback evidence alone. Such an idea could also be found in current theories of test validation, such as, in Bachman's (2005) test validity argument and test utilization argument models, test's consequences are regarded as one type of significant warrant to be considered in validity argument but not the only one. Thus, it could be said that washback is closely connected with test validity, but the actual validation requires a much broader consideration. Using test validity is insufficient to describe the nature of washback.

The most accepted characterisation of the nature of washback to date refers to its *complex nature*. According to Cheng and Curtis (2004), the realization of the complexity of test impact was one of the most important theoretical developments in the field of *Language Testing* in the past 30 years. In the late nineties, with the accumulation of knowledge on washback, many scholars began to realize that the assertions and assumptions about washback in previous literature might have oversimplified its nature (Alderson & Hamp-Lyons, 1996; Hamp-Lyons, 1997; Watanabe, 1996, 1997; Wall, 1999, 2000). Washback is actually "a highly complex rather than a monolithic phenomenon. The influence has been observed on various aspects of learning and teaching…" (Watanabe, 2004, p.19). The

phenomenon of washback influences nearly all the things happening in the classrooms as well as the self-study activities outside the classrooms, which could involve numerous uncontrolled variables. Watanabe (1997) identifies some of these variables and classifies them into five categories: *personal variable* (factors of stakeholders), *test variable* (e.g. test contents and test methods), *activity/interaction variable* (e.g. teaching methods and learning methods), *macro-context variable* (society where the test is used) and *micro-context variables* (school settings where the preparation is been done) (p.72-74). Later on, he (2004, p.22) modified this classification by incorporating the *activity/interaction variable* with the *micro-context variable* and adding a new one as the *prestige factors*, which refers to the factors of '*test stakes'* or '*test status'*. This modified version will be applied in the following chapter as a basis for analysing the contextual factors (see Chapter 3).

In light of the establishment of washback variables, Watanabe (1997) conceptualizes five dimensions as guidelines for describing washback complexity: *Specificity*, *Intentionality*, *Covert/overt behaviors*, *Positive/negative* and *Strong/weak* (pp.74-77). He (2004) improves this framework in his subsequent work by deleting the *Covert/overt Behaviors* dimension, adding the *Length* dimension, replacing the title of *Positive/negative* with *Value* and changing the title of *Strong/weak* to *Intensity*. With a clear acknowledgment of the complex nature of Washback, this study, as a small-scale research, will address such nature

through particularly the Value and Intensity dimensions. Here is Watanabe's revised model aligning with the rationales of such selection.

a) **Specificity**: Watanabe (2004) divides this dimension into two clear-cut categories: *general* and *specific*. "*General washback* means a type of effect that may be produced by any tests…*specific washback*, on the other hand, refers to a type of washback that is related to only one specific aspect of a test or one specific test type" (Watanabe, 2004, p.20). For example, the motivation for study harder could be regarded as the *general* washback because most of tests have this influence, while it is *specific* washback when a *Reading* test affects students to concentrate on doing *Reading* exercises. This study concerns washback phenomena of the target test from both *general* and *specific* perspective, such as how the test influence test-takers' attitude in general and how test affect the ways students choose to practice language skills. Thus, this dimension will not be regarded as a particular focus of this investigation.

b) **Intensity**: In general, washback *intensity* refers to the continuum between *strong* effects and *weak* effects. *Strong* washback means that a test plays the role of determining everything that happens in the classroom and teachers instruct nothing but test-related contents; whereas *weak* washback indicates that a test rarely affects teaching and learning. This dimension is closely relevant to this study. In the exploration of test-takers' attitudes, it is necessary to identify how

intense the washback effects is, especially in their selections and uses of language learning activities.

c) **Length**: Washback effects may last for a *long* period of time or a *short* one. For example, washback effects of a 'gatekeeping' test might cease when test-takers enter the institution, and such washback is known as the short-term effects. When test-takers retain their language learning habits that they developed during the period of test preparation after taking the test, the washback could be labeled as long-term effect. Studies show that this enduring kind of washback effects appears dynamic and changes indefinitely over time (Shohamy et al., 1996; Zhan & Wan, 2013). This study will not consider this dimension, as it requires a longitudinal approach.

d) **Intentionality**: This dimension includes *intended washback* and *unintended washback*. Messick (1989) implies the existence of this dimension through making suggestions for consequential validation, which requires the "…evaluation of the intended and unintended social consequences of test interpretation and use" (p.84). McNamara (1996) also supports the presence of both unintended and intended effects of assessments and suggests collecting stakeholders' behaviors and comparing them with the educational goals to identify whether it is intended or unintended. Some of others scholars evaluate the occurrence of this dimension by gathering test designers' intentions and compared

their intentions with empirical washback phenomena (e.g. Andrews, Fullilove & Wong, 2002; Qi, 2005; 2007). Investigating the educational goals from teachers or the intentions from test designers is beyond the scope of this study; hence, this study will not consider this dimension.

e) **Value**: This dimension is accordance with the *bipolar nature* of washback. It supports the existence of both *positive* and *negative* washback effects. The identification of either positive or negative requires to be built upon a thorough consideration of the contextual factors, which may include "*who* it is that actually conducts the investigation within a particular education context, as well as *where*, the school or university contexts, *when* the time and duration of using such assessment practices, *why* the rationale, and *how*, the different approaches used by different participants within the context" (Cheng & Curtis, 2004, p.8). That is to say, similar washback phenomena could have different directional effects in different contexts. This study particularly seeks to identify the occurrence of presumed negative washback phenomena, thus, a process of analyzing the contextual factors is indispensable.

2.3 Theoretical Development in Washback Research Models

To systematize the investigation on washback and capture an accurate description of the washback mechanism, researchers have made greatly theoretical

contributions on developing washback research models. Prominent models include Hughes's (1993) *The Trichotomy Backwash Model*, Alderson and Wall's (1993) *Washback Hypotheses*, Bailey's (1996) *A Basic Model of Washback*, and Wall's (1999) *Henrichsen's Hybrid Model of the Diffusion/Implementation Process*. These models have guided washback studies for more than three decades. In modern studies, inspired by the existing theories in both language education and general education， numbers of new models appeared with ever-growing attention on identifying the mediating factors to the washback effects, such as Green's (2006, 2007) *Model of Washback Direction, Variability and Intensity* and Shih's (2007) *A Washback Model of Students' Learning*. In the next section, these models will be described chronologically.

2.3.1 The Trichotomy Backwash Model

In Hughes' (1993) unpublished but influential paper, he proposes three ways of viewing the phenomena of washback: *participants, processes,* and *products* (see Table 2.3).

Participants	Students, classroom teachers, administrators, materials developers and publishers, whose perceptions and attitudes toward their work may be affected by a test.
Processes	Any actions taken by the participants which may contribute to the process of learning.
Products	What is learned (facts, skills, etc.) and the quality of learning.

Table 2.3 The Trichotomy Backwash Model (Hughes, 1993, cited in Bailey, 1996, diagrammatized in Cheng & Curtis, 2004, p.12)

This model concerns how washback works through three aspects: *who is involved*, *what they do* and *what is learned*. It significantly classifies the types of stakeholders ('*Participants*') and marks how participants' attitudes, perceptions and behaviors ('*Processes*') towards teaching and learning may reflect test impact. Despite it to some extent bypassing the influential factors within the contexts and being unclear of how to examine the '*Products*' component, the contribution of an early development among washback theories could not be denied. This model has been widely applied in washback studies for individual research design or as a basis for developing new model.

2.3.2 The Washback Hypotheses

The Washback Hypotheses (Alderson & Wall, 1993, p.120-121) refers to a set of possibilities generated from literature and personal experiences in order to build a feasible framework for research, which explicitly includes 15 hypotheses:

1) *A test will influence teaching.*
2) *A test will influence learning.*
3) *A test will influences **what** teachers teach;* and
4) *A test will influence **how** teacher teach;* and therefore by extension from (2) above

5) *A test will influence **what** learners learn;* and
6) *A test will influence **how** learners learn.*
7) *A test will influence the **rate and sequence** of teaching;* and
8) *A test will influence the **rate and sequence** of learning*
9) *A test will influence the **degree and depth** of teaching;* and
10) *A test will influence the **degree and depth** of learning.*
11) *A test will influence attitudes to the content, method, etc. of teaching and learning.*
12) *Tests that have important consequences will have washback;* and conversely
13) *Tests that do not have important consequences will have no washback.*
14) *Tests will have washback on **all** learners and teachers.*
15) *Tests will have washback effects for **some** learners and **some** teachers, but **not** for others.*

This model demonstrates washback's *contents* (teaching and learning), *characteristics* (rate, sequence, degree and depth), *variables* (attitudes and test stakes), and *stakeholders* (teachers and learners) from a micro-aspect of teaching and learning. The basic intention of this mode is not to completely capture how washback works but to '*lay out the territory*' (Tsagari, 2006, p.22). The fifteenth hypothesis was expanded later on as "*tests will have different amounts and types of washback on some teachers and learners than on other teachers and learner*" by additionally considering the influences of '*the status of the test*', '*the extent to which the test being counter to current practice*', '*the extent to which teachers and textbook writer thinking about appropriate methods for test preparation*', and '*the extent to which teachers and textbook writers being willing and able to*

innovate' on the washback effects (Alderson & Hamp-Lyons, 1996, p. 296). This model has successfully pointed out the direction of conducting washback research at that time, inspired early considerations on the influential factors affecting washback and positively affects the subsequent empirical studies for decades (e.g. see Alderson & Hamp-Lyons, 1996; Watanabe, 1996, 1997; Stoneman, 2006; Li, 2009; Xiao, Sharpling & Liu, 2011).

2.3.3 A Basic Model of Washback

Inspired by Hughes's (1993) *Trichotomy Model* and Alderson and Wall's (1993) *The Washback Hypotheses* model, Bailey (1996) combines the merits of both and proposes an advanced model with directional connections between components (see Figure 2.2).

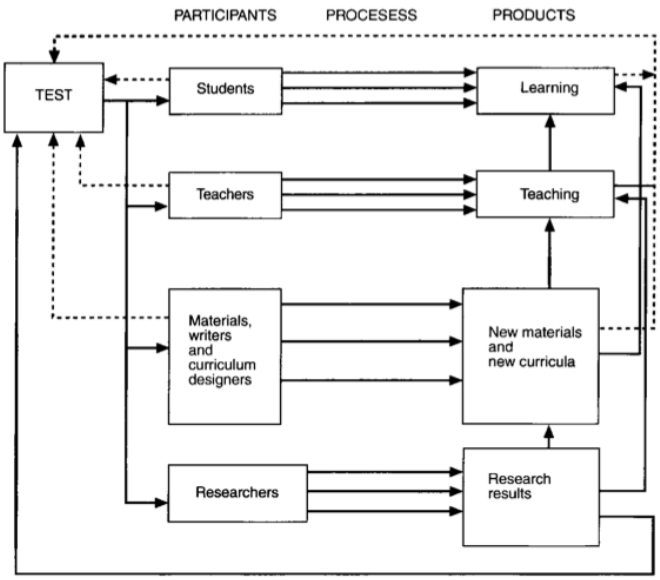


Figure 2.2 A Basic Model of Washback (Bailey, 1996, p.264)

This model changes the primary groups of stakeholders in Hughes's model (see Table 2.3) to *the publishers, the administrators, the curriculum designer* and *the researchers*. It improves the previous models by clearly demonstrating how test impact flows and how teaching and learning evidence reflect test impact backward. It also suggests when dealing with individual groups of stakeholders, the process involved in washback should be considered accordingly. This model to some extent captured the complexity of washback's mechanism, but to somewhat it failed to accurately depict the '*process*' component. It demonstrates the '*process'* components as a linear style of relationship with no consideration of other forces, such as social factors or cultural factors. According to the current knowledge, washback is a complex phenomenon interrelated with diverse variables, thus some of the connections in the '*process*' component are oversimplified as a linear connection.

2.3.4 Henrichsen's Hybrid Model of the Diffusion/ Implementation Process

Unlike previous models, Wall's (1999, 2000) model has a distinctive purpose: to implement an *assessment-driven reform*. The recognition of the probable effectiveness by using tests to drive educational reforms could date back to the late 1980s, such as the proposition of MDI (Popham, 1987). Inspired by previous ideas and her own investigation with Alderson in Sri Lankan (Wall & Alderson, 1993), Wall (1999, 2000) argues that the existing strategies at that time of using the power of examination to facilitate desirable changes in the educational

practices were challenged because the developers may have overlooked the complex nature of test impact. In order to build a more feasible model and minimize the unintended consequences, Wall (2000) combines test impact theories with innovation theories and generates the *Henrichsen's Hybrid Model of the Diffusion/Implementation Process* (see Figure 2.3).

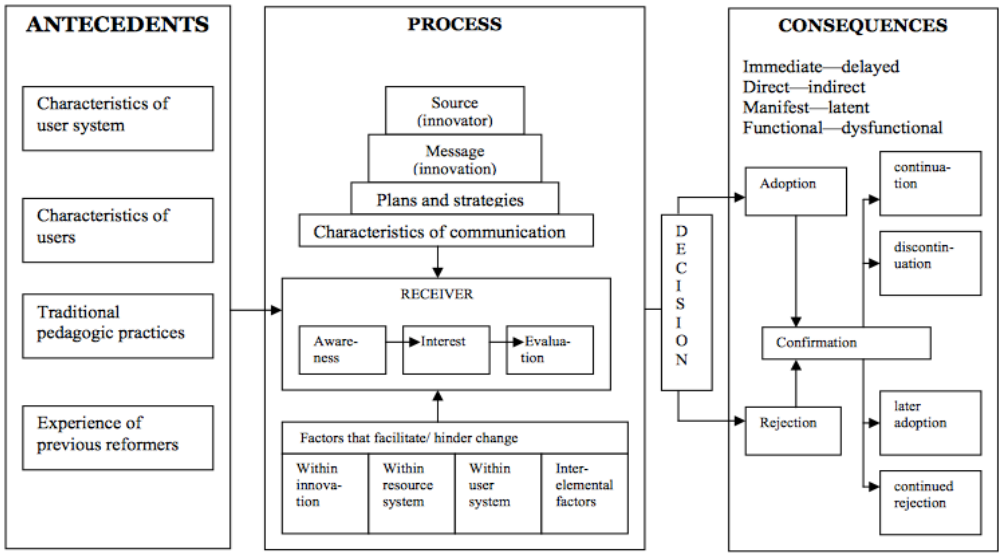


Figure 2.3 Henrichsen's Hybrid Model of the Diffusion/Implementation Process (Wall, 2000, p.504; Wall & Horák, 2008, p.8)

This model could be interpreted as an explanation of washback mechanism, as well as a guiding framework for implementing educational innovations. The process of implementation is broken down to three different components: (1) *The Antecedents*: the existing conditions of contexts before innovation, or known as *The Baseline Study* (e.g. see Wall & Horák, 2007, 2008), (2) *The Process*: the

investigation of both facilitating and hindering factors of innovation during implementation and (3) *The Consequences*: the outcomes of implementation and the results of the interaction between two former components. This model particularly concerns the intricate factors which could either facilitate or hinder the implementation and emphasizes the investigation on those factors should be conducted throughout three components.

Due to the purpose of introducing intended educational changes, this model primarily focuses on the feasibility of implementation and the management of the procedures. It puts less emphasis on making comparisons between factors, such as the differences between different stakeholders or whether the stakeholders' factors are interwoven. Besides, applying this model is less practical for small-scale studies. Researchers with limited time duration or financial restriction may find it impractical to apply this model.

2.3.5 Some Newly-developed Models

2.3.5.1 A Model of Washback Direction, Variability and Intensity

In this century, an increasing amount of consideration has been devoted to examining the influential factors that could alter the washback phenomena. Those factors could be widely found in the newly-developed models. For instance, Green's (2006) model particularly links the *value* and *intensity* dimension of

washback phenomena with *test factors* and *personal factors* with the purposes of predicting washback's direction (same as '*value*') and *intensity* (see Figure 2.4).

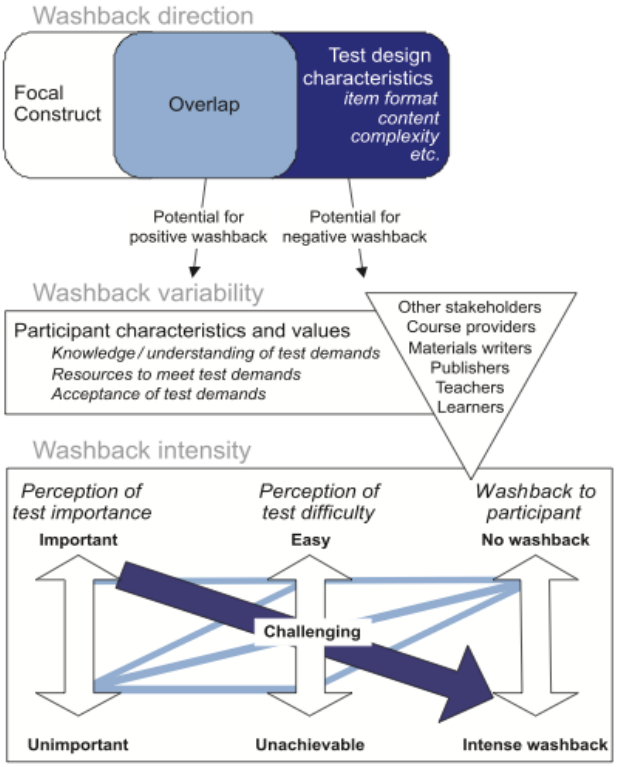


Figure 2.4 A Model of Washback's Direction, Variability, and Intensity (Green, 2006, p. 340; Green, 2007, p.24)

Firstly, as stated above, this model links washback with test validity by highlighting the occurrences of positive washback within the overlapped area between *test construct* and *test characteristics*, while negative washback within the area of '*construct under-representation*', which can be said to occur when "the test is too narrow and fails to include important dimensions or facets of focal constructs" (Messick, 1996, p.17). According to Green's (2013) own explanation, this model essentially outlines the washback mechanism by presenting : i) the *test*

design as a key determinant of *washback direction*, ii) *participants' values, motivations* and *resources* as the major determinants of *washback variability* and iii) *participants' perceived importance and difficulty of the test* as the determinants of *washback intensity*. This model explicitly links the washback variability and intensity with *personal factors* and *test factors*, however, some recent studies find that the *context factors*, or the *extrinsic factors*, also had significant impacts on washback's occurrence.

2.3.5.2 A Washback Model of Students' Learning

Unlike Green's model, this model takes a broad range of influential factors into account in portraying washback phenomena. The assumption that high-stakes tests usually have significant washback effects was stated in many previous publications, (Madaus, 1988; Shohamy, et al., 1996; Cheng, 1997, 1998; Shohamy, 2001; Qi, 2005). However, in Shih's (2007) washback investigation on a high-stakes test, the outcomes have disagreements. Results show that the high-stakes test had only little washback effects on learners' language learning. Shih (2007) argues that such outcome may be affected by the contextual factors, which could not be explained by the existing models in the literature. Hence, a new model is required, in which these contextual factors are classified into three categories: *Extrinsic Factors*, *Intrinsic Factors* and *Test Factors* and each category contain a detailed inventory of sub-factors (see Figure 2.5).

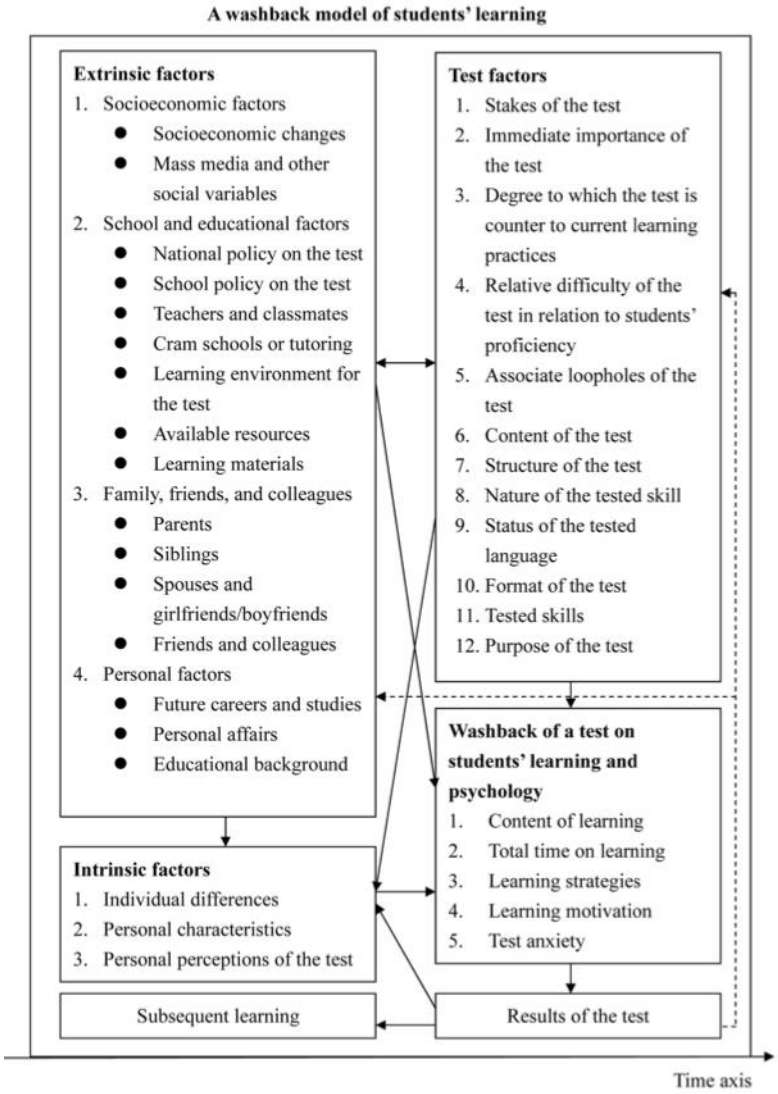


Figure 2.5 A Washback Model of Students Learning (Shih, 2007, p.151)

This model supports the *complex nature* of washback and stresses the significance of contextual impacts on washback's occurrence. Shih's (2007) study further suggests that historical assumptions should not be taken directly for granted without considering the contextual situations. This model also imply that, despite the concept of 'washback' referring specifically to the *micro-level* of test's

influences on teaching and learning, the understanding of it still requires the consideration of *macro-level* context.

2.4 Empirical Studies on Washback

As Cheng and Curtis (2004) note, "there seems to be at least two major types or areas of washback or backwash studies—those relating to traditional, multiple-choice, large-scale tests, which are perceived to have had mainly negative influences on the quality of teaching and learning, and those studies where a specific test or examination have been modified and improved upon, in order to exert a positive influence on teaching and learning" (p.3). Taking this classification and the status quo of washback's research into account, the major characteristics of empirical studies concerning washback could be mapped as:

Types	*Type One*: Traditional large-scale tests	*Type Two*: Test reform and curricular innovation
Main Purposes	To evaluate or investigate washback phenomena understanding what it looks like	To encourage positive washback and inhibit negative washback using the knowledge of how washback mechanism works
Applied Framework	Models or Frameworks capturing the complexity of washback phenomena	Implementation models or washback mechanism models for implementing innovations
Methodological characteristics	Exploratory or descriptive studies, such as case study or ethnography using one-shot or cross-sectional approaches	Longitudinal studies, mostly applying mixed methods and collecting data at different times.

Table 2.4 A General Description of Two Major Types of Washback Empirical studies

Literature shows that *Type One* studies are often applied as the baseline study for *Type Two*. For example, Wall and Horák (2007, 2008) conducted both types of studies, namely one for investigating the effects of an existing test by documenting the educational setting and stakeholders' perceptions and the other for investigating the effects of a new test which had been revised substantially,

and then compared the results with the preceding baseline study results to evaluate the reform. The present study is a typical *Type One* study with main purpose of investigating how washback looks like in this context, which may provide baseline evidence for a further test reform (details see 3.1.2).

2.4.1 Washback of High-stakes Tests

Despite the developments and changes of central topics among the empirical studies of washback, the test impact of high-stakes tests has never left the spotlight. The attention on high-stakes testing could be traced back to the very beginning of test impact studies. The *stake* of a test is defined as 'the extent to which the outcomes of a test can affect the candidates' futures' (Davies et al., 1999, p.185), and the *high-stakes tests* are the tests "whose results are seen—rightly or wrongly—by students, teachers, administrators, parents, or the general public, as being used to make important decisions that **immediately** and **directly** affect them" (Madaus, 1988, p.87). These important decisions is categorized into four types: "(a) graduation, promotion, or placement of students; (b) the evaluation or rewarding of teacher or administrators; (c) the allocation of resources to schools or school districts; and (d) school or school system certification" (Madaus, 1988, p.87). In the current century, three types of high-stakes language tests have attracted most of the empirical attention: *the nationwide, large-scale tests with the purpose of selection*, *the international language proficiency tests* and *the nationwide exit tests for graduation*. Studies

have found that these three types of language tests attaching with high-stakes functions usually produce detrimental effects to the related teaching and learning (see Table 2.5).

Test Types	Examples	High-stakes Decisions	Findings Regarding the Negative Impacts of High-stakes Tests
International Proficiency Tests	*IELTS* (International English Language Testing System); *TOEFL* (Test of English as a Foreign Language)	Higher Education Admission; Immigration; Overseas Employment	• Lack of authenticity of target language use as a gatekeeping test (Moore & Morton, 2005; Green, 2006). • The commercial pressures for successes in language schools have induced negative washback effects on teachers (Hayes & Read, 2004). • There are mismatches between stakeholders' perceptions and what IELTS is claimed to assess which have led to misuses of test scores (Murray, Cross & Cruickshank, 2014); • IELTS 'band' has used by most institutional decision-makers by its symbolic value, which is considered as negative by overlooking the diverse academic environments in tertiary educations (Smith & Haslett, 2007). • IELTS was found producing strong impact on the content of preparation courses, or known as '*teaching narrowly to the exam*'(Moor et al., 2009)
University Entrance Exams	China: *NMET* (National Matriculation English Test); Indonesia: *ENE* (English National Examination)	University Admission; Secondary School Graduation	• Unintended teaching and learning foci (Qi, 2007); • Teaching to the test and narrowing the curriculum (Jones & Egley, 2004; Sukyadi & Mardiani, 2011); • Mainly focusing on test-taking strategies (Sukyadi & Mardiani, 2011); • Invalid in measuring proposed capabilities (Rowland, 2011); • Text anxiety and pressures would cause demotivation for language learning and counterproductive activities (Jones & Egley, 2004); • Increasing dropout and retention rates (Madaus & Clarke, 2001; Haney, 2000) • Students' negative perceptions towards the test itself and the preparation for it (Yildirim, 2010).
College Exit Tests	*CET Band 4* (College English Tests); *GEPT*(General English Proficiency Test)	Graduation, or sometimes Employment	• *CET Band 4* is gainful in promoting receptive learning but not productive learning, such as mainly focusing on learning vocabulary by rotes than on developing actual language skills (Cai, 2006); • Almost no positive impact on motivating learners for long-term post-test learning (Zhang & Wan, 2013) • The GEPT seems to fail in promoting the claimed 'lifelong learning and English acquisition' (Cai, 2006);

Table 2.5 Studies Investigating the Negative Impacts of High-stakes Tests in the New Millennium

As Table 2.5 indicates, many detrimental effects have directly linkages with the high-stakes decisions. In other words, '*test stakes*' has been increasingly recognized as an important factor that could mediate the washback effects. However, according to both previous and recent literature, such importance is solely explained by a widely accepted assumption that the higher the test stakes or the bigger the decisions attached to a particular test, the more significant test impact the test would produce (Alderson & Hamp-Lyons, 1996; Shohamy et al., 1996; Cheng, 2005; Qi, 2005). That to say, current understandings on the concept of '*test stakes*' encompass only a dichotomous distinction between '*high-stakes*' and '*low stakes*', and also the recognition of the relationship between test stakes' levels and washback's extent entails only a presumed positive correlation. Such situations virtually make the analysis of '*test stakes*' factor in washback empirical studies difficult. To overcome such difficulty, Stoneman (2006) developed a 3-factor framework for analysing '*test status*' (a similar term to '*test stakes*). This framework was generated based on an empirical comparison between two high-stakes English tests in terms of test statuses. The framework includes: *name recognition/respect (the extent to which the test is known and respected)*, *official sanction/mandate (whether the test has the explicit sanction from the local government)* and *extent of implementation (the number of school/institutions that send learners to partake in the test and/or the size of the candidature)* (Stoneman, 2006, p.394). The target test of this study is a high-stakes test with the function of selection. The washback investigation of it requires an evaluation of the test

variables. Thus, this study will apply this framework in the analyses of the '*test stakes*' factor (see section 3.1.2).

2.4.2 Washback and Different Stakeholders

Another leading empirical topic among washback studies is the investigation of washback effects on different stakeholders. In literature, scholars identify the groups of stakeholders in various ways. Hughes (1993, cited in Bailey, 1996) classifies the types of *participants* in washback phenomena as: *students*, *classroom teachers*, *administrators*, *materials developers* and *publishers* (see Table 2.3). According to Bailey's (1996) research model (see Figure 2.2), there are mainly five types of stakeholders involved in washback phenomena: *students*, *teachers*, *materials writers*, *curriculum designers* and *researchers*, while Hamp-Lyons (1997) lists the five key groups of stakeholders as: *learners*, *teachers*, *parents*, *government and official bodies*, and *the market place*. Rea-Dickens (1997) expands the list into: *language testers*, *teachers*, *parents*, *administrators*, *teacher educators*, *sponsors and funding bodies*, *government bodies*, *the public, various national and international examination authorities*, *members of working parties*, *curriculum committees*, *test administers* and *test users*. Regardless of these various classifications, *teachers* and *learners/test-takers* are always the key types of stakeholders. Drawing together the ranges of potential test stakeholders, I propose a tentative framework to demonstrate the different distances between them and the test within the testing community (see Figure 2.6).

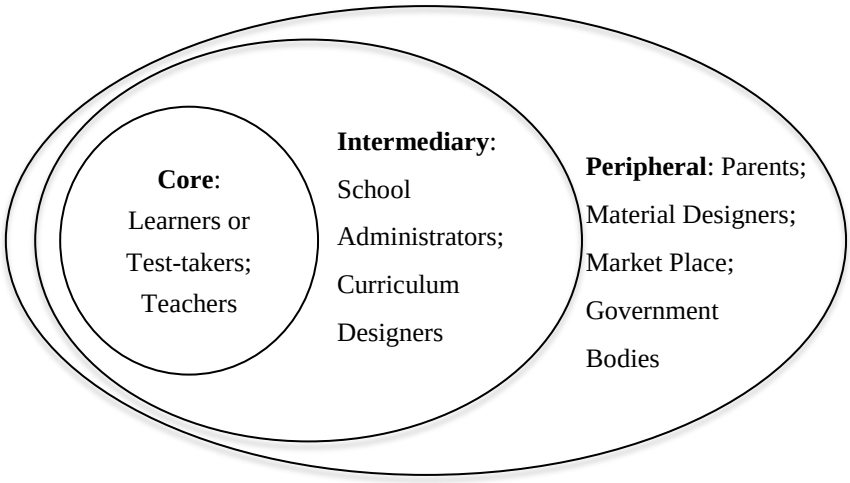


Figure 2.6 Stakeholders in the Language Testing Community

However, comparing with the number of studies using *teachers* as the major source of data, studies focusing on hearing *students'* voices are considerably less. Such imbalances have been noticed since the late 1990s, but the washback studies on learners are still relatively few (Tsagari, 2006; Stoneman, 2006; Green, 2007, 2013; Xie & Andrews, 2013). As Spratt (2005) reviews and summarizes, the existing empirical studies have identified a broad range of *teacher-related factors* affecting the degrees and kinds of washback, such as 'teacher beliefs', 'teacher's attitudes', 'teachers' education and training' and other factors (p.29). However, the *student-related factors* have not been well explored, especially the relationship between washback effects and students' learning outcomes (Hughes, 1993; Wall, 2000; Alderson & Banerjee, 2001). Motivated by this fact, this study aims to explore the washback phenomenon specifically from the perspectives of

students' attitudes to identify any occurrence of negative washback and the possible relationship between test-affected attitudes and the learning product[1].

2.5 Summary and Research Questions

In the first section of this chapter, I have demonstrated how the notion of 'washback' was defined in the literature, in which Bailey's (1996) general definition of *'washback'* referring to *the influence a test has on teaching and learning done for the purpose of test preparation* was selected to be adopted in the present study. In the second section, this chapter has examined the historical exploration on washback's nature, highlighted that washback being a complex phenomenon was the commonly accepted explanation for the nature, and selected Watanabe's (1997, 2004) multi-dimensional framework to guide the consideration of washback's complexity and the analyses of contextual factors of this study.

In the third section, there was a chronologically sequenced outline of several research models that theorists had developed throughout the years to underpin systematic washback investigations. Those models have diverse foci, such as exploring the educational aspects affected by washback, applying washback theories in implementing educational innovations and illustrating factors affecting

[1] 'The learning product' in this study will be represented by participants' average test performances on the NMET mock tests instead of specific language skills or abilities (details see section 3.3).

washback's occurrence. Current trend shows increasing emphasizes on the macro-contextual factors that could mediate washback effects.

In the fourth section, this chapter reviewed some of the noteworthy empirical achievements in the past three decades with slightly more weight on the current findings. Firstly, empirical washback studies could be mainly classified into two types: the type investigating large-scale tests' washback effects and the type which focus on the implementation of a curricular innovation, and there are some methodological differences between them. Secondly, high-stakes tests have attracted a large proportion of empirical attention within washback studies. Most studies have contributed in identifying the exact washback effects within various contexts, but the current understanding has not fully explored the relationship between '*test stakes*' factor and test impact. Finally, among the empirical attention of washback, the investigations on *learners/test-takers*, as one of the key types of stakeholders in the testing community, appear to be inadequately conducted. Motivated by this gap, this study specifically aims to explore the washback phenomenon through test-takers' attitudes and behaviors.

Hence, there are three overarching research questions for this study:

- *Did students report that the NMET[2] as a high-stake test have negative washback effects on their attitudes toward the test, test preparation and English learning? If it did, how intense were the effects and why?*
- *Did students report that the NMET as a high-stake test have negative washback effects on their behaviors as test preparation activities? If it did, how intense were the effects and why?*
- *Was there a relationship between the ranges of students' average scores in the monthly NMET mock tests and their negative attitudes to the NMET?*

Chapter 3 Research Context

As a complex phenomenon, washback investigations require the consideration of contextual factors that could potentially mediate the effects (Cheng & Curtis, 2004; Spratt, 2005; Watanabe, 2004; Shih, 2007, 2010). As stated above, Watanabe (2004) provides a useful framework to systematize such consideration (see section 2.2.2). Following this framework, this chapter will consider the *test factors* (section 3.1: test review), *prestige factor* (section 3.1.2: test stakes), *personal factors* (section 3.2.2: school context), *micro-context factors* (section 3.2.2: school context), and *macro-context factors* (section 3.2.1: socio-cultural context) within the domain of washback phenomena that this study seeks to explore. In the analyses of the *test factors*, this chapter will provide a short test

[2] The NMET is short for the National Matriculation English Test, and test review is presented in section 3.1.

review including the description of test characteristics and the evaluation of test stakes. Concerning the candidates' *personal factors*, this chapter will present a general illustration of the English teaching and learning situations in China to demonstrate how and why the participants of this study may choose to learn English and prepare for an English test. In the consideration of *context factors*, including both micro and macro, this chapter will identify some of the social, educational and cultural characteristics of *Jiangxi* province, where the research site was, and then discuss how these characteristics might affect the washback phenomena.

3.1 Test Review

3.1.1 Test Description

The NMET is short for the National Matriculation English Test, which is a compulsory subject in the test battery of National College Entrance Examinations (NCEE) (*'Gaokao'*) in China. The NCEE is authorized by the National Education Examinations Authority (NEEA) directly under the monitoring of the Ministry of Education (MOE) of China.

The NCEE has two overlapping sets of test batteries, namely *the Tests for Liberal Arts* and *the Tests for Sciences*. The overlapping subjects, which are also the

compulsory ones, are *Chinese Language and Literature*, *Mathematics* and *English* (or sometimes other foreign languages). The differences between these two sets are that *the Tests for Liberal Arts* also include *History*, *Politics* and *Geography* while *the Tests for Sciences* include *Physics*, *Chemistry* and *Biology*. In China, the actual administration of the NCEE is established according to a two-level authoritative management system. At the first level, the MOE issues a national test syllabus and two sets of ready-to-use tests per battery each year. Some administrative divisions apply the ready-to-use tests directly but some of others are authorized with autonomy for test-designing to avoid regional discrimination. In 2014, there were 16 administrative divisions used the self-design NMETs and *Jiangxi* province was one of them. Moreover, since 1998, the testing system has known as the '3+x' test program. The '3' refers to the three compulsory subjects but the filling-in of 'x' varies to provinces. In *Jiangxi* provinces, the authoritative department in charge of the NCEE is the *Jiangxi Provincial Education Examination Authority* (JPEEA), which filled the 'x' with '1'. This '1' refers to a form of integrated test as a systematic integration of three test-battery-specific subjects (see Table 3.1).

Subject	Test Date	Test Time (mins)	Mark
Chinese Language and Literature	June 7th	150	150
Mathematics		120	150
Integrated Liberal Arts/Integrated Sciences	June 8th	150	300
English		120	150

Table 3.1 The NCEE's Test Dates, Test Duration and Marks

All the NMETs are large-scale tests, and the ones used by several administrative divisions are, as Hamp-Lyons identifies, 'ultralarge-scale' tests (Jin, 2014, p.344). According to the 2014 National Matriculation Test Annual Report (NMTAR), the number of NMET test-takers nationwide in 2014 was approximately 9.39 million (http://www.eol.cn/html/g/report/2014/index.shtml). Among them, there were around 0.3 million of them took the *2014 Jiangxi* NMET (http://gaokao.eol.cn/). Thus, the *Jiangxi* NMET is also a large-scale test.

Additionally, all the NMETs are norm-referenced standardized tests (Qi, 2007). All the NMETs are designed according to the national standardized test syllabus. Universities establish their benchmarks for entry each year using statistical analyses, including such as the rankings of achieved scores, the provincial or regional normal distribution, and the multiple comparisons between provinces.

That means, even though test-takers take the different NMETs, the competition of top-class universities admission is actually nationwide.

The *Jiangxi NMET* contains two volumes: *volume one* contains only objective questions (multiple-choice items), whereas *volume two* contains subjective questions (non-multiple-choice items) (see Table 3.2 & Table 3.3). (See *Appendix One*: *The Translated Jiangxi NMET 2014*)

Volume	**Language Ability**	**Marks**	**Percentage (%)**
1	Listening	30	20.0
	Grammar	15	10.0
	Reading + Vocabulary (Integrated Task)	30	20.0
	Reading	40	26.7
2	Reading + Writing (Integrated Task)	10	6.7
	Writing	25	16.7
Total		150	100.0

Table 3.2 The NMET's Testing Language Abilities and the Percentages of Marks

Item Types	**Contents**		**Number of Items**	**Marks**	**Approx Time for Completion**
Multiple-Choice Items	Part 1	Listening	20	30	20
	Part 2	Language Application: Clozes	35	45	25
	Part 3	Reading Comprehensions	20	40	35
Non-multiple-choice Items	Part 4	Reading & Writing	5	10	40
	Part 5	Writing	1	25	
Total			80+1	150	120

Table 3.3 Composition of the *2014 Jiangxi NMET*

As Table 3.2 and Table 3.3 show, *reading*, as a type of testing abilities, is tested both individually and integratedly. It appears in several different test parts and makes up of a comparatively large percentage of the total mark. Such test structure implies the possibility that the test designers to some extent value the *reading* ability more strongly. Therefore, it might be reasonable to postulate that the candidates' test preparation was affected by such underlying value and spent more of their efforts and time practicing *reading* skills and doing reading exercises than other skills.

3.1.2 Test Stakes

Applying Madaus's (1988) definition of 'high-stakes tests' (see 2.4.1), the NMET is unquestionable a high-stakes test. The NMET's test scores are used to make decisions about both high-school exit and university admission. As noted above, '*test stakes*' has been widely regarded as an important factor to be considered in washback research and this study has selected Stoneman's (2006) model for analyzing the test stakes factor (see section 2.4.1). The model includes: *name recognition/respect*, *official sanction/mandate*, and *extent of implementation*. Inspired by this model, the analysis of the test stakes factor of this study includes (a) '*the size of the candidature* and the competition' (*extent of implementation*), (b) 'the normal preparation duration and the *once-per-year* test-taking opportunity' (*official sanction/mandate*) and (c) '*the meanings of the test scores* in the socio-cultural context of China' (*name recognition/respect*). Combining the analyses of these three elements, I argue that the test stakes level of the NMET is **extremely high**.

For the *first* element, according to the NMTAR, the national university admission rate in 2013 was around 75%, and the rate for the top-class universities ('*yiben*') was less than 25%. Notwithstanding, the competition was much fiercer than 25% appears when more than 9 million test-takers were involved. A candidate with a

one mark difference in score could conceivably have surpassed hundreds or even thousands of competitors. That is why Chinese people always say '*never take one mark lightly because it may change your life*". Secondly, the NMET is a mandated test for students seeking university admission. Candidates can take the tests multiple times till they are satisfied with their results. Normally, the minimum preparation for the NCEE is three years, which begins at the moment students enter high schools. Chinese high-school students always metaphorically compare the preparation for the NCEE to '*running a marathon*'. However, as mentioned above, the '*marathon*' could be extended for one more year or years if candidates achieved unsatisfactory scores. This annually test-taking policy is fundamentally a catalyst of the high-degree test stakes by causing more test-taking anxiety and pressure.

Thirdly, the scores of the NCEE are often associated with twofold cultural identities embedded in the Chinese society. At surface level, the scores represent intellectual superiority, which have the power of determining educational opportunities and sometimes future employment. But more deeply, being accomplished in the NCEE is believed as a mean of freeing themselves from social inequality. Candidates from middle classes or lower are entitled to break the social stratification by entering the top-class universities. Many educationalists in China argue that this ideology of earning social privilege through the NCEE has significantly increased the fierceness of the competition

(Davey et al., 2007). 'Brutal' is an adjective commonly used in the news media to describe such competition. Thus, taking these three aspects into account, I argue that the NMET, as the compulsory subject of NCEE, represents an extremely high level of test stakes.

In 2013, the MOE affirmed a future test reform specifically innovating the NMET and one of the policies involves a substantial reduction in the level of test stakes by changing the *once-per-year* test-taking opportunities to *multiple* times each year and the university admission offices will only receive the most satisfactory scores (www.gaokao.com/). It is also notified that the actual implementation of this reform might need years to accomplish nationally. Located in the southeast China, *Jiangxi* is relatively far away from the political centre. Thus, some educationalists presumed the test reform for the *Jiangxi NMET* might occur within the next two or three years (http://www.edu.sina.com.cn/). From this point, this study might be useful in providing baseline evidence for examining the effect of future changes on test impact.

3.2 Research Context

3.2.1 Socio-cultural Context

The research site is located in *Jiangxi*, a southern *inland* province of China. Since

the nationwide '*Four Modernization*' reform signified by Chairman Deng in 1983, tremendous amounts of effort and resources have been allocated to the foreign language education in China (Mao & Min, 2004). However, unlike the *coastal* region, where studies on language tests have been mostly conducted such as Shanghai, Guangdong, and Hong Kong, the inland region has attracted scant empirical attention up to this point, and also receives fairly limited educational resources due to its less developed economic background (Hu, 2003). Furthermore, despite *Jiangxi* being geographically close to the coastal line on the map, the international exposure it receives is considerably less than the coastal region due to the topographical features of this area as hilly and wooded which to some extent limits the access. Thus, the social factors, as the limited empirical exploration, the less developed economic background and the little international exposure, indicate that the development of English education there, involving both teaching and learning, is significantly less rapid than the coastal region.

For the cultural context, as found in Garrott's (1991) extensive survey collecting Chinese language learners' attitudes and cultural values, Chinese people tend to interpret '*learning*' as "…the accumulation of knowledge and the reading of books rather than the practical process of using the knowledge" (p.35). In other words, Chinese people culturally prefer '*learn to use*' to '*learn by use*'. Although it has been argued that such value is virtually a resistance to the implementation of communicative language-teaching approach, this type of cultural value could

still be found in how modern Chinese high-school students learn English (Hu, 2002; Shi, 2006). Thus, because of the possible impact of these social and cultural factors, this study decided to focus more strongly on the traditional types of learning attitudes and activities rather than the more modern ones (see section 4.2.1).

3.2.2 School Context

The research site is one of the *province-wide key high schools* in *Jiangxi*. The concept of *key high school* has existed in Chinese society since the beginning of the NCEE's history, and refers to the high schools with comparatively high-qualified staff, top-class university admission rates and high-demanding educational goals. *Key high schools* in *Jiangxi* are authorized by the provincial Department of Education according to pre-determined criteria. According to these criteria, *Key high schools* are classified into province-wide level, district-wide level and city-wide level. Schools at the same level apply similar educational systems, employ staff with equivalent qualifications, and admit students with homogenous academic backgrounds. As one of the provinces-wide key high schools, this research site shares close scholastic resemblances with other key high schools in this level. Thus, the washback effects of the NMET on test-takers in this school may be somewhat representable in making inferences for other province-wide key schools in this area.

For the contemporary Chinese high-school students, English learning normally starts at the third grade of primary school, and *English* becomes one of the key subjects. Thus, the participants of this study, as senior-year high-school students, may normally have been taught English for more than seven or eight years. Besides, as it is a compulsory subject, the achieved scores on *English* are always powerfully determinative in both secondary schools entry and high schools entry[3]. All the participants in this study had been required to take high-stakes English exams since primary school. In other words, they were well experienced in taking high-stakes English tests.

Other particular *school factor* that needed to be considered is the monthly testing policy. As noted above, the 3-year high-school studying is essentially the test preparation for the NCEE (see section 3.1.2). In order to monitor students' studying, such a policy is applied throughout the period. The design of these periodic tests follows the exact test format, test content and task types as the NMET. Hence, students' average scores on the NMET monthly mock tests may provide an indicator of their probable eventual scores in the real NMET.

[3] In Chinese educational system, the secondary and high-school education are separated each lasting 3 years, and for both levels students are required to take entrance exams to be admitted (Jin & Cortazzi, 2006).

3.3 Significance of this Study

Among the empirical attention on the NMET's washback effects, this study is significant in **three** ways. Firstly, it was conducted at a time, which was close to a test reform. The data collected in this study could be viewed as baseline evidence for evaluating the changes brought by this reform. Secondly, this study was one of the few washback studies particularly exploring how a test affects *students/ test-takers*' learning and also linking their attitudes with learning outcomes. It has been suggested that many previous studies address the issues of washback on test-takers without making link to the '*product of learning*' (Wall, 2000; Alderson & Banerjee, 2001; Green, 2013). The reason for such a phenomenon may vary from case to case. In the case of the NMET, one possible explanation is that most of the parents consider their children's achieved scores as a private matter and unwilling to provide them. Therefore, facing this situation, this study is taking a step forward by collecting participants' average scores on the monthly mock tests. Despite the undeniable differences between the average scores and the actual scores, under the circumstance, use of the mock test scores could be an effective alternative.

Moreover, as Cheng and Qi (2006) state, it is a big challenge for language testers developing the NMET as a test technique for both evaluating language ability and servicing as a selecting device in a vast country like China with millions of test-

takers. One of the contributing factors for this challenge is derived from the geographical variations within China. There are virtually a variety of washback contexts for the NMET in China, which requires individual exploration. To enhance the understanding of washback's complexity, more empirical studies in different contexts need to be carried out (Spratt, 2005; Shih, 2007, 2010; Green 2013). Studies on the NMET-related issues could be found conducted in Hubei (e.g. see Xiao & Carless, 2013), Heibei (e.g. see Andreas, 2004), Ganshu (e.g. see Hannum et al., 2011), Guangdong (e.g. see Qi, 2005; 2007), Jinan (e.g. see Zhang, 2013), Shanghai (e.g. see Zou, 1995) and so forth. This study was the first NMET-related empirical study as well as the first washback research conducted in *Jiangxi*. It provides solid evidence manifesting the washback effects of the NMET in this particular context and somewhat contributes to the general understanding of washback.

Chapter 4 Research Method

4.1 Introduction

This study applies a *survey* methodology using *quantitative questionnaires*. "*Survey* research is a quantitative research method which aims to collect self-report data from individuals, and the typical instrument used for this purpose is the written questionnaire…the basic idea behind the survey research is the recognition of the characteristics, opinions, attitudes, and intended behaviours of a

large population…" (Dörnyei & Csizér, 2012, P.75). Questionnaires as a research instrument have outstanding advantages in collecting factual, attitudinal and behavioural types of respondents from a sample population, which has been substantially applied in a wide variety of language-learning-related research. (Dörnyei, 2003). This methodology strongly aligns with the research purpose of this study, which is to examine any occurrence of negative washback through the identification of an overall attitudinal tendency to agree or disagree with specific statements. Moreover, participants in this study were young adult language learners who might lack of the professional and academic knowledge to verbally describe their attitudes toward the test and their behaviors of test preparation precisely and clearly. They might even take participation carelessly. Using questionnaires with pre-determined items could maintain a high quality of data and avoid unnecessary waste of research resources. For these reasons, questionnaires were judged to be a suitable method for this study in data collection.

4.2 Questionnaire Design

4.2.1 The Constructs of *Attitudes* and *Behaviors*

Attitudes and *behaviours* are both important constructs for this study. In the field of *Social Psychology*, *'attitude'* is "the most distinctive and indispensable concept" (Allport, 1935, p.798), but the usage of '*attitudes*' did not limited to the areas of

psychology. For example, in the area of *Second Language Acquisition*, the importance of learners' attitudinal characteristics on language learning was identified since the 1960s (Gardner, 1968). However, despite its popularity, the definition of '*attitudes*' is somewhat indefinite in regardless the field of psychology or other attitude-focused areas. As an abstract predisposition of human beings, the formation of '*attitudes'* connotes variations. According to Garrett (2010), the definitions of '*attitude*' vary in the "…degree of elaboration and in the weighting given to different features of attitudes" (p.19). Because of these variations, in many attitude-related studies, the applications of '*attitude*' concept normally depend on disciplinary inclinations and the specific purposes of individual studies. For instance, in the area of *Language Testing*, stakeholders' '*attitude'* is widely viewed as a significant element in the *test usefulness* and *construct validity* evaluations (Bachman, & Palmer, 1996; Messick, 1996; Elder & Lynch, 1996; Stricker, Wilder & Bridgeman, 2006; Cheng & Deluca, 2011). In these kinds of evaluations, researchers usually develop a hypothetical construct of '*attitudes'* with measurements and scales, such as *Likert* scales, to examine participants' attitudes explicitly. The most cited construct of '*attitude*' in this field consists of three components: *affective* (sympathetic nervous responses: verbal statements of affect), *behavioral* (overt actions: verbal statements about behaviors) and *cognitive* (perceptual responses: verbal statements of beliefs), known as the *ABC model of Attitudes* (Rosenberg & Hovland, 1960; Garrett, 2010).

However, this representation of the construct appears less suitable for the present study. This study aims at exploring a washback phenomenon through the overall attitudinal pattern from a sample population. This model of '*attitude*' is proved to be too complex to operationalize. This model requires a two-layer consideration when generating an overall attitude: the surface layer of 3-component and the second layer of duality underlying each component. For instance, regarding *attitude* as *cognition,* the second layer requires considering a distinguishable boundary between objects and social values, regarding *attitude* as *affect*, an explicit distinction between favorability and non-favorability is needed, and *attitude* as *behavior* requires the dual categorization of approval and disapproval actions. Thus, those sub-dualities basically increased the difficulty of generating an overall attitudinal tendency from this model, and so an alternative was sought.

Recently, Murray et al. (2012) have conducted a quantitative study with the purpose of investigating the attitudes of test candidates. In contrast with the *ABC* model, their model consists of 3 different components: *Beliefs*, *Opinions* and *Emotions* (see Figure 4.1). *Beliefs* refer to the dispassionate statements describing test-related status, such as 'English is the easiest subject to learn'. *Opinions* refer to the suggestions or descriptions concerning test design, language learning, test preparation and so forth with a hypothetical tone such as 'NMET should have a *speaking* section'. *Emotions* component is defined as the feeling or affective statements regarding the test, test impact or test preparation, such as 'I felt

stressed studying for the NMET'. The interaction and interconnection of these three formulate the construct of '*overall attitudes*' (see Figure 4.1). This model seems to be reasonably suitable for the current study. Firstly, it is less complex than the other one by requiring no sub-categories. Secondly, its usefulness has already been demonstrated in a study for collecting test candidates' attitudes. The third aspect of suitability lies in the common goal of obtaining an overall attitudinal tendency from the participants.

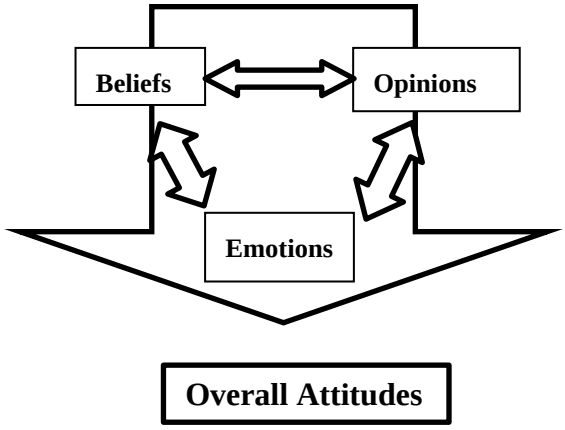


Figure 4.1 Overall Attitudes Model (Murray et al., 2012, p.583)

However, despite the suitability of this model, some of potential problems regarding the issues of applicability also require consideration. Firstly, although participants in both studies were candidates of a language test, they had strong demographic differences (see Table 4.1).

	Murray, et al. (2012)	Current Study
Occupation	Overseas Trained teachers	Local high-school Students
Countries of Origins	Australian Immigrants from a broad range of countries	All from the south non-coastal region of China
Age	Adulthood	Early adulthood
Language Background	Bilingual or Multi-lingual speakers with different kinds of mother tongues	English as Foreign Language Learners speaking Mandarin as mother tongue

Table 4.1 Demographical Differences of Participants between Two studies

It could not be assumed that attitudes would have the same meaning for participants in this study, or resolve into the same three sub-components. Therefore, it was decided that a factor analysis would be applied to the data in order to examine the applicability. The second concern is that, in Murray et al.'s (2012) study, attitudes regarding participants' behavioural patterns were not included. They used two strands of cognitive components: *Beliefs* and *Opinions*. However, *behaviours* are one of the central concerns of this study. In order to avoid confusion, when the questionnaire for this study was designed, the *behavioural attitudes* of this study were divided into two sub-components with the working labels of "*Behavioural Beliefs*' and '*Behavioural Opinions*'. Questionnaire items following these two notions were worded applying the definitions of '*Beliefs*' and '*Opinions*' accordingly with *behaviour*-related contents (see section 4.2.2).

In washback studies, the term '*Behaviors*' primarily pertains to the activities conducted prior to the target tests with the purposes of achieving satisfactory scores, or known as the '*Test Preparation Activity*' (TPA). Smith (1991) defines TPA as the 'meanings-in-action', which particularly refers to the human behaviours that "…is not a response to a fixed stimulus, but an active construction that is based on an interpretation or symbolizing of a particular situation" (p.525). Smith (1991) summarizes eight main categories of such human behaviours from literature as '*non special preparation*', '*teaching test-taking skills*', '*exhortation*', '*teaching the content known to be covered by the test*', '*teaching to the test*', '*stress inoculation*', '*practicing on the items of the test itself or parallel forms*', and '*cheating*' (p.526-536). One category is particularly relevant for this study: the '*practicing on items of the test itself or parallel forms*' (short as '*practicing'*). According to Matoush and Fu (2012), this type of TPA is surprisingly common in most of the preparation classes for high-stakes tests in China, in which students were asked to practice mock tests ('*parallel forms*'), stimulated tests ('*parallel forms*') and even used tests ('*the test itself*'). Most Chinese students appear to believe that this TPA would help them to be familiar with the test format and then effectively improve their test performances. According to Pan (2009), the TPA of '*practicing*' is listed as a negative washback by potentially causing the neglect of real-life knowledge (see Table 2.2). Thus, considering the popularity and

negativity, this study decided to apply the concept of '*practicing*' in investigating participants' test-preparing behaviors.

Besides '*practicing*', this study also applied '*cramming*' and '*test-wiseness'* as the investigating TPAs. The selection of these two was based on the analyses of the *socio-cultural factors* and the *test factors* in the prior chapter (see section 3.1 and 3.2.1). From the analysis of socio-cultural factors, this study found the need to take more of the traditional methods and strategies into account in the selection of investigating TPAs. As a typical representative of the traditional TPA, the criticism of '*cramming*' occurs as early as the 1870s (see 'an examiner', 1871). '*Cramming*' generally refers to the actions of hastily memorizing or learning by rote. In China, '*cramming*' is also a technical term for a traditional teaching and learning method, known as the '*stuffing-the-duck method*' (Jin & Cortazzi, 2006). In Pan's (2009) inventory of negative washback effects, '*cramming*' is regarded as one by inducing negative attitudes and demotivation (see Table 2.2) Therefore, '*cramming*' was applied in this study by not only qualifying as a traditional method but also being a widely criticized negative language-learning activity.

As one particular branch in the family of *test-taking strategies, test-wiseness strategies* have a long history of academic attention (Cohen, 2006). Early attention dates back to the 1970s within the fields of *Educational Measurement*

and *Psychology* (e.g. see Fueyo, 1977; Lange, 1981). The most quoted definition of it referring to 'a subject's capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a high score" (Millman et al., 1965, p.707). In washback studies, *test-wiseness* is widely argued to be a type of negative TPA because it is "…logically independent of the subject matter for which the items are supposedly measures" (Millman et al, 1965, p.707). Meanwhile, *test-wiseness strategies* are enormously popular in the contemporary test-preparing industries, especially in the cases of high-stakes language tests involving multiple-choice items such as the NMET in China, the TOEFL and the IELTS (e.g. see Yang, 2000; Mahdavi & Ghabelju, 2013; Tavakoli & Samian, 2014). For tasks with multiple-choices items, test-takers could choose the correct answers by finding clues or strategically guessing without using the testing language skills. As the analyses of *test factors* show, the NMET is a test with a large proportion of multiple-choice items and a high level of test stakes. It is reasonable to ask whether participants in this study have already overused the *test-wiseness strategies.* Thus, this type of TPA was also selected.

4.2.2 Constructing and Piloting the Questionnaire

Built upon the constructs of '*attitudes*' and '*behaviors*', the actual writing of the questionnaire items was guided by Dörnyei's (2003) procedures of effective questionnaire construction, which includes:

- "Deciding on the general features of the questionnaire, such as the length, the format, and the main parts;
- Writing effective items/questions and drawing up an item pool;
- Selecting and sequencing the items;
- Writing appropriate instructions and examples;
- Piloting the questionnaire" (Dörnyei, 2003, P.16-17)

The length of questionnaire needed to be short because of the limited amount of time participants of this study could spare. According to the research purposes, there was no necessity of obtaining participants' personal information, such as names or class numbers. Thus, the questionnaire was designed to be anonymous.

For the structure of the questionnaire, it was decided to include three parts: the information letter, the multi-choice items (*Part One*) and the 4-scale *Likert* items (*Part Two*). The wording of the information letter strictly followed the guideline from the Human Research Ethics Committee of Macquarie University with detailed information of researchers' identities, research purposes, background of this study and a clear clarification of the total voluntary participation (see *Appendix Four*). The multiple-choice items in *Part One* were applied to collect the information of participants' experiential characteristics and academic backgrounds with only categorical variables. The *Likert scales* in *Part Two* made

up the major part of this questionnaire, which was assigned with a 4-point scale, namely ‘strongly disagree’, ‘disagree’, ‘strongly agree’ and ‘agree’. The exclusion of a neutral scale (e.g. ‘neither agree nor disagree’) was decided according to the cultural preference found by many researchers that Chinese participants always attempt to select the neutral one in spite of having different opinions (Nunnally, 1978; Robson, 1993; Dörnyei, 2003). For studies inviting only Chinese participants like this one, such preference might be threatening to the validity of research outcomes. Besides, this research aims to identify the existence of a certain phenomenon, thus, a response of agreement or disagreement would be more meaningful.

In drafting an item pool, the screening of applicable information was guided by the construct of ‘*attitudes*’ and ‘*behaviours*’, the analyses of both *test factors* and the *context factors* and some informal conversations with students and English teachers from other local high schools. Moreover, this research aimed to examine the *Value* and *Intensity* of washback, and then all the items had to be worded with such reflection, but, in order to avoid a situation in which respondents mark only one side of the rating scale, some items reflecting positive washback effects were added. Hence, an additional step of reversing the responses of positive washback items was needed in the data analysis procedure. Table 4.2 shows the rationale for each item and their related washback dimensions. Items worded from different constructs were mixed up in the final questionnaire to create the sense of variety

and enable the checking of internal consistency (see Table 4.3). The final version of the questionnaire was in participants' first language: Chinese. The translation was conducted with the assistance of a Chinese native speaker who comes from the same region as the research site, with clear awareness of avoiding potential pragmatic miscommunication, and also to be easy to read and reader-friendly.

Item No.	Purposes	Washback Dimensions
1-5	To investigate washback's intensity on teaching.	Value, Intensity
6	To examine students' opinions about the uses of test scores.	Value
7, 9	To investigate the use of two types of TPA: *practicing* and *test-wiseness*	Value, Intensity
8, 10	To investigate the use of one type of TPA: *test-wiseness*	Value
11,12	To investigate the use of two types of TPA: *test-wiseness* and *practicing*	Value, Intensity
13	To investigate the use of one type of TPA: *practicing*	Value, Intensity
14, 28	To examine students' opinions on test reform	Value
15, 16	To investigate the use of one type of TPA: *test-wiseness*	Value
17	To investigate the use of one type of TPA: *cramming*	Value
18, 19	To investigate students' understanding of test requirements	Value
20-23	To examine students' attitudes towards test validity in making inferences of language abilities	Value
24-26	To investigate students' motivation triggered by the test	Value
27	To investigate washback's effect on the level of demotivation	Value, Intensity
29, 30	To examine students' overall attitudes	Value

Table 4.2 Rationales of Questionnaire Items

Construct	Item Number	Total number
Attitudes: *Beliefs*	1, 2, 3, 4, 5, 18, 19, 20, 21, 29, 30, *7, 8, 9, 11, 12, 13,*	17
Attitudes: *Opinions*	6, *10*, 14, *15, 16, 17*, 22, 23, 28	9
Attitudes: *Emotions*	24, 25, 26, 27	4
Total		30

Table 4.3 Item Distribution of the Attitudinal Construct (Italic numbers represent the *Behavioral Attitudes* items)

The pilot study was conducted during 13th to 18th, May 2014, and around 51 participants in one class were invited. The main purpose of this pilot study was to examine whether the questionnaire items were easy to read, whether any sensitive topics were involved, and whether the duration of completion was appropriate, and also to rehearse the administration procedures. The piloting class was randomly selected from the senior-year grade classes. This class did not participate in the following main data collection. In order to evaluate the questionnaire, a *Part Three* containing 4 multiple-choice items with open-ended questions was added in the end to collect constructive suggestions for improvement (See *Appendix Two*). 31 copies of questionnaire were collected with approximately 62% recovery rate. The results are presented in Table 4.4 and Table 4.5.

Questions	Total No. of Responses (Rates) & Translated Answers	
	Yes	**No**
1: are they clear and understandable?	29 (94%)	2 (6%): • 'Sentences are too long and too many complex sentences'; • 'Some sentences are difficult to read';
2: are they easy to interpret?	1 (3%): Item 27	30 (97%)
3: are there any sensitive topics?	2 (6%): Item14; Blank	29 (94%)

Table 4.4 Pilot Study Results of *Part Three*: Question 1-3

Question No.	Total No. of Responses (Response Rates)	Translated Answers
4: Suggestions	7 (22%)	Positive (or constructive): • 'Good enough' • 'Rewrite some of the sentences that is too complex to understand from the first sight'
		Negative: • 'Too long' • 'Boring' • 'I do not understand how this could help me' • 'Waste of my time'

Table 4.5 Pilot Study Results of *Part Three*: Question 4

The conduct of pilot study implies **three** constructive suggestions for revising the questionnaire. Firstly, to increase the recovery rate, the local research assistant suggested asking the grade dean to be present when the questionnaires were distributed, and then the students might take the participation more seriously. Secondly, some of the items, such as Item 11, Item 15, Item 16 and Item 27, were reported as less easy to read, which may cause by probable misplacements of sentential components during translation. Thus, these items were reworded. Thirdly, participants reported that Item 14 might involve sensitive issues. Therefore, this item was replaced with a different statement asking for opinions about the possible test form (revised Item 14: the NMET should have *speaking* section).

4.3 Participants

The participants of this study were all senior-year students from one high school in *Jiangxi*. For their language background, all the participants were English-as-a-foreign-language learners with Mandarin as first language. They all had been preparing for the NMET for more than 2 years. Besides, according to the prior establishment of the *school factors*, there are two class-types in this grade: *the class of Liberal Arts* and *the class of Sciences*. In order to avoid potential sample

bias, the study invited a similar number of students from each class-type. The total number of invited participants was around 310.

4.4 Data Collection

After being briefed on the general information of this research, the principal of this high school provided researchers with an oral consent for research-conducting in January 2014. During the ethics application, the official approval was obtained through email communication later in March 2014. The ethics approval was granted at 12th of May 2014. The pilot study was conducted in the following week and the modified questionnaire according to the pilot results was distributed in later May, 2014. Except the *information letter*, the revised questionnaire contained 3 pages: page one for *Part One*, page two and three for *Part Two* (see *Appendix Three*).

During the data collection, the questionnaires were distributed at *the self-study hours* (*'wanzixi'*). *The self-study hours* were a part of the school curricular, and were held in the evening normally from 7:30 pm to 10:00 pm. Students were asked to organize their own study during those times and teachers who attend the class as supervisors were required not to deliver any form of teaching. Therefore, completing the questionnaire during *the self-study* hours would not be considered to impinge on class time. The grade dean was asked to be present during

questionnaire distribution so that the students would take the participation more seriously. But the participation was still totally voluntarily. The dean was required only to be present, but not to join in any part of the distribution or collection or impose any kinds of potential coercion. Since the questionnaire was totally anonymous, no personal information was asked from participants. Participants could hand in the completed questionnaire right after completion or they could hand in the next morning. The completed questionnaires were collected by a local research assistant. Approximately 310 copies of questionnaires were distributed, and most of them were collected directly on the night of distribution and only were a few handed in the next day. 285 copies were collected with approximately 92% of recovery rate. All 285 copies of them were mailed directly after collection, and successfully arrived in Australia in June 2014.

4.5 Data Analysis Procedures

The software for quantitative data analyses used in this study was IBM SPSS 2.0. Data processed in the SPSS was totally in a numerical form, and each response was assigned to a code (see *Appendix five: the codebook*). A total of 7 steps were involved in the data analysis procedure:

a) In the coding for *Part One* items, the answers for *gender* were coded with '1=male' and '2=female'. Likewise, *the class types* were assigned with '1=classes of liberal arts' and '2=classes of sciences'. The status of test-taking

experiences were coded with '1=yes' and '2=no'. The four selectable answers of the following five categorical variables were codified from 4 to 1. For example, for the variable of '*how many years have you been taught English*', the answers were assigned with '4= 9 or more', '3= 7 to 8', '2= 5 to 6' and '1= 4 or less'. The coding for *Part Two* items was also straightforward: '1=strongly disagree', '2=disagree', '3=agree' and '4=strongly disagree' (see *Appendix Five*: *The Codebook*).

b) After saving the initial data into a SPSS file (.sav), the procedure of *data cleaning* was subjected, in which involved correcting typing errors, excluding the implausible data and checking the internal consistency. Blanks in the SPSS file were automatically recognized as missing data by the software. Therefore, the missing data in the *Part One* were not filled, while the missing data in *Part Two* variables were filled with *means* for inferential analysis. *Part Two* data used the Cronbach's Alpha as the index for examining the internal consistency. The result of Cronbach's Alpha is .786 indicating a high internal consistency.

c) In the stage of questionnaire design, some of the items were worded with the meanings of positive washback effects to avoid the potential bias of one-side answering. Before following steps, the responses of those items were reversed as '4 to 1', '3 to 2', '2 to 3', and '1 to 4'. The reversed items were Item 18, 19, 22, 23, 24, 25 and 29. Thus, after this procedure, all statements in *Part Two* were expressed as statements or agreements indicating a unidirectional

washback phenomenon.

d) Exploratory factor analysis was used in exploring the internal relationship with the *Likert* items. The specific technique applied was the *Principal Component Analysis* (PCA). Direct oblimin was chosen and applied as the method for factor rotation due to the advantage of needing no assumption that the rotated factors were uncorrelated. Factor scores were computed by calculating as the *means* of the items within each factor.

e) To particularly address Research Question 1 and 2, both descriptive and frequency analyses were performed for the *Likert* items and the factors that retained from PCA. An additional variable was added and labeled as '*Overall Attitudes*', and the scores in this variable were the *means* of the remaining items from the factor analysis.

f) To answer Research Question 3, a series of *t*-tests were conducted to examine whether there were statistically significant differences between groups of high-achievers and the low-achievers in the NMET in terms of those extracted attitudinal factors. Before the *t*-tests, an additional step was applied to merge the 4 categories in the variable of '*the average NMET mock test scores ranges*' (*Part One*: Item 5) into dichotomous categories, as the 'high-achievers' (merge '150-120' and '119-90') and the 'low-achievers' (merge '89-60' and 'lower than 60').

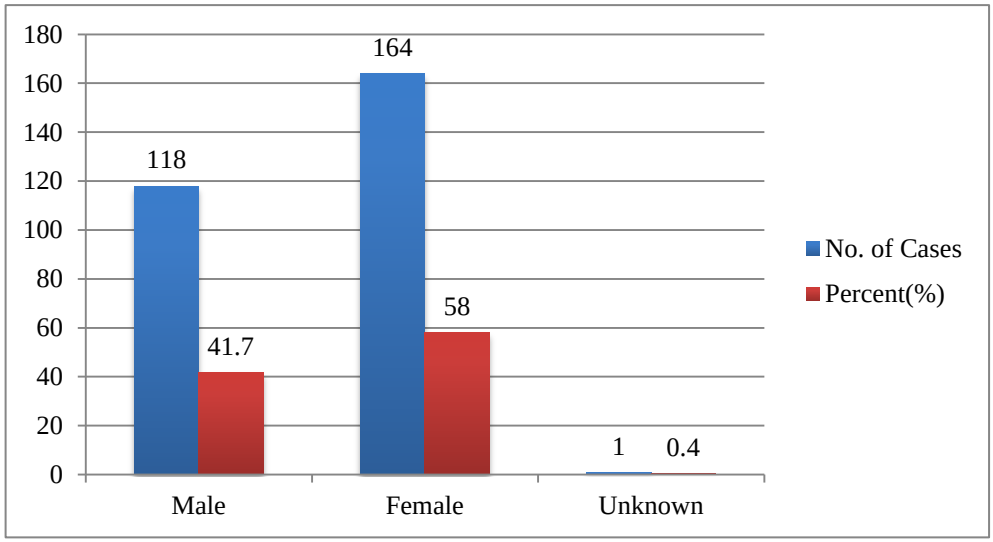
g) To further examine whether there were statistically significant differences among participants within four score ranges (independent variables) in term of

attitudinal factors (dependent variables), MANOVA was performed to avoid inflated Type 1 error instead of a series of one-way ANOVAs separately for each dependent variables. The suitability for MANOVA was assessed through several assumption tests, which included descriptive statistics for reasonable sample size, multivariate normality, linearity satisfaction, multi-collinearity checking, and homogeneity of variance-covariance matrices. One-way ANOVA was also applied later to further explore the significant differences by making multiple comparisons between groups.

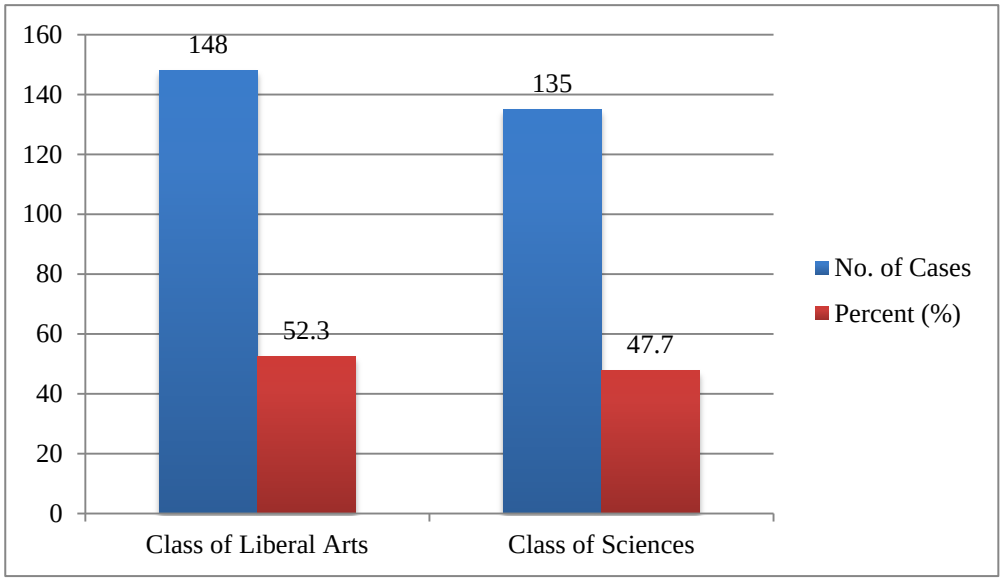
Chapter 5 Data Analysis

5.1 Descriptive Results

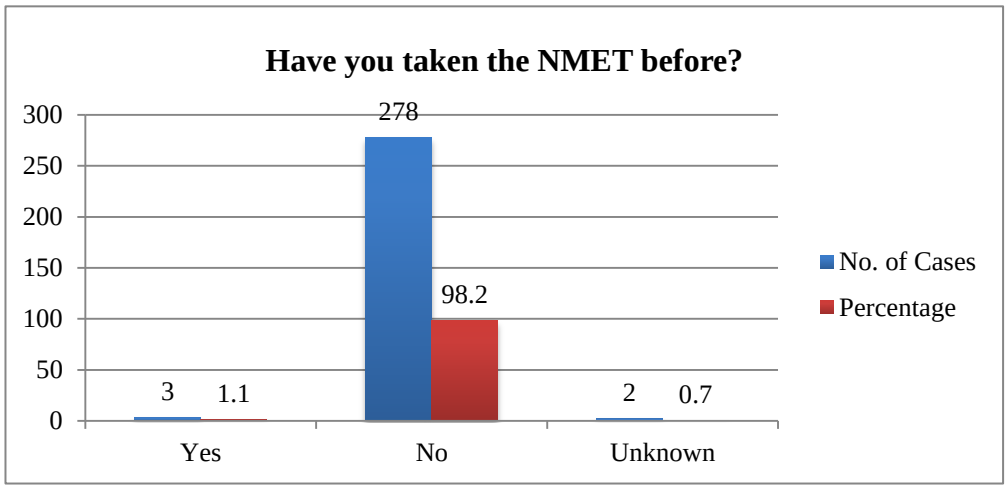
A total of 285 students participated in this study, in which 2 cases were considered as invalid. One of them selected multiple answers for each items (ID=116) and the other only had one type of selection (ID=267). These two were excluded from the following analysis. The remaining 283 participants consisted of 41.7% male (N=118) and 58% female (N=164) with one unknown case (see Table 5.1). Roughly equal numbers of participants came from the *Class of Liberal Arts* (N=148) and the *Class of Sciences* (N=135) (see Table 5.2). Only 3 out of 283 participants had taken the NMET before, and thus 98.2% of the participants were pre-test candidates (see Table 5.3).



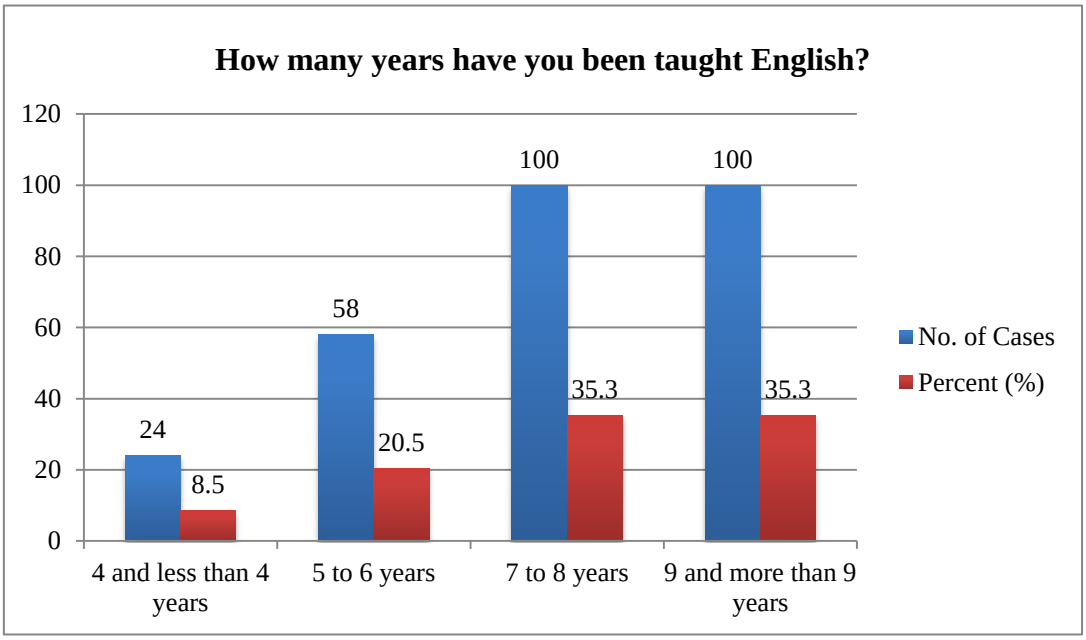
Bar 5.1 Gender



Bar 5.2 Class Types



Bar 5.3 The NMET Test-taking Experiences



Bar 5.4 Years of English Taught and Learning Experiences

Most of participants had been taught English for '*7 to 8*' (N=100, 35.3%) or '*9 or more than 9*' (N=100, 35.3%) years (see Table 5.4). This outcome is consistent with the *school factors* that most of current high-school students began to learn English at the 3th grade of primary school. Figure 5.1 shows that 15.2% of the participants' average scores on the monthly NMET mock tests were in the range of '*Below 60*' (N=43), 47% of them were in '*60-89*' (N=133), 31.1% in '*90-119*' (N=88) and 6.4% in '*120-150*' (N=18). After being connected, the line appears close to a positive skewed distribution (see Figure 5.1). Moreover, regarding participants' average scores on the monthly English exams since freshmen year, 10.6% of them (N=30) reported a generally decrease, 26.1% (N=74) had big fluctuation without changes overall, 59% (N=167) had small fluctuation with also no change overall and only 4.2% (N=12) had generally increased scores (see

Table 5.5). Thus, more than 75% of the participants reported that no big change, neither increase nor decrease, occurred on their test scores during their years of test preparation.

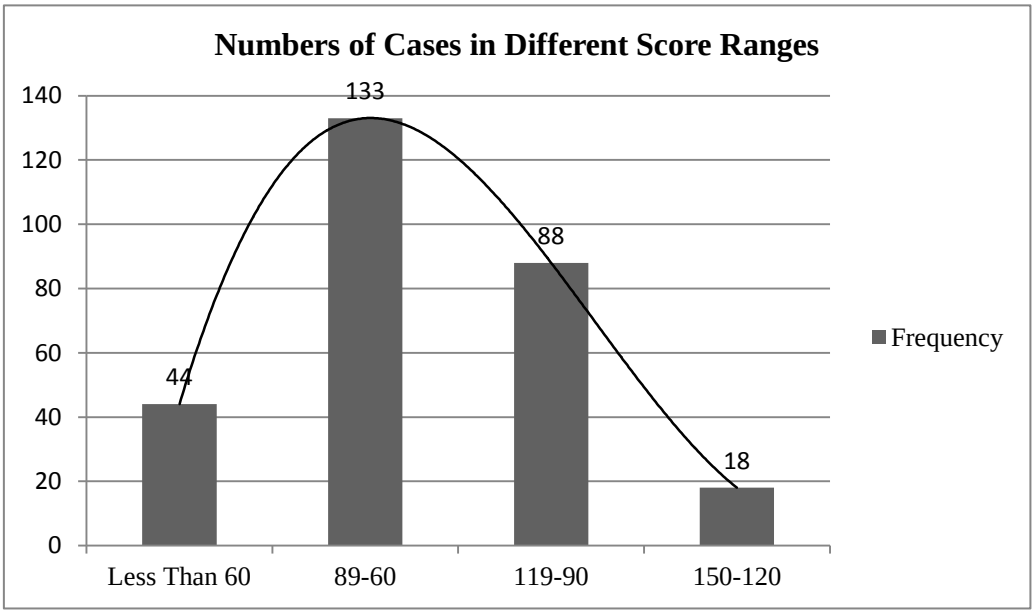
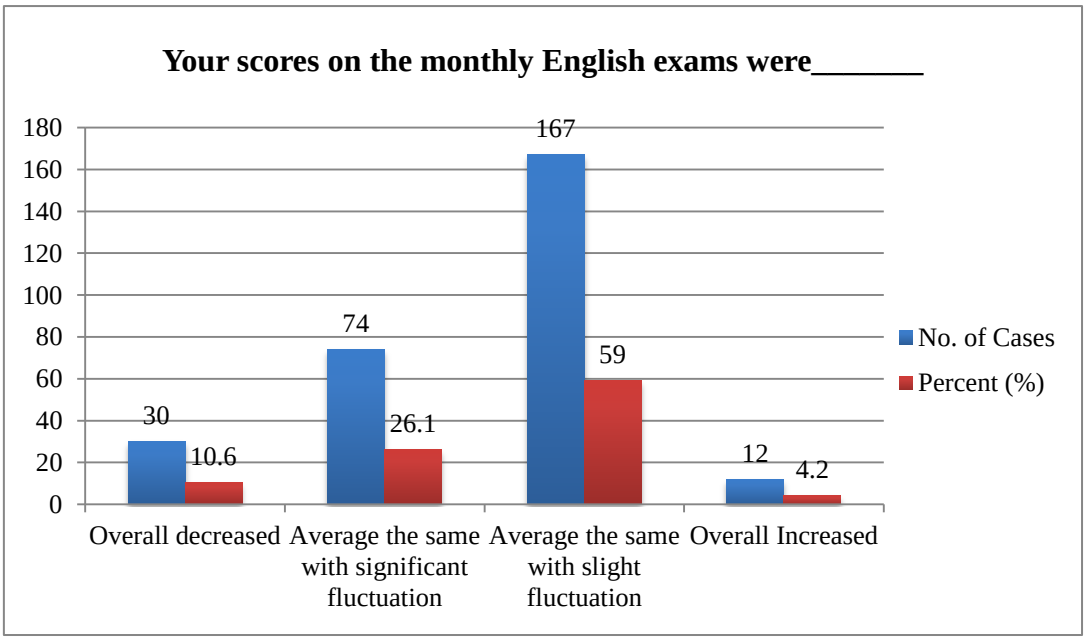
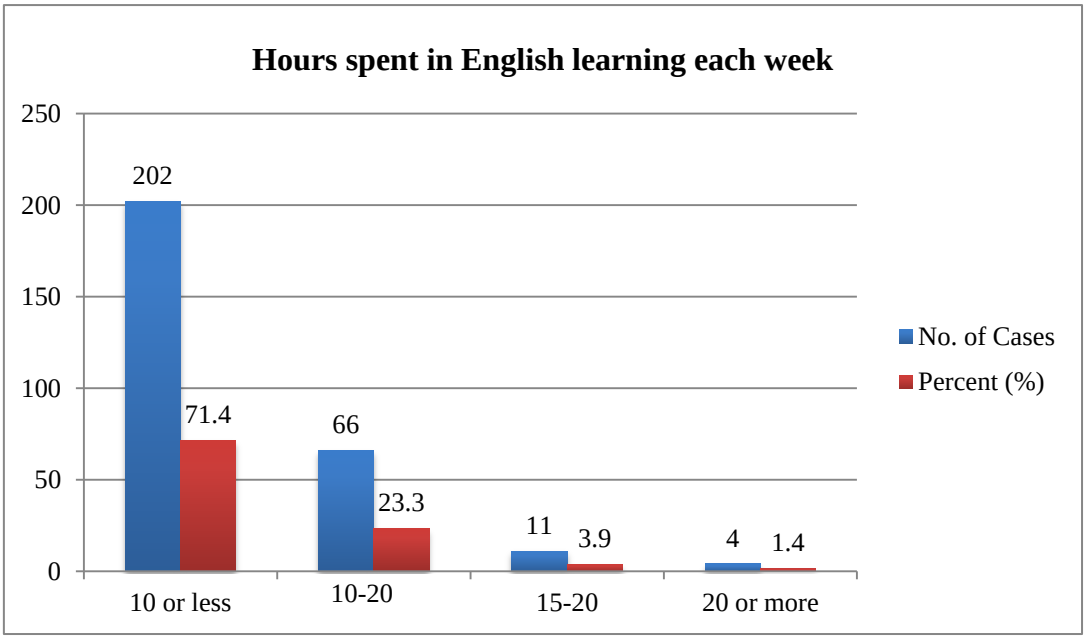


Figure 5.1 Population Distributions in Four Ranges of Average Scores in the NMET Monthly Mock Tests.

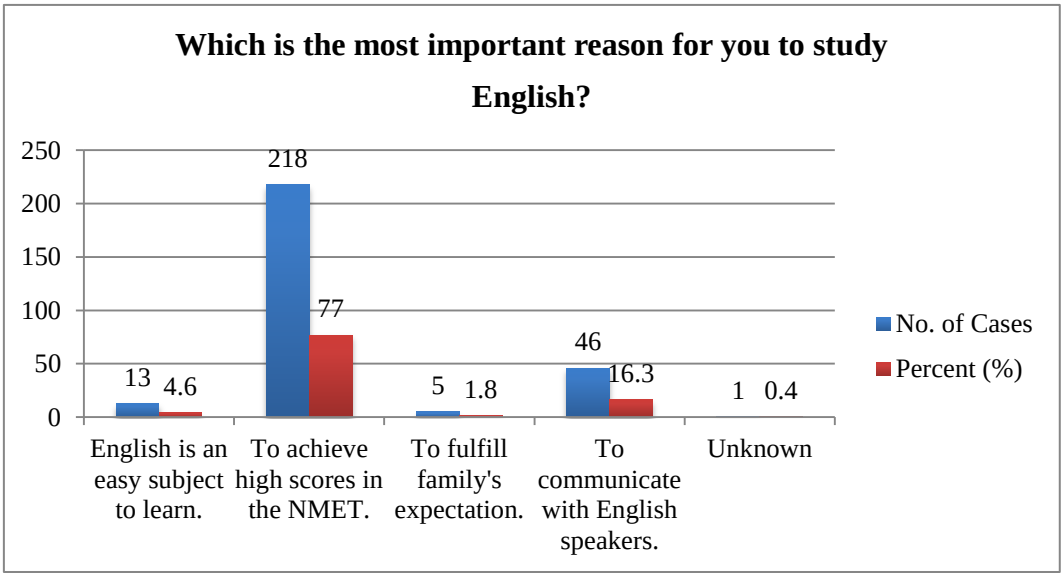


Bar 5.5 Participants' Average Scores Statuses in the Monthly English Tests

According to Table 5.6, 202 out of 283 students spent '*10 or less*' out-of-class hours in studying English. That means that close to 72% of students spent less than 1.5 hours daily on English-learning. This is an interesting result, which seems to be inconsistent with some of the *test factors*. Detailed discussion of this inconsistency is presented in the next chapter (see section 7.1). Furthermore, 218 out of 283 participants thought "*to achieve high scores in the NMET*" was the most important reason to study English comparing with other three (see Table 5.7). It implies a general *test-driven-motivation* for language learning. More discussion of this phenomenon is presented in the next chapter (see section 6.1).



Bar 5.6 Numbers of Hours Participants Normally Spent on English-learning



Bar 5.7 The Most Important Reason for Studying English

The following three tables demonstrate the results of frequency analysis (Table 5.8a and Table 5.8b) and the scale-level descriptive analysis (Table 5.9) of 30 *Likert* items. In Table 5.8a and Table 5.8b, the bolded numbers represent the ones having over 50% rates of selection. The italic items are the *behavioural attitudes* items. The signs of '+' in the end of some items represent that the total percentage of *Agree* ('agree' pluses 'strongly agree') is larger than the *Disagree* ('disagree' pluses 'strongly disagree') whereas the '-' account to the total percentage of *Agree* is less than *Disagree*. Due to the fact that Item 18, 19, 22, 23, 24, 25 and 29 were worded as the reflections of positive washback, the signs in these seven items were treated with the opposite ones. For example, the '+' sign in item 18 was considered as '-'. Likewise, the descriptive analysis for all 30 items was using the reversed coding scores of these seven. The new seven variables were labelled with a 'Re-' as prefix (see Table 5.9). The bolded rows in Table 5.9 are the items written particularly to reflect the washback's *Intensity* dimension.

Items	Strongly Disagree (%)	Disagree (%)	Agree (%)	Strongly Agree (%)
1，NMET-related contents are the dominant topic in English classes. +	23 (8.1%)	55 (19.4%)	**151 (53.4%)**	54 (19.1%)
2，Almost all the assigned exercises are NMET-related. +	28 (9.9%)	78 (27.6%)	114 (40.3%)	63 (22.3%)
3，English teachers mention NMET frequently during classes. —	31 (11%)	119 (42%)	79 (27.9%)	54 (19.1%)
4，Textbooks for the English subject are essentially the NMET preparation books. +	20 (7.1%)	91 (32.2%)	105 (37.1%)	67 (23.7%)
5，Content that is irrelevant to the NMET is seldom mentioned in the classes. —	31 (11%)	123 (43.5%)	79 (27.9%)	50 (17.7%)
6，NMET scores take too much percentage in university admission decision-makings. +	29 (10.2%)	64 (22.6)	80 (28.3%)	110 (38.9%)
7，*I practice multiple-choice tasks all the time.* —	37 (13.1%)	109 (38.5%)	101 (35.7%)	36 (12.7%)
8，*Memorizing the high-frequency words is enough for passing the NMET.* —	73 (25.8%)	**146 (51.6%)**	44 (15.5%)	20 (7.1%)
9，*Focusing on practicing reading exercises is an effective way to improve the overall score.* +	32 (11.3%)	100 (35.3%)	113 (39.9%)	38 (13.4%)
10，Sometimes, test-taking strategies are more important than actual language abilities. +	38 (13.4%)	63 (22.3%)	108 (38.2%)	74 (26.1%)
11，*My English learning methods will be modified constantly according the countdown to the NMET.* +	26 (9.2%)	96 (33.9%)	126 (44.5%)	35 (12.4%)
12，*Comparing reading and listening, hours spent in doing writing exercises are less.* +	38 (13.4%)	62 (21.9%)	122 (43.1%)	61 (21.6%)
13，*The main English-learning activity is practicing NMET stimulated exercises.* +	31 (11%)	97 (34.3%)	114 (40.3%)	41 (14.5%)
14，The NMET should have speaking section. +	45 (15.9%)	80 (28.3%)	98 (34.6%)	60 (21.2%)
15，Sometimes, I could choose the correct answer without comprehending the listening materials. +	18 (6.4%)	57 (20.1%)	**152 (53.7%)**	56 (19.8%)

Table 5.8a Frequency Results of the *Likert* Items (Item 1-15)

Items	Strongly Disagree (%)	Disagree (%)	Agree (%)	Strongly Agree (%)
16，Sometimes, I could choose the correct answer without comprehending the reading materials. +	17 (6%)	60 (21.2%)	**147 (51.9%)**	59 (20.8%)
17，*Memorizing the written models is an effective way to achieve high scores in the Writing section.* +	29 (10.2%)	44 (15.5%)	**161 (56.9%)**	49 (17.3%)
18，Teachers have explained the NMET rationale and requirements to me. + (-)	30 (10.6%)	111 (39.2%)	100 (35.3%)	42 (14.8%)
19，I am fully aware of the NMET's rationale and requirements. - (+)	65 (23%)	**161 (56.9%)**	44 (15.5%)	13 (4.6%)
20，'Being good English learners' does not equate 'achieving high scores in the NMET'. +	15 (5.3%)	38 (13.4%)	133 (47%)	97 (34.3%)
21，'Achieving high scores in the NNET' does not equate 'being good English learners'. +	19 (6.7%)	25 (8.8%)	133 (47%)	106 (37.5%)
22，The NMET is a valid test in making inferences of my language ability. — (+)	103 (36.4%)	139 (49.1%)	34 (12%)	7 (2.5%)
23，Studying for the NMET is directly related to my future needs in university. — (+)	87 (30.7%)	131 (46.3%)	58 (20.5%)	7 (2.5%)
24，The NMET motivates me to study harder. + (-)	54 (19.1%)	86 (30.4%)	114 (40.3%)	29 (10.2%)
25，Preparing for the NMET helps me developing confidence. — (+)	78 (27.6%)	129 (45.6%)	60 (21.2%)	16 (5.7%)
26，I am feeling under stressed preparing for the NMET. +	18 (6.4%)	63 (22.3%)	125 (44.2%)	77 (27.2%)
27，Sometimes, I felt that I simply wanted to give up English learning. —	59 (20.8%)	105 (37.1%)	63 (22.3%)	56 (19.8%)
28，I would prefer that the NMET could be taken more than once a year. +	47 (16.6%)	81 (28.6%)	90 (31.8%)	65 (23%)
29，Generally speaking, the NMET positively influences my studying and me. + (-)	49 (17.3%)	87 (30.7%)	113 (39.9%)	34 (12%)
30，Generally speaking, the NMET negatively influences my studying and me. —	45 (15.9%)	122 (43.1%)	56 (19.8%)	60 (21.2%)

Table 5.8b Frequency Results of the *Likert* Items (Item 16-30)

Descriptive Statistics				
	N	Sum	Mean	Std. Deviation
Reitem22	283	904.00	3.1943	.73973
Item 21	283	892.00	3.1519	.84313
Item 20	283	878.00	3.1025	.82520
Reitem23	283	864.00	3.0530	.78144
Reitem19	283	844.00	2.9823	.75539
Item 6	283	837.00	2.9576	1.01320
Reitem25	283	835.00	2.9505	.84490
Item 16	283	814.00	2.8763	.80484
Item 15	283	812.00	2.8693	.79930
Item 1	**283**	**802.00**	**2.8339**	**.82770**
Item 17	283	796.00	2.8127	.84020
Item 4	**283**	**785.00**	**2.7739**	**.89031**
Item 10	283	784.00	2.7703	.98584
Item 2	**283**	**778.00**	**2.7491**	**.91323**
Item 12	**283**	**772.00**	**2.7279**	**.94915**
Item 14	283	739.00	2.6113	.99153
Item 28	283	739.00	2.6113	1.01626
Item 11	**283**	**736.00**	**2.6007**	**.82057**
Reitem24	283	731.00	2.5830	.91248
Item 13	**283**	**731.00**	**2.5830**	**.86868**
Item 9	**283**	**723.00**	**2.5548**	**.86275**
Item 3	**283**	**722.00**	**2.5512**	**.92254**
Reitem29	283	717.00	2.5336	.91565
Item 5	**283**	**714.00**	**2.5230**	**.90820**
Item 7	**283**	**702.00**	**2.4806**	**.87649**
Item 30	283	697.00	2.4629	.99709
Reitem18	283	695.00	2.4558	.87153
Item 27	**283**	**682.00**	**2.4099**	**1.02876**
Item 8	283	577.00	2.0389	.83490

Table 5.9 Descriptive Statistics of *Part Two*: *Likert* Items

5.2 Exploratory Factor Analysis

After data cleaning, the suitability for a factor analysis was evaluated, in which includes steps of checking the sample size and examining the interrelationships among the items. Firstly, the sample size of this study (30 variables × 283 respondents) was found as agreeing to the suitability according to Stevens' (1996) recommendations of the reliability of factor structure and sample size requirements. Secondly, the examination of the correlation showed many coefficients of .3 and above (see *Appendix six: The correlation matrix*). As Table 18.0 indicates, the value for Kaiser-Meyer-Olkin Measure is .793, which exceeded the recommended value of .6, and the value of Bartlett's Test of Sphericity revealed statistical significances. Hence, the results of these two steps support the factorability. Then, 30 *Likert* items were subjected to a *principal component analysis* (PCA).

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.793**
Bartlett's Test of Sphericity	Approx. Chi-Square	2531.892
	df	435
	Sig.	**.000**

Table 5.10 Examination of Interrelationship among *Likert* Items

The preliminary PCA indicated 9 components having initial eigenvalue greater than 1, which explained 61.541% of cumulative percentage of variance. The following inspections of the scree test (Cattell, 1966) and the *Parallel Analysis* (30 variables × 283 respondents × 100 replications) revealed a better solution of 4 components. The scree plot showed a clear break after the forth component (see Figure 5.2), and the total initial eigenvalue in the variance explanation exceeded the random eigenvalue value in the *Parallel Analysis* till the fourth component.

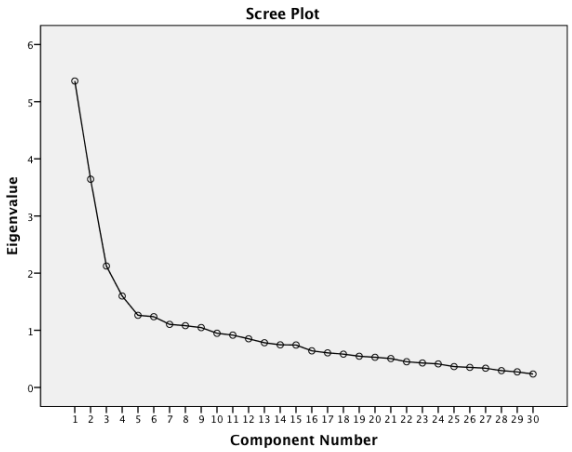


Figure 5.2 Scree Plot of the Initial PCA

Accordingly, a 4-component solution with oblimin method of rotation was performed resulting in a total explanation of 42.431% of the variance. However, the rotation did not reveal a clear and applicable item distribution. Several items had factor loadings in multiple components. For factor retaining, the exclusion of those items might be needed. The remaining items were insufficient in generating a convincing '*overall attitudes'*. Therefore, an additional 3-component solution

was subjected with a purpose of comparison. The following Table 5.11 and Table 5.12 demonstrate the pattern matrix and structure matrix for 4-component solution and 3-compononent solution respectively. The results showed that the 3-component solution explained overall 37.099% of variance with 5.332% lower than the 4-component solution, but the pattern matrix and the structure matrix revealed a clearer and simpler structure with many strong loadings and little cross-factor loadings. Besides, the items distribution in the 3-component solution was more applicable for this study due to the fact that the interpretation of this solution appears to be closely corresponding to the construct of attitude guiding the questionnaire design. As Table 5.13 show, this correspondence specifically refers to the same construct structure but with different item distribution. Therefore, the 3-component solution was retained, but Item 9 needed to be excluded in the following analyses because its factor loading was too weak to show. The new attitudinal model is displayed in Table 5.13.

4-component Solution	**Pattern Matrix (Component)**				**Structure Matrix (Component)**			
	1	2	3	4	1	2	3	4
Item 1		.620				.590		
Item 2		.709				.676		
Item 3		.755				.740		
Item 4		.734				.702		
Item 5		.632			.304	.612		
Item 6	.719				.724			
Item 7				-.401				
Item 8	.395	.307			.427	.409		
Item 9			.306				.344	
Item 10					.448	.316		
Item 11		.312				.431	.342	
Item 12							.350	
Item 13		.453				.553	.318	
Item 14			.314		-.373		.301	
Item 15			.704				.690	
Item 16			.716				.733	
Item 17		.476		-.366		.412	.370	-.413
Reitem 18		-.445	-.306			-.512	-.408	
Reitem 19		-.365				-.407	-.320	
Item 20			.652				.636	
Item 21			.597				.587	
Reitem 22				.623				.614
Reitem 23				.697				.689
Reitem 24	.585			.410	.622			.457
Reitem 25	.638			.424	.672			.481
Item 26	.702				.705			
Item 27	.747				.740			
Item 28			.334			.317	.317	
Reitem 29	.621			.336	.657	.729		.384
Item 30	.773				.788	.775		-.418

Extraction Method: Principal Component Analysis

Rotation Method: Oblimin with Kaiser Normalization

a. Rotation converged in 12 iterations.

Table 5.11 Pattern Matrix and Structure Matrix of the 4-component solution

3-component Solution	Pattern Matrix (Component)			Structure Matrix (Component)		
	1	2	3	1	2	3
Item 1		.633			.590	
Item 2		.646			.631	
Item 3		.667			.675	
Item 4		.627			.628	
Item 5		.535		.313	.541	
Item 6	.663			.678		
Item 7		.387			.359	
Item 8		.488			.491	
Item 9						.330
Item 10		.383		.317	.404	
Item 11		.454			.495	
Item 12		.316			.363	.315
Item 13		.569			.601	
Item 14			.357			.335
Item 15			.686			.688
Item 16			.691			.719
Item 17		.476			.515	.321
Reitem 18		－.455			－.506	－.381
Reitem 19		－.355			－.393	－.302
Item 20			.656			.649
Item 21			.587			.589
Reitem 22	.491			.487		
Reitem 23	.482			.472		
Reitem 24	.733			.732		
Reitem 25	.789			.780		
Item 26	.647			.662		
Item 27	.649			.663		
Item 28			.322			.316
Reitem 29	.732				.729	
Item 30	.765				.775	

Extraction Method: Principal Component Analysis

Rotation Method: Oblimin with Kaiser Normalization

a. Rotation converged in 10 iterations.

Table 5.12 Pattern Matrix and Structure Matrix of the 3-component Solution

Factors	Item numbers	Total number
Factor One: Emotional statements regarding the *scores use*, *scores interpretation*, and *motivation* (**Emotions**)	6, 22, 23, 24, 25, 26, 27, 29, 30	9
Factor Two: Dispassionate statements regarding *TPA*s and *classroom teaching* (**Beliefs**)	1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 17, 18, 19	14
Factor Three: Hypothetical statements regarding *test reform* and *TPA*s (**Opinions**)	14, 15, 16, 20, 21, 28	6
Total		29

Table 5.13 Structure of the New Attitudinal Model

The scores of those extracted factors were the *means* of associated items. The new variable of the '*Overall Attitudes*' was generated and its scores were calculated as the *means* of retained items in three factors (see Table 5.14).

	Minimum	Maximum	Mean	Std. Deviation	Variance
Factor1: *Emotions*	1.22	4.00	2.7852	.61294	.376
Factor2: *Beliefs*	1.50	3.86	2.6345	.38038	.145
Factor3: *Opinions*	1.00	4.00	2.8704	.50927	.259
Overall Attitudes	1.90	3.79	2.7301	.32605	.106

Table 5.14 Descriptive Statistics of the New Attitudinal Factors

5.3 Significantly Attitudinal Differences between Students in Different Ranges of Scores

5.3.1 *T*-tests

Four independent-samples *t*-tests (*df*=280) were conducted to compare the attitudinal differences between high-achievers (N=176) and low-achievers (N=106). One case (ID=199) was excluded from this analysis due to a missing answer on the test scores variable. The group of high-achievers contains students whose average scores on the NMET monthly mock tests were in the range of 90 to 150, whereas the group of low-achievers refers to students whose average score in the NMET monthly mock tests was below 90. The *t*-tests found significant differences between high-achievers and low-achievers in terms of the *Emotions* factor, the *Opinions* factor and the *Overall Attitudes* (see Table 5.15). The scores in the column of eta squared indicated the effect sizes. According to the guidelines proposed by Cohen (1988) for interpreting effect size values, .103 could be regarded as a largely moderate effect, while both .016 and .032 could be viewed as small effect. That indicates that the high-achievers and low-achievers regarding their emotional attitudes were differentiated most significantly.

Factors	Score Range	Mean	SD	Sig. (2 tailed)	eta squared
Factor One: *Emotions*	Low-achievers	2.5314	.56801	**> .000**	**.103**
	High-achievers	2.9381	.59102		
Factor Two: *Beliefs*	Low-achievers	2.6071	.36955	.350	
	High-achievers	2.6510	.38793		
Factor Three: *Opinions*	Low-achievers	2.9497	.49770	**.037**	**.016**
	High-achievers	2.8191	.51048		
Overall Attitude	Low-achievers	2.6545	.32151	**.003**	**.032**
	High-achievers	2.7749	.32204		

Table 5.15 Descriptive Statistics and Independent-sample *t*-tests with Attitudinal Scores between High-achievers and Low-achievers

5.3.2 MANOVA and ANOVAs

A one-way between-groups multivariate analysis of variance (MANOVA) was subjected to the SPSS to investigate the attitudinal differences among students in four ranges of scores ('*150-120*': N=18, '*119-90*': N=88, '*89-60*': N=133 and '*Below 60*': N=43). Three dependent variables were: the *Emotions* factor, the *Beliefs* factor and the *Opinions* factor. Assumption tests were conducted to check the normality, linearity, univariate and multicollinearity with no violations in the results. Then, after being subjected to the MANOVA, results showed the presences of significant differences in the combined variables with a *large* effect

size, F (3, 282) =7.23, $p \leq .000$; Wilks' Lambda=.80; partial eta squared=.072. Only the *Emotions* factor reached statistical significance applying a Bonferroni adjusted alpha level of .017(.05/3) (see Table 5.16 & Table 5.17). Thus, according to this result, a one-way ANOVA was performed specifically using the *Emotions* factor as the dependent variable to make further multiple comparisons between groups.

Source	**Dependent Variables**	**Type III Sum of Squares**	***df***	**Mean Square**	**F**	**Sig.**
The four scores ranges on the monthly NMET tests.	Factor One: *Emotions*	17.607	3	5.869	18.470	**.000**
	Factor Two: *Beliefs*	.829	3	.276	1.922	.126
	Factor Three: *Opinions*	1.630	3	.543	2.124	.097

Table 5.16 MANOVA: Tests of Between-Subjects Effects

Dependent Variable	**The Ranges Test Scores**	**Mean**
Factor One: *Emotions*	Below 60	3.222
	89-60	2.846
	119-90	2.595
	150-120	2.222

Table 5.17 MANOVA: Estimated Marginal Means of *Emotions* Factor

In the ANOVA using the *Emotions* factor as the dependable variable, results show interesting outcomes (F (4, 283) =18.470, $p<0.005$, eta squared=0.166). Post-hoc comparisons using the Tukey HSD test indicated that both group '*Below 60*' and '*89-60*' had significant differences with the remaining three, and no such difference was found between group '*119-90*' and '*150-120*'(see Table 5.18). The mean plots shown in Figure 5.3 appear as a nearly straight line. Since the variable of test scores was collected as a categorical variable rather than a continuous one, the data could not be submitted to a correlation analysis, but this result imply a potential linear relationship between test scores and the *Emotions* factor.

Test Score Ranges	**Test Score Ranges**	**Mean Difference**	**Std. Error**	**Sig.**
Below 60	89-60	.37594*	.09889	**.001**
	119-90	.62753*	.10489	**.000**
	150-120	1.00000*	.15825	**.000**
89-60	Below 60	-.37594*	.09889	**.001**
	119-90	.25159*	.07746	**.007**
	150-120	.62406*	.14157	**.000**
119-90	Below 60	-.62753*	.10489	**.000**
	89-60	-.25159*	.07746	**.007**
	150-120	.37247	.14582	.054
150-120	Below 60	-1.00000*	.15825	**.000**
	89-60	-.62406*	.14157	**.000**
	119-90	-.37247	.14582	.054

Table 5.18 ANOVA: Multiple Comparisons between Groups of *Emotions* Factor

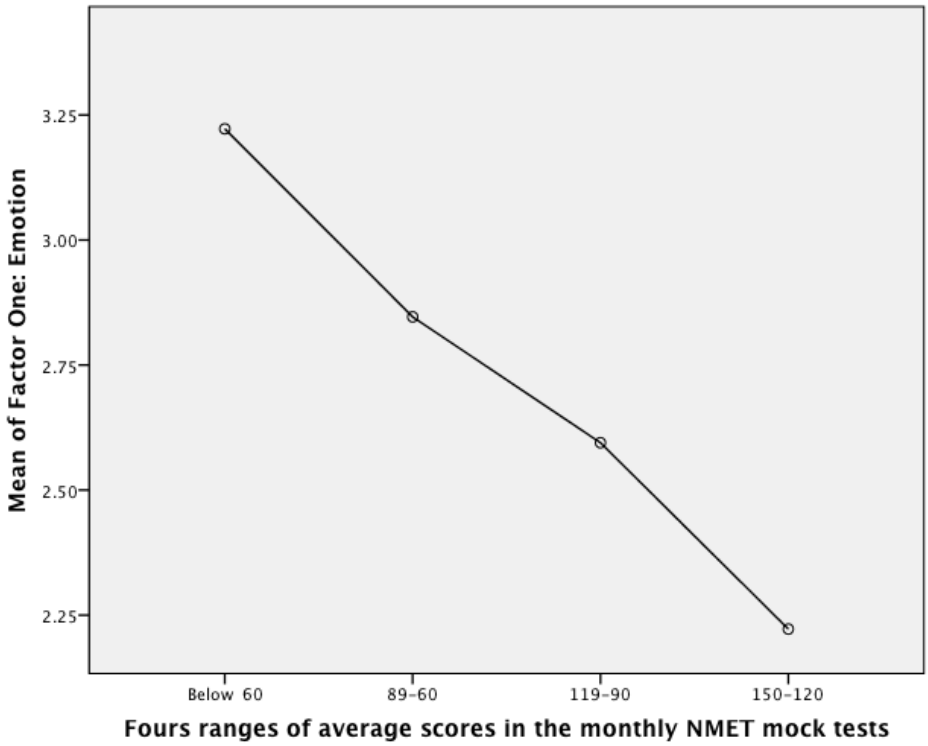


Figure 5.3 ANOVA: Mean Plots of the *Emotions* Factor in terms of the NMET score ranges

Furthermore, the one-way ANOVA on the *Overall Attitudes* factors shows rather different results (F (4, 283) = 8.575, $p<0.00$, eta squared=0.0085). According to the Tukey HSD test, both group '*Below 60*' and '*150-120*' were the ones having significant differences with the rest of three, and no such difference was found between group '*119-90*' and '*89-60*'(see Table 5.19). As the mean plots shown in Figure 5.4, generally, high score groups tended to have comparatively low attitudinal scores and low score groups had comparatively high attitudinal scores.

Test Score Ranges	Test Score Ranges	Mean Difference	Std. Error	Sig.
less than 60	89-60	.15888*	.05510	**.022**
	119-90	.20216*	.05844	**.003**
	150-120	.42751*	.08817	**.000**
89-60	less than 60	-.15888*	.05510	**.022**
	119-90	.04327	.04316	.748
	150-120	.26863*	.07888	**.004**
119-90	less than 60	-.20216*	.05844	**.003**
	89-60	-.04327	.04316	.748
	150-120	.22536*	.08125	**.030**
150-120	less than 60	-.42751*	.08817	**.000**
	89-60	-.26863*	.07888	**.004**
	119-90	-.22536*	.08125	**.030**

*. The mean difference is significant at the 0.05 level.

Table 5.19 ANOVA: Multiple Comparisons between Groups of '*Overall Attitudes*'

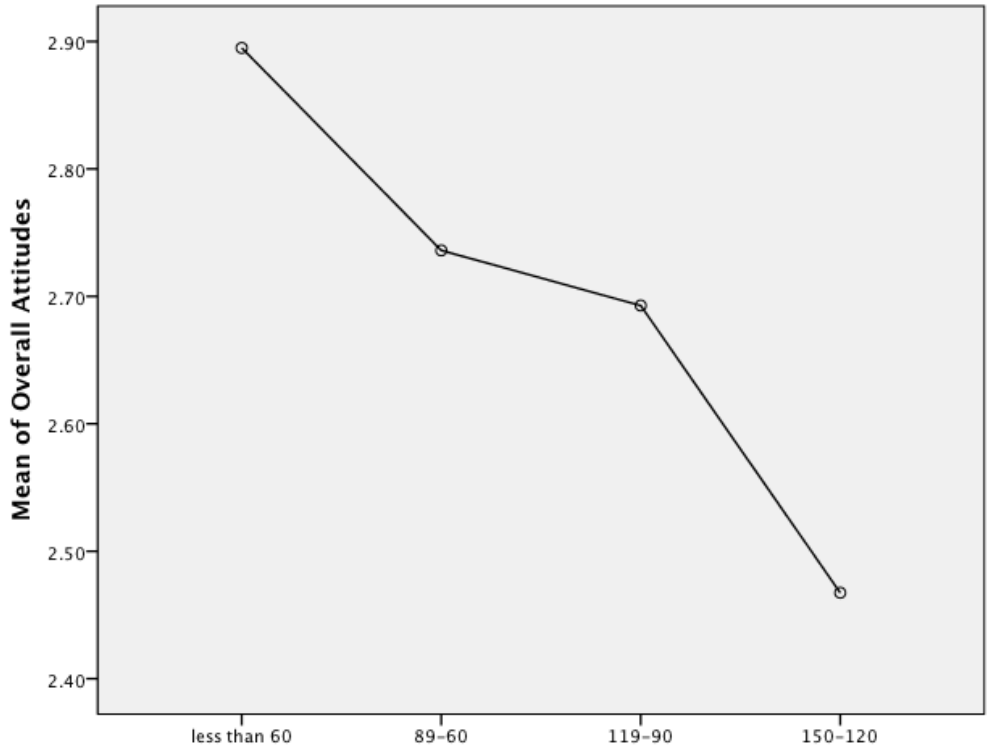


Figure 5.4 ANOVA: Mean Plots of '*Overall Attitudes*' in terms of the NMET score ranges

5.4 A Brief Summary

This chapter presents the detailed procedures and related findings of data analysis. For the analyses of *Part One* items, the descriptive and frequency results draw a basic background profile of the participating group. Most of the results are congruent with the characteristics identified in the analyses of the context factors and the research design, which include such as most of participants had been taught English for more than 7 years and roughly equal numbers of participants came from each class-type. Moreover, the descriptive and frequency of *Part Two* items reveals *an overall tendency to agree*. Most of the statements in the *Likert* were agreed by the participating group and some of them were strongly agreed. Furthermore, in the examination of the applied attitudinal construct guiding the questionnaire design, the factor analysis found a best solution of 3-component structure with the same sub-constructs: *Emotions*, *Beliefs*, and *Opinions* but with different item distribution. Applying this new construct, a new variable labelled the '*Overall Attitudes*' was generated. The statistical techniques applied to examine the relationship between attitudinal variables and the test-score variable (*t*-tests, MAVONA and AVONAs) indicate that the emotion-related attitudes had the strongest relationship with the test scores, and in general, students achieved within the high-test-score ranges attempted to have less negative overall attitudes, whereas students scored within low-test-score ranges, especially below 60, attempted to have more negative attitudes.

Chapter 6 Discussions

6.1 Research Question 1 and 2

- Research Question 1: "*Did students report that the NMET as a high-stake test have negative washback effects on students' attitudes toward the test, test preparation and English learning? If it did, how intense were the effects and why?"*
- Research Question 2: "*Did students report that the NMET as a high-stake test have negative washback effects on students' behaviors as test preparation activities (TPA)? If it did, how intense were the effects and why?"*

The results of data analysis are informative. First of all, the answer to the first part of Research Question 1 is **yes**. More negative washback effects than postive were found particularly on test-takers/students' attitudes toward *the test*, *test preparation* and *English learning in general*. According to the results, the frequency of responses to *Part One* Item 8 show that 77% of students believed that achieving high scores was the most important reasons for studying English comparing with '*to fulfil family expectation*', '*to communicate with English speakers*' and because '*English is an easy subject*' (see Table 5.7). Such a large percentage denotes the existence of a nearly unanimous *test-driven-motivation* for language learning. In accordance with the aforementioned detrimental effects of

high-stakes tests found in empirical studies (see Table 2.5), *test-driven-motivation* is found as negative by potentially causing terminations of continuous learning after the test being taken, in other words, candidates harbouring such motivation are less likely to become long-term English learners. Other evidence indicating the phenomenon of more negative than positive washback emerge from the scale-level descriptive analyses of both the '*Overall Attitudes*' (see Table 5.13) and the *Likert* items (see Table 5.8a, Table 5.8b & Table 5.9), in which both manifest *an overall preference to 'agree'*. As Table 5.13, Table 5.8a, Table 5.8b and Table 5.9 demonstrate, the minimum, the maximum and the mean scores of the '*overall attitudes*' variable are namely 1.90, 3.79 and 2.73, the range of mean scores of all 30 *Likert* items is from 3.1943 to 2.0389, and also 21 out of 30 *Likert* items had more people (≥50%) selecting 'agree' or 'strongly agree' than people selecting 'disagree' or 'strongly disagree'. Based on the coding system for questionnaire responses which is '1' for 'strongly disagree', '2' for 'disagree', '3' for 'agree' and '4' for 'strongly agree', these three batteries of descriptive evidence all support the overall preference to 'agree'. Thus, combining the fact that all *Likert* items incorporate negative aspects of washback effects, an overall preference to 'agree' substantiates the phenomenon of overarching negative washback.

Likewise, the answer to the first part of Research Question 2 is also **yes**. Research Question 2 particularly concerns the aspect of *Behavioral Attitudes*. The presence of negative washback effects was found particularly on the aspects of *how*

students learned English and *how they prepared for the test*. Firstly of all, for the usage of the three types of TPAs, data show 74.2% of participants agreed to the statement in Item 17 regarding the negative use of '*cramming'* in doing writing tasks, Item 7, 12 and 13 representing the excessive use of '*practicing*' were agreed by 47.9%, 64.7% and 54.8% respectively, and the overuse or misuse of the *test-wiseness strategies* was agreed by 22.6% (Item 8), 53.3% (Item 9), 64.3% (Item 10), 56.9% (Item 11), 73.5% (Item 15) and 72.5% (Item 16) with an average rate of 57.18%. Despite the statement in Item 8 which was weakly agreed, the overall tendency to 'agree' prevails. Thus, it is evidentially supported that the negative uses of *cramming*, *practicing* and *test-wiseness strategies* occur as the negative washback effects on behaviors in this case.

For answering the second part of both Research Question 1 and 2, data corroborate a **moderate** level of washback intensity. As prior established in questionnaire design, there were 11 *Likert* items entailing the indication of high washback intensity (see Table 4.2). The mean scores for those 11 items ranges from 2.40 (Item 27) to 2.83 (Item 1) (see Table 5.9), which is situated between the score of '2' as 'disagree' and '3' as 'agree'. Thus, high intensity washback seems to be less often reported. Furthermore, Item 7, 9, 11, 12 and 13 were worded in particular with indication of high washback's intensity on TPAs, and their means scores are 2.48, 2.55, 2.6, 2.72 and 2.58 respectively. These numbers support the overall preference to 'agree', but it is hardly a strong tendency, rather a moderate

one. This moderate level of intensity further indicates that the aforementioned assumption in respect of the positive correlation between high levels of test stakes attached and the significance of related test impacts is implausible in explaining washback phenomena (see sections 2.3.5.2 & 2.4.1 & 3.1.2). Increasing numbers of washback cases on high-stakes tests were found supporting such implausibility (e.g. see Shih, 2007; Li, 2009). Those studies, including the present one, strongly extrapolate that the current understanding in response to the relationship between test stakes and test impact is deficient in both elaborating theoretical explanation and guiding empirical applications. It is the time to question the sufficiency of using the dichotomous system as high-stakes tests and low-stakes tests in explaining the complexity of related washback phenomena and ask whether the relationship between the level of test stakes and the extent of test impact is positively correlated.

Meanwhile, despite the general moderate level, data show high intensity occurred particularly in response to the issues of *classroom teaching* (Item 1, 2 and 4). Item 1 ('*NMET-related contents are the dominant topic in English classes*') had the strongest support (72.5%) and also the highest mean scores (2.8339) among 11 intensity items. Item 4 ('*Textbooks for English subject are essentially the NMET preparation book*') had the second largest mean score as 2.7739 and an agreement rate of 60.8%. Item 2 ('*almost all assigned exercises are NMET-related*') had the third largest mean score and was supported by 62.6%. Those numbers may

sufficiently indicate an intense occurrence of the MDI, and also support one of the listed negative impacts of high-stakes tests: '*narrowing the curriculum down to test-oriented*' (see Table 2.2 & Table 2.5).

6.2 Research Question 3

The best solution resulting from the PCA shows close resemblances with Murray et al.'s (2012) model, which somewhat supports its applicability for this study. The new model is also a 3-component structure with same sub-constructs: *Beliefs*, *Opinions* and *Emotions* but slightly different item distribution (see Table 4.3 and Table 5.13). This outcome confirms the presumption that, due to some of the demographic differences (see Table 4.1), participants in this study and Murray et al.'s (2012) study might view the construct to some extent differently.

Using the new 3-component structure, the answer to Research Question 3, which is: '*was there a relationship between the ranges of students' average scores in the monthly NMET mock tests and their negative attitudes to the NMET*', is **yes**. The analyses result in **four** major findings. Firstly, when considering the independent variable with dichotomous ranges of test scores as 'high-achievers' (90-150) and 'low-achievers' (Below 90), statistically significant differences were found in three dependent variables: the *Emotions* factor， the *Opinions* factor and the *Overall Attitudes*, among which the *Emotions* factor had the largest effect size.

This result resonates with many previous testing studies and reviews which found a strong relationship between test performance and emotional factors or, as many researchers refer to them, the '*psychological factors*' (Harlen & Crick, 2003; Sundre & Kitsantas, 2004; Cheng & Deluca, 2011). For example, in Zhao's (2006) exploratory study on the relationship between Chinese university students' attitudes toward the CET Band 4 and their test performances, stepwise regression analysis showed significant impacts of students' emotional attitudes on their test performances, specifically the *attitudes of motivation*, *beliefs of the test*, and *test anxiety*. A similar result was also found in Cheng et al.'s (2014) recent examination of the interrelationship between motivational and test anxiety constructs on test performances of three different language tests across three distinctive contexts. Results strongly support the interrelationships between motivation and test performances, and between test anxiety and test performances. Their study also generated an important model of how *test-takers' characteristics* and *test events* (a similar concept to '*test status*' and '*test stakes*') interact with *emotional factors* leading to difference test performances in different contexts.

For the *second* major finding, when dealing with four ranges of students' scores: "*150-120*", "*119-90*", "*89-60*" and "*Below 60*", the *Emotions* factor was the only one out of three factors to result in significant differences. This result further supports the existence of a strong relationship between test performances and emotional factors in this context. Furthermore, in the multiple comparisons using

a one-way ANOVA between *four* groups in terms of the *Emotions* factor shows that both the group of '*Below 60*' and the group of '*89-60*' had significant differences with the other three but no such difference was found between the groups of '*119-90*' and '*150-120*'. Interestingly, the mean plots figure resulting from this analysis appears closely to be a straight line (see Figure 5.3), which could imply a linear relationship between students' test scores and their negative emotions. In order to verify or refute this preliminary finding, a follow-up investigation collecting students' test scores as a continuous variable is needed.

In contrast, thirdly, in the one-way ANOVA for the variable of the '*Overall Attitude'*, group '*Below 60*' and group '*150-120*' were the two having statistically significant differences with the other three and no such difference was found between group '*89-90*' and group '*119-90*'. Comparing these two ANOVA results, it might mean that scores '90' and '60' were the two breaking points in terms of students' negative emotions, while scores '60' and '120' were the breaking points of an overall negative attitude.

Finally, it is notable that the *Beliefs* factor was the only factor that resulted in no significant difference dealing with either two ranges of test scores or four ranges. Reviewing the definitions of the *Beliefs* factor, which refers to the dispassionate statements regarding issues of *test preparation in the classroom*, *TPA*s and *perceptions of the test's requirements and rationale*, it is likely that participants,

regardless of their test scores, shared a homogenous agreement of these beliefs. According to the descriptive analyses discussed in this chapter, items in the *Beliefs* factors relate to three specific negative washback effects: *strongly intense washback effects in classroom teaching*, *misuses and overuses of negative TPAs especially the common acceptance of test-wiseness's effectiveness* and *the lack of understanding in test requirements and rationale*. The data reveals no significant relationship between test performances and those three *Beliefs*, which means that these three specifically attitudinal washback effects were commonly accepted by students with various test-scoring backgrounds. Interestingly, such washback phenomenon could not be fully explained by any dimensions in Watanabe's (2004) five-dimension model for analyzing washback's complexity. Except for three of them which have been discussed as unsuitable or irrelevant for this research (see section 2.2.2), concerning the remaining two, neither *Value* nor *Intensity* used in this study as investigating targets could fully explain this phenomenon. It could be partially explained by *Value* because statements in the questionnaire referring to it were intentionally worded in negative terms. It could not be considered as *Intensity* because this phenomenon basically reflects the breadth of washback rather than depth or degree. One possible way of explaining of this phenomenon is using the concept of washback's *Scope*, which could be frequently seen in the literature as a term in referring to the multi-faceted nature of washback or its complex mechanism (e.g. see Cheng, 2004, 2005; Green, 2007; Pan, 2009; Aftab, et al., 2014). Nonetheless, this concept has not been well-defined in the existing

literature. It could be considered hypothetically as a new dimension of washback phenomenon referring to the breadth of its impact on a particular group of people, but, as Wall and Alderson (1993) point out that 'washback needs to be studied and understood, not asserted' (p. 69), and then more empirical exploration in the future is needed to further verify the occurrence of this dimension and examine to what extent it could explain washback's complexity.

6.4 Students' Perceptions on Test Validity

It is noted that test-takers' perceptions could usually provide abundant information for test validation (Cheng & DeLuca, 2011). In this study, participants' perceptions were pertinent with one specific type of test validity: *Face Validity*. *Face Validity* (FV) refers to "the degree to which a test appears to measure the knowledge or abilities in claims to measure, as judged by an untrained observer' (Davies, et al., 1999, p.14). The evaluation of FV does not require a deep analysis of how the test has represented the construct, but an overall judgment with regard to the impression. Gauging FV is somewhat important in test validation because it reveals what test 'looks like' from the point of view of non-professional observers, and in most testing communities, the majority stakeholders are comprised of people with low assessment literacy. For high-stakes test like the NMET, it is even more necessary to have a good public

credibility; otherwise, public criticisms in Chinese society with a massive population would certainly induce severe consequences.

The evaluation of FV in this study included two types of perceptions: an overall judgment of the test influences (Item 29 and 30) and whether the scores were valid in making proficiency inferences (Item 20, 21, 22 and 23). However, the results show contradictory outcomes. Selecting frequencies of Item 29 and 30 show that 51.9% of participants agreed that the test had generally positive effects on both themselves and their English learning, and 59% of them disagreed with the statement that the test had negatively impacted them and English learning in general (see Table 5.8b). These two numbers imply an acceptable degree of FV. It seems that most of participants praised and supported the role that the NMET had played in their lives, which seems to be a positive and reassuring signal for the test designers. However, on the subject of the other perception, 81.3% of participants agreed that 'being good English learners' did not equate to 'achieving high scores in the NMET' (Item 20), and 84.5% of them supported the corresponding statement with a reversed order (Item 21). Additionally, 85.5% of them disagreed that the NMET was a valid test making inferences of their language proficiency (Item 22) and 77% of them thought that studying for the NMET had no direct connection with their future needs (Item 23). These four percentages indicate that the NMET had actually low degree of FV, which is

incongruent with the result of the prior perception. One possible explanation for this incongruence may derive from the cultural influences. As stated before, China has a history with thousands of years using test taking to earn not only the recognition of intellectual prestige but the tickets to social superiority (see section 3.2.1). That may be why, despite acknowledging those negative aspects of the NMET, in the deep side of test-takers' perception, they followed the cultural traditions and accepted that the test did not negatively impact them in general. Nonetheless, the conclusive explanation of this paradox may need to hear from other groups of stakeholders, such as the parents.

6.5 Test-takers' Beliefs in the Effectiveness of Test-wiseness Strategies

As stated above, a large proportion of multiple-choice items and a high level of test stakes may cause the overuse or misuse of test-wiseness strategies (see section 4.2.2 and Table 2.5). The following results somewhat confirms this likelihood. Both *Listening* and *Reading* sections in the NMET contain only multiple-choice tasks, as the analyses of Item 15 and 16 show, nearly 74% of participants agreed that they could select the correct answer without comprehending the listening or reading materials, and even the *Writing* section had also fallen victim to the misuse of test-wiseness strategies (Item 17 in Table 5.8b). Additionally, there were 64.3% of participants agreeing that the *test-wiseness strategies* were more important than the actual language skills (Item 10

in Table 5.8a), which indicates a strong awareness of distinguishing between the uses of *test-wiseness strategies* and the application of actual language skills. In other words, there were still widely negative uses of the *test-wiseness strategies* among participants in spite of being conscious that it may lead to 'inflated score gains'. Such phenomenon could be somehow triggered by the high-stakes benefits brought by achieving high scores in the NMET, such as being admitted to top-class universities, receiving social recognition and so forth. Actual language acquisition seems to be less important compared with the huge benefit brought by score gains.

More importantly, this study found that the *test-wiseness strategies* participants had applied might not be as effective as they believed. The responses' frequency of *Part One*: Item 5 show that only 4.2% of the participants had overall increased scores in the nearly three years of test preparation study, and 85% of them had the average same scores with either big or small fluctuation (see Table 5.5). These percentages could connote that either the *test-wiseness strategies* were too difficult to master or they were simply not effective in score gains. Ascertaining which of these possibilities was the more influential factors requires further investigations because the relationships between the uses of test-taking strategies and test performances could be more complicated than what has been found (Song & Cheng, 2006). Despite that, the current evidence is sufficient to cast doubts on

the effectiveness of *test-wiseness strategies* that participants used in improving scores, and also encourage participants to reconsider their current practices.

Chapter 7 Conclusions

7.1 Implications for Testing and Pedagogical Practices

The results of this study lead to several suggestions for both test developers and English teachers. Firstly, as revealed in the result of Item 19, there were 79.9% of participants though that they were not fully aware of the test requirements and rationale (see Table 5.8b). That could indicate that students were preparing for the test without a clear focused orientation, which make them even more susceptible to ungrounded *test-wiseness strategies* for score gains. Moreover, only 50.1% of participants agreed that teachers had provided them with a full explanation of test requirements and rationale (Item 18 in Table 5.8b). It is reasonable to ask *why* only around 50% of students were instructed by teachers and even then most of them did not fully understand, and *why* only teachers of those 50% provided such kind of instructions and the rest did not. The first '*why*' indicates an instructional failure regarding the method of delivering and explaining test-related information. The second '*why*' unveils the possible presence of a discrepancy or lack of disciplinary communication within the English teachers' community. Both potential problems should not be ignored and require the collaboration of test developers and teachers to overcome. As Fan and Jin (2013) point out, in modern

China, stakeholders outside the examination board are often kept '*in the dark*' of the professional standards of test developments and the quality of such test. Thus, to ensure the test quality and improve the professional practices in language testing communities, test developers should be obligated to provide clear instructions and comprehensible explanations to the public and teachers within the same disciplinary community and the same working context ought to exchange the pedagogical ideas related to what should be instructed and how to undertake it effectively.

Secondly, data shows that, excluding class hours, most students spent 10 or fewer hours on studying English each week, which means average less than 1.5 hours a day were spent in studying English (see Table 5.6). This number appears to be divergent with the significance of the NMET, which is a high-stakes test, one of three compulsory subjects, and a determinative factor in their lives in a highly competitive society. For a test with such significance to attract so little devotion of time from the test-takers appears to be unreasonable. Such inconsistency could not be accurately explained by guessing. It requires some in-depth investigations finding out how exactly students managed their time, but it should be enough to draw local educators' attention to the arrangement of school courses and the rationality of students' autonomy learning.

7.2 Limitations of the present research

The present study is limited in two aspects. Firstly, this study is limited by the scope of data. Triangulated mixed method with both quantitative and qualitative data, collecting from classroom observations, interviews and questionnaires from multiple types of stakeholders, is the most commonly suggested methodology in investigating washback phenomena. Ideally, the complex nature of washback requires such holistic perspective and dynamic documenting. However, due to the limitation of resources and research duration, the present study could only collect one type of data from one type of stakeholders, which might not present a complete picture of washback. For example, from both the exploration of literature and the analyses of context factors, three types of TPAs: *practicing*, *cramming* and *test-wiseness*, were selected as particularly suitable for this investigation. However, due to the limited time for participants to complete the survey, many issues regarding these TPAs could not be covered by the questionnaire. The data pool would be richer with some follow-up surveys or interviews. It would be worthwhile to extend the present study into a longitudinal study, especially during the time of a coming top-down nationwide test reform. Secondly, this study was also somewhat limited because of its use of the convenience sampling method. This washback investigation was conducted in one high school. A wider-scale sampling with participants from different schools could provide additional information to the issues that could not be fully explained in this study, such as the paradoxical perceptions regarding FV and

students' time management. These issues could be better explained by a washback investigation with a larger scale.

7.3 Suggestions for Future Research

Generally, this study reveals **two** main suggestions for further research. Firstly, this study further emphasizes the significance of hearing the learners' voice in understanding the phenomena of washback and advocates more such hearing. As a core type of stakeholders in the language testing community, *learners/test-takers'* perspectives provide researchers with a unique insight to how the test produces influences on them. Taking this study as an example, the contradiction revealed in the evaluation of test FV highlights a noteworthy question of how test-takers themselves thought about the test's impacts. Regardless of the drawbacks, most of participants still believed the NMET had positively influenced them in general. Further studies could aim at exploring the potential causes of this somewhat paradoxical perception by hearing from different types of stakeholders in the same or different contexts. Another aspect reflecting the importance of gathering students' perceptions lies in participants' clear awareness of the potential negative consequence of using *test-wiseness strategies*: the 'inflated score gains'. In other words, to some extent the participants of this study preferred the 'inflated score gains' to actual language acquisition. This finding potentially raises an important question to modern language testers: what would happen if

more and more professional test designers and researchers focused on developing communicative language tests which may bring beneficial washback to test-takers' actual language acquisition, but the majority test-takers themselves still preferred 'inflated score gains'? Such potential contradiction between learners' and test designers' preferred outcomes requires more empirical attention on gathering test-takers' attitudes toward the test-related issues and identifying what 'positive washback' is to them.

Secondly, this study also points out a need of developing a systematic framework to guide the analyses of the '*test factors*', especially the '*test stakes*' factor, in washback empirical investigations. As state above, it is noticed that the gradual accumulation of individual washback studies involving high-stakes tests are calling for a systematic technique for analyzing the concept of '*test stakes*' in order to synthesise various empirical findings to create a coherent body of knowledge. The comparisons between high-stakes tests in terms of their different test stakes levels are the basis of such synthesis. For examples, in Shih's (2007) washback investigation on the GEPT (General English Proficiency Test), only *little* and *indirect* washback effects were found on test-takers' learning, which appears to be somewhat surprising and inconsistent with the popular assumption regarding the positively correlated relationship between high-stakes tests and significant test impact. After considering the influential factors, Shih (2007) argues that this test fundamentally entailed an insufficiently high level of test

stakes to produce significant test impact. Whereas, in the washback investigation of the NMET by Qi (2007), the findings reveal that the two intended purposes of the NMET, which are (a) as a selection device and (b) to make inferences of language proficiency, are actually conflicting with each other and have brought unintended washback effects to the related English education. Both as high-stakes tests, the test stakes level of the GEPT is too low to make significant changes, while the test stakes level of the NMET is too high to make intended changes. Thus, using the concept of '*test stakes*' provides a different but substantial explanation to the contradictive findings in both studies and more importantly, could enhance the general understanding on washback phenomena.

Currently, the *school-based assessment* or *classroom-based assessment* as new means to inhibit the negative washback effects of high-stakes tests has been frequently mentioned in the literature (Cheng, 2014; Xiao, Sharpling & Liu, 2011; Jin, 2014; Yu, 2010; Wall, 2012). According to Stoneman's (2006) model of '*test status*' (see section 2.4.1), this type of assessment achieves the substantial inhibition of negative washback through the reduction of test stakes' level because it fundamentally limits the '*extent of implementation*' to only schools or classrooms. Thus, the investigation of washback on this type of assessment also requires a rather explicitly analysis of test stakes' levels and the consideration of how the change of such levels links with the minimizing of negative washback.

7.4 A Summing up

This study sought to explore the washback phenomena of a high-stakes English test in China from the perspectives of test-takers' attitudes and test-preparing behaviours. Quantitative data collected from nearly 300 questionnaires was analysed using both descriptive and inferential statistical techniques. The main findings are:

- Results show positive presences of **negative** washback effects on test-takers' attitudes toward the test, test preparation, English learning in general and their use of test-preparing activities.
- Evidence also supports the existence of intense washback effects. However, the degree of such intensity was to some extent **moderate**, which was a somewhat surprising finding when the investigating test had an **extremely** high level of test stakes. Thus, the study suggests that it is the time to question the sufficiency of using the dichotomous system as high-stakes tests and low-stakes tests in explaining the complexity of related washback phenomena and also reconsider the assumed relationship between the level of test stakes and the extent of test impact.
- This study also examined the *Face Validity* (FV) from test-takers' perceptions on how the test influenced them in general and whether the test scores were valid in making inferences of their language proficiency. Results indicate a **paradox** that the test had an acceptable degree of FV in

general but it met with strong disapproval regarding its validity as a language proficiency test. This study argues this paradox may be caused by cultural influences, but a more conclusive explanation needs further investigations comparing this paradox with other types of stakeholders' perceptions on the same issues.

- Factor Analysis showed reasonable level of applicability of *The 'Overall Attitudes' Model* for analyzing test-takers' attitudes in this study (Murray et al., 2012). It may also be applicable for future studies with similar purposes.
- This study found a positive relationship between test-takers' attitudes and their presumed test performances, and the strongest one liay in the **Emotions** factor, namely test-takers' negative emotions on the *scores use*, *scores interpretation*, and *motivation*. Low-achievers tend to have more negative emotional attitudes than high-achievers. This finding resonates with many previous studies finding the strong relationship between emotional attitudinal and test performance.
- Results show discrepancies of communication among this testing community. To over such problem, this study suggests that *test developers* ought to provide clear instructions and comprehensible explanations regarding test requirements and rationale, and t*eachers* within the same disciplinary community need to communicate more to provide test-takers with effective instructions and unified explanations of test-related

information.

- This study addressed the complexity of washback through the dimensions of '*Value'* and '*Intensity*' selected from Watanabe's model. However, one of the phenomena this study revealed could not be explained fully by any dimensions in this model. Thus, this study suggests considering the '*Scope*' dimension to explain the breath of washback's complexity and proposes further exploration on the occurrence of this dimension in furture studies.
- For further research, this study emphasizes the significance of hearing the learners' voice in understanding the phenomena of washback and advocates more such studies in the future.
- For further research, this study proposes a need for developing a systematical tool for analyzing the degree of '*test stakes*' in order to enhance understanding of the washback phenomena and synthesis the diverse empirical findings on washback issues into a more coherent body of knowledge.

References List

"An Examiner" (1871). Cramming for Examination. *NATURE, 13*(July), 202.

Aftab, A., Qureshi, S., & William, I. (2013). Investigating the Washback Effect of the Pakistani Intermediate English Examination. *International Journal of English and Literature, 5*(7), 149-154.

Alderson, J. C. (2004). Foreword. In: Cheng, L., Watanabe, Y. & Curtis, A. (Ed.), *Washback in Language Testing* (pp.ix-xii). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Alderson, J. C., & Banerjee, J. (2001). Language Testing and Assessment (Part I). *Language Teaching*, *34*(04), 213-236.

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*(2), 115-129.

Allport, G. (1935). Attitudes. In C. Murchison (Ed.), *A Handbook of Social Psychology* (pp. 798-844). Worcester, MA: Clark University Press.

Andreas, J. (2004). Levelling the Little Pagoda: The Impact of College Examinations, and Their Elimination, on Rural Education in China. *Comparative Education Review*, *48*(1), 1-47.

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting Washback—a Case Study. *System*, *30*(2), 207-223.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly: An International Journal*, *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.

Bailey, K. M. (1996). Working for Washback: A Review of the Washback Concept in Language Testing. *Language Testing, 13*(3), 257-279.

Cai, Y. (2006). *Is the College English Test Band Four Promoting or Impeding College Students' English Language Learning? Findings from Classroom-based Research at a University in Central China.* (Unpublished master's thesis), Memorial University of Newfoundland, Canada.

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cheng, L. & Curtis, A, (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In: Cheng, L., Watanabe, Y. & Curtis, A. (Ed.), *Washback in Language Testing: Research Contexts and Methods* (pp.3-18). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cheng, L. (1997). How does Washback Influence Teaching? Implications for Hong Kong. *Language and Education, 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. (1999). Changing Assessment: Washback on Teacher Perceptions and Actions. *Teaching and Teacher Education, 15*, 253-271.

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: Lawrence Erlbaum Associates, Inc.

Cheng, L. (2005). *Changing Language Teaching through Language Testing: A Washback Study* (Vol. 21). Cambridge: Cambridge University Press.

Cheng, L. (2008). The Key to Success: English Language Testing in China. *Language Testing, 25*(1), 15-37.

Cheng, L. (2014). Consequences, Impact and Washback. In J. A. Kunnan (Eds.), *The Companion to Language Assessment: Evaluation Methodology and Interdisciplinary Themes (Vol. 3)*. UK: John Willey & Sons, Inc.

Cheng, L., & DeLuca, C. (2011). Voices from Test-Takers: Further Evidence for Language Assessment Validation and Use. *Educational Assessment, 16*(2), 104-122.

Cheng, L., & Qi, L. (2006). Description and Examination of the National Matriculation English Test. *Language Assessment Quarterly, 3*(1), 53-70.

Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and Test Anxiety in Test Performance Across Three Testing Contexts: the CAEL, CET, and GEPT. *TESOL Quarterly, 48*(2), 300-30.

Cohen, A. D. (2006). The Coming of Age of Research on Test-Taking Strategies. *Language Assessment Quarterly, 3*(4), 307-331.

Cohen, J. W. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davey, G., Lian, D. C., & Higgins, L. (2007). The University Entrance Examination System in China. *Journal of Further and Higher Education, 31*(4), 385-396.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., and McNamara, T. (1999). *Dictionary of Language Testing* (Vol. 7). Cambridge: Cambridge University Press.

Dörnyei, Z. (2003). *Questionnaires in Second Language Research: Construction, Administration, and Processing*. Mahwah: Lawrence Erlbaum Associates, Inc.

Dörnyei, Z., & Csizér, K. (2012). How to Design and Analyze Surveys in Second Language Acquisition Research. In A. Mackey & M. S. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (Vol. 11, pp. 74 - 94): Blackwell.

Elder, C., & Lynch, B. (1996). Public Perceptions of Basic Skills Tests and Their Ethical Implications. Paper presented at the 20th Language Testing Research Colloquium, Monterey.

Fan, J., & Jin, Y. (2013). A Survey of English Language Testing Practice in China: The Case of Six Examination Boards. *Language Testing in Asia*, *3*(1), 1-16.

Frederiksen, J. R., & Collins, A. (1989). A Systems Approach to Educational Testing. *Educational Researcher, 18*(9), 27-32.

Fueyo, V. (1977). Training Test-taking Skills: A Critical Analysis. *Psychology in the Schools, 14*(2), 180-184.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book.* Canada: Routledge Applied Linguistics.

Gardner, R. C. (1968). Attitudes and Motivation: Their Role in Second-language Acquisition. *TESOL Quarterly, 2*(3), 141-150.

Garrett, P. (2010). *Attitudes to Language.* Cambridge: Cambridge University Press.

Garrott, J. R. (1991). *Chinese Students' Cultural Values and Their Attitudes toward English-language Learning and Teaching* (Unpublished doctoral dissertation), The University of Texas, Austin.

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

Green, A. (2007). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge: Cambridge University Press.

Green, A. (2013). Washback in Language Assessment. *International Journal of English Studies, 13*(2), 39-51.

Hamp-Lyons, L. (1997). Washback, Impact and Validity: Ethical Concerns. *Language Testing, 14*(3), 295-303.

Hamp-Lyons, L. (2007). The Impact of Testing Practices on Teaching: Ideologies and Alternatives. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching Part 1* (pp. 487-504). New York: Springer Science+Business Media, LLC..

Hannum, E., An, X., & Cherng, H. S. (2011). Examinations and Educational Opportunity in China: Mobility and Bottlenecks for the Rural Poor. *Oxford Review of Education, 37*(2), 267-305.

Harlen, W., & Crick, R. (2003). Testing and Motivation for Learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207.

Hayes, B., & Read, J. (2004). IELTS Test Preparation in New Zealand: Preparing Students for the IETLS Academic Module. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 97-112). Mahwah: Lawrence Erlbaum Associates, Inc.

Hu, G. (2002). Potential Cultural Resistance to Pedagogical Imports: The Case of Communicative Language Teaching in China. *Language, Culture and Curriculum, 15*(2), 93-105.

Hu, G. (2003). English Language Teaching in China: Regional Differences and Contributing Factors. *Journal of Multilingual and Multicultural Development, 24*(4), 290-318.

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

Hughes, A. (2013). *Testing for Language Teachers* (13th ed.). Cambridge: Cambridge University Press.

Jin, L., & Cortazzi, M. (2006). Changing Practices in Chinese Cultures of Learning. *Language, Culture and Curriculum*, *19*(1), 5-20.

Jin, Y. (2014). The Accidental Language Tester: An Interview with Liz Hamp-Lyons. *Language Assessment Quarterly*, *11*(3), 338-351.

Jones, B. D., & Egley, R. J. (2004). Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing. *Education Policy Analysis Archives, 12*(39), 1-34.

Lange, B. (1981). ERIC/RCS: Promoting Test-wiseness. *Journal of Reading*, 740-743.

Li, H. (2009). Are Teachers Teaching to the Test: A Case Study of the College English Test (CET) in China. *International Journal of Pedagogies and Learning, 5*, 25-36.

Li, X. (1990). How Powerful can a Language Test be? The MET in China. *Journal of Multilingual and Multicultural Development, 11*, 393–404.

Madaus, G. E. (1988). The Influence of Testing on the Curriculum. In L.N. Tanner (Eds.), *Critical Issues in Curriculum: Eight-seven Yearbook of the National Society for the Study of Education, Part 1*(pp.83-121).Chicago, IL: University of Chicago Press.

Mahdavi, A., & Ghabelju, M. (2013). Test Taking Strategies Instruction and Iranian EFL Learners' Performance on International Language Tests. *International Journal of Applied Linguistics & English Literature*, *2*(1).

Mao, L., & Min, Y. (2004). Foreign Language Education in the PRC. In M. Zhou & H. Sun (Eds.), *Language Policy in the People's Republic of China: Theory and Practice since 1949* (pp. 319-329). Dordrecht, NLD: Kluwer Academic Publisher.

Matoush, M. M., & Fu, D. (2012). Tests of English Language as Significant Thresholds for College-bound Chinese and the Washback of Test-preparation. *Changing English, 19*(1), 111-121.

McNamara, T. (1996). *Measuring Second Language Performance.* London: Longman.

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

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational Measurement* (3 ed., pp. 13-103). New York: Macmillan.

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

Millman, J., Bishop, H., & Ebel, R. (1965). An Analysis of Test-wiseness. *Educational and Psychological Measurement*, *25*, 707-726.

Moore, S., Stroupe, R., & Mahony, P. (2009). Perceptions of IELTS in Cambodia: A Case Study of Test Impact in a Small Developing Country. *IELTS Research Reports, 13,* 1-109. Retrieved from http://www.ielts.org/PDF/vol13_Report6.pdf.

Moore, T., & Morton, J. (2005). Dimensions of Difference: A Comparison of University Writing and IELTS Writing. *Journal of English for Academic Purposes, 4*(1), 43-66.

Morrows, K. (1986). The Evaluation of Tests of Communicative Performance. In M. Portal (Ed.), *Innovations in Language Testing: Proceedings of the IUS/NFER conference* (pp. 1-13). London: NFER/Nelson.

Murray, J. C., Cross, J. L., & Cruickshank, K. (2014). Stakeholder Perceptions of IELTS as a Gateway to the Professional Workplace: The Case of Employers of Overseas Trained Teachers. *IELTS Research Report Online Series 2014/1*.

Murray, J. C., Riazi, A. M., & Cross, J. L. (2012). Test Candidates' Attitudes and Their Relationship to Demographic and Experiential Variables: The Case of Overseas Trained Teachers in NSW, Australia. *Language Testing*, *29*(4), 577-595.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Pan, Y. C. (2009). A Review of Washback and Its Pedagogical Implications. *VNU Journal of Science Foreign Languages, 25*, 257-263.

Pearson, I., (1988). Tests as Levers for Change. In Chamberlain, D. & Baumgardner, R. J., *ESP in the Classroom: Practice and Evaluation* (pp.98-107). London: Modern English.

Popham, W. J. (1987). The Merits of Measurement-driven Instruction. *Phi Delta Kappan*, 679-682.

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.

Rea-Dickins, P. (1997). So, Why do We Need Relationships with Stakeholders in Language Testing? A View from the UK. *Language Testing, 14*(3), 304-314.

Robson, C. (1993). *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*. Oxford: Blackwell.

Rosenberg, M. J., & Hovland, C. I. (1960). Attitude Organization and Change. In M. J. Rosenberg & C. I. Hovland (Eds.), *Cognitive, Affective, and Behavioral Components of Attitudes* (pp. 1-14). New Haven: Yale University Press.

Rowland, B. (2011). *The Influence of High Stakes Testing and Test Preparation on High School Students' Perspectives on Education and Lifelong Learning* (Unpublished doctoral dissertation), Walden University, US.

Shepard, L. A. (1990). Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test?. *Educational Measurement: Issues and Practice, 9*(3), 15-22.

Shepard, L. A. (1992). What Policy Makers Who Mandate Tests should Know about the New Psychology of Intellectual Ability and Learning. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing Assessment: Alternative Views of Aptitude, Achievement and Instruction* (pp. 301-327). London: Kluwer Academic.

Shi, L. (2006). The Successors to Confucianism or a New Generation? A Questionnaire Study on Chinese Students' Culture of Learning English. *Language, Culture and Curriculum, 19*(1), 122-147.

Shih, C. M. (2007). A New Washback Model of Students' Learning. *The Canadian Modern Language Review, 64*(1), 135-162.

Shih, C. M. (2010). The Washback of the General English Proficiency Test on University Policies: A Taiwan Case Study. *Language Assessment Quarterly, 7*(3), 234-254.

Shohamy, E. (1997). Testing Methods, Testing Consequences: Are They Ethical? Are They Fair?. *Language Testing, 14*(3), 340-349.

Shohamy, E. (1999). Language Testing: Impact. In B. Spolsky (Ed.), *Concise Encyclopaedia of Educational Linguistics* (pp. 711-714). Oxford Pergamon.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. Harlow: Longman.

Shohamy, E. (2007). The Power of Language Tests, the Power of the English Language and the Role of ELT. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching Part 1* (pp. 521-532). New York: Springer Science+Business Media, LLC. .

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test Impact Revisited: Washback Effect over Time. *Language Testing, 13*(3), 298-317.

Smith, H., & Haslett, S. (2007). Attitudes of Tertiary Key Decision-makers towards English Language Tests in Aotearoa New Zealand: Report on the Results of a National Provider Survey. *IELTS Research Report, 7*.

Smith, M. L. (1991). Meanings of Test Preparation. *American Educational Research Journal, 28*(3), 521-542.

Song, X., & Cheng, L. (2006). Language Learner Strategy Use and Test Performance of Chinese Learners of English. *Language Assessment Quarterly, 3*(3), 243-266.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.

Spratt, M. (2005). Washback and the Classroom: The Implications for Teaching and Learning of Studies of Washback from Exams. *Language Teaching Research, 9*(1), 5-29.

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Stoneman, B. W. H. (2006). *The Impact of an Exit English Test on Hong Kong Undergraduates: A Study Investigating the Effects of Test Status on Students' Test Preparation Behaviors* (Unpublished doctoral dissertation). The Hong Kong Polytechnic University, Hong Kong.

Stricker, L. J., Wilder, G. Z., & Bridgeman, B. (2006). Test Takers' Attitudes and Beliefs about the Graduate Management Admission Test. *International Journal of Testing, 6*, 255–268.

Sukyadi, D., & Mardiani, R. (2011). The Washback Effect of the English National Examination on English Teachers' Classroom Teaching and Students' Learning. *K@ta: a Biannual Publication on the Study of Language and Literature, 13*(1), 96-111.

Sundre, D. L., & Kitsantas, A. (2004). An Exploration of the Psychology of the Examinee: Can Examinee Self-regulation and Test-taking Motivation Predict Consequential and Non-consequential Test Performance?. *Contemporary Educational Psychology, 29*(1), 6-26.

Tavakoli, E., & Samian, S. H. (2014). Test-wiseness Strategies in PBTs and IBTs: The Case of EFL Test Takers, Who Benefits More?. *Procedia-Social and Behavioral Sciences, 98*, 1876-1884.

Tsagari, K. (2006). *Investigating the Washback Effect of a High-stakes EFL Exam in the Greek Context: Participants' Perceptions, Material Design and Classroom Applications* (Unpublished doctoral dissertation), Lancaster University, UK.

Wall, D. (1997). Impact and Washback in Language Testing. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of Language and Education: Vol. 7 Language Testing and Assessment* (pp. 291-302). Dordrecht: Kluwer Academic.

Wall, D. (1999). *The Impact of High-stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory* (Unpublished doctoral dissertation), Lancaster University, UK.

Wall, D. (2000). The Impact of High-stakes Testing on Teaching and Learning: Can this be Predicted or Controlled?. *System, 28*, 499-509.

Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 79-92). USA, Canada: Routledge.

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

Wall, D., & Horák, T. (2007). Using Baseline Studies in the Investigation of Test Impact. *Assessment in Education: Principles, Policy & Practice, 14*(1), 99-116.

Wall, D., & Horák, T. (2008). The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe: Phase 2, Coping with Change. *Princeton, NJ: Educational Testing Service*.

Watanabe, Y. (1996). Investigating Washback in Japanese EFL Classrooms: Problems of Methodology. In G. Wigglesworth & C. Elder (Eds.), *The Language Testing Cycle: From Inception to Washback* (pp. 208-239), Canberra: ALAA.

Watanabe, Y. (1997). *The Washback Effects of the Japanese University Entrance Examinations of English: Classroom-based Research* (Unpublished doctoral dissertation), University of Lancaster, UK.

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: Lawrence Erlbaum Associates, Inc.

Xiao, Y., & Carless, D. R. (2013). Illustrating Students' Perceptions of English Language Assessment: Voices from China. *RECL Journal*, *44*(3), 319-340.

Xiao, Y., Sharpling, G., & Liu, H. (2011). Washback of National Matriculation English Test on Students' Learning in the Chinese Secondary School Context. *Asian EFL Journal*, *13*(3), 103-29.

Xie, Q., & Andrews, S. (2012). Do Test Design and Uses Influence Test Preparation? Testing a Model of Washback with Structural Equation Modelling. *Language Testing*, *30*(1), 49-70.

Yang, P. (2000). *Effects of Test-wiseness upon Performance of the Test of English as a Foreign Language* (Unpublished doctoral dissertation), University of Alberta, Edmonton.

Yildirim, O. (2011). Washback Effects of a High-stakes University Entrance Exam: Effects of the English Section of the University Entrance Exam on Future English Language Teachers in Turkey. *The Asian EFL Journal Quarterly*, *12*(2), 92-116.

Yu, Y. (2010). *The Washback Effects of School-based Assessment on Teaching and Learning: A Case Study* (Unpublished doctoral dissertation), The University of Hong Kong, Hong Kong.

Zhan, Y., & Wan, Z. H. (2013). Dynamic Nature of Washback on Individual Learners: The Role of Possible Selves. *Assessment & Evaluation in Higher Education*, 1-19.

Zhang, Y. (2013). Does Private Tutoring Improve Students' National College Entrance Exam Performance?—A Case Study from Jinan, China. *Economics of Education Review*, 23, 1-28.

Zhao, J. (2006). *Exploring the Relationship between Chinese University Students' Attitudes toward College English Test and Their Test Performance* (Unpublished doctoral dissertation), Queen's University, Kingston, Canada.

Zuo, H. (1995). *Analysis of the Experimental College Entrance English Examination in the People's Republic of China with a Proposal for Revision* (Unpublished doctoral dissertation), Georgetown University, Washington, D.C.

Appendix One:

The 2014 *Jiangxi* National Matriculation English Test

Volume One (Multiple-choice questions: 115 Marks)

Part One: Please write down your answers on the test paper. At the end of the recording, you will be given 2 minutes to transfer your answers to the answer sheet.

Section One (Question 1-5; 1.5 marks for each question; total 7.5 marks):

Listen to the 5 conversations and answer the following questions. Select the correct item from A, B and C and write it down on the corresponding place. After each conversation, you will be given 10 seconds to answer the question and read the next one. Each conversation will be played once only.

Example: How much is the shirt?

A. £19.15

B. £9.18£

C. £9.15

The Correct Answer is C. Now listen carefully.

1. What does the woman want to do?

A. Find a place. B. Buy a map. C. Get an address.

2. What will the man do for the woman?

A. Repair her car. B. Give her a ride. C. Pick up her aunt.

3. Who might Mr. Peterson be?

A. A new professor. B. A department heard. C. A company director.

4. What does the man think of the book?

A. Quite difficult. B. Very interesting. C. Too simple.

5. What are the speakers talking about?

A. Weather. B. Clothes. C. News.

Section Two: (Question 6-20; 1.5 marks for each question; total 22.5 marks)

Listen to the next five conversations or monologues and select the correct answer from A, B and C. Before the recording, you will be given 5 seconds to read each question, and after the recoding, you will be given 5 seconds to write down your answers. Each conversation or monologue will be played twice.

Listen to the sixth recording and answer questions 6 and 7.

6. Why is Harry unwilling to join the woman?

A. He has a pain in his knee. B. He wants to watch TV. C. He is too lazy.

7. What will the woman probably do next?

A. Stay at home. B. Take Harry to hospital. C. Do some exercises.

Listen to the seventh recording and answer questions 8 and 9.

8. When will the man be home from work?

A. At 5:45. B. At 6:15. C. At 6:50.

9. Where will the man be home from work?

A. The Green House Cinema. B. The New State Cinema. C. The UME Cinema.

Listen to the eighth recording and answer questions 10 to 12.

10. How will the speakers go to New York?

A. By air. B. By taxi. C. By bus.

11. Why are the speakers making the trip?

A. For business. B. For shopping. C. For holiday.

12. What is the probably relationship between the speakers?

A. Driver and passenger. B. Husband and wife. C. Fellow workers.

Listen to the ninth recording and answer questions 13 to 16.

13. Where does this conversation probably take place?

A. In a restaurant. B. In an office. C. In a classroom.

14. What does John do now?

A. He's a trainer. B. He's a tour guide. C. He's a college student.

15. How much can a new person earn for the first year?

A. $10,500. B. $12,000. C. $15,000.

16. What kind of life does the speaker seem to like most?

A. Four. B. Three. C. Two.

Listen to the tenth recording and answer questions 17 to 20.

17. How long has the speaker lived in a big city?

A. One year. B. Ten years. C. Eighteen years.

18. What is the speaker's opinion on public transport?

A. It's comfortable. B. It's time-saving. C. It's cheap.

19. What is good about living in a small town?

A. It's safer. B. It's healthier. C. It's more convenient.

20. What kind of life does the speaker seem to like most?

A. Busy. B. Colorful. C. Quiet.

Part Two: Knowledge Application (Two sections; total 45 marks)

Section One: Multiple-Choice Questions (Questions 21-35; 1 mark for each question; total 15 marks)

21.----Could I use this dictionary ?

----_____.It's a spare one.

A. Good idea B. Just go ahead C. You're welcome D. You'd better not

22. They chose Tom to be ___captain of the team because they knew he was __smart leader.

A. a; the B. the; the C. the; a D. a; a

23 Thanks for your directions to the house; we wouldn't have found it ___.

A. nowhere B. however C. otherwise D. instead

24.----Tony , why are your eyes red ?

---I __ up peppers for the last five minutes.

A. cut B. was cutting C. had cut D. have been cutting

25. Starting your own business could be a way to achieving financial independence. ____, it could just put you in debt.

A. In other words B. All in all C. As a result D. On the other hand

26. When it comes to __ in public, no one can match him.

A. speak B. speaking C. being spoken D. be spoken

27. Anyway, we're here now, so let's ___some serious work.

A. come up with B. get down to C. do away with D. live up to

28. Among the many dangers_-- sailors have to face, probably the greatest of all is fog.

A. which B. what C. where D. when

29. I don't believe what you said, but if you can prove it, you may be able to __-me .

A. convince B. inform C. guarantee D. refuse

30. Life is unpredictable; even the poorest __become the richest.

A. shall B. must C. need D. might

31. ___nearly all our money, we couldn't afford to stay at a hotel.

A. Having spent B. To spent C. Spent D. To have spent

32. ---When shall I call, in the morning or afternoon?

----___. I'll be in all day.

A. Any B. None C. Neither D. Either

33. It is unbelievable that Mr. Lucas Leads a simple life __his great wealth.

A. without B. despite C. in D. to

34. He is thought ___foolishly .Now he has no one but himself to blame for losing the job .

A. to act B. to have acted C. acting D. having acted

35. It was the middle of the night __ my father woke me up and told me to watch the football game.

A. that B. as C. which D. when

Section Two: Cloze (Question 36-55; 1 mark for each question; total 30 marks)

Read the following passage and select the correct answers from A, B, C and D.

"Mum, can I invite my classmate Brett over to stay tomorrow night, please? It's Friday, and we don't have any _36_. Can I, please?" Mum was sitting at the kitchen table. Dad was_37_next to her, resting his head on his arms. Mum could _38_that James wanted so badly to have his friend over.

“I’m so sorry, James,” she said.

“I’m never allowed to have friends come to the house? Why, Mum?” James asked sadly, almost in _39_.

“I know it’s _40_ for you,” Mum said softly. “But I’m just worried other people might think we’re a little… strange. And then they would make fun of you. ”

“No, they wouldn’t, Mum,” James protested. “We’re not _41_ at all. We’re just ordinary people. ”

Mum sighed heavily. “To tell you the truth, James, my neck has been so painful that it’s given me a heavy _42_. And your poor father –he doesn’t feel _43_. He really needs a rest. ”

“I can help, Mum!” James said. “ _44_ I can make you and Dad feel better, can Brett come over? Please?”

“Well…”Mum began.

“Great! Thanks, Mum!” James almost shouted. “Just sit there, don’t move. ” He rushed over to the kitchen drawer and _45_ what he needed –two spanners. “Hang on, Mum,” he said. “This won’t take a second. ” After some _46_, James was finished. With a smile of _47_ on his face he said, “There! How does that feel?” “Oh, James,” Mum said. “That’s a much better! How did you do it?”

“Easy,” James said _48_. ” Dad had tightened your neck bolts too much! I just _49_ them slightly! I learned that in robotic science at school. ”

"What about you father? Can you _50_ him?" asked Mum.

"I'll try," James replied. He _51_ up Dad's hair at the back of his neck. and plugged the electric wire into Dad's head. Then he turned the _52_ on. Dad opened his eyes and _53_ immediately. "He just let his _54_ run too low, that's all," James said, "Shall I tell Brett to come over straight after school tomorrow?"

"I guess so," replied Mum. "Your friends will just have to _55_ that we are a very unusual family. Thanks, son!"

36. A. chance B. message C. homework D. difficulty

37. A. a sleep B. reading C. alone D. standing

38. A. explain B. see C. agree D. doubt

39. A. terror B. tears C. surprise D. silence

40. A. fair B. easy C. good D. hard

41. A. strange B. normal C. popular D. anxious

42. A. headache B. loss C. task D. day

43. A. ill B. funny C. sorry D. well

44. A. As B. If C. Since D. Before

45. A. kept B. controlled C. found D. returned

46. A. requests B. thoughts C. repairs D. instructions

47. A. sympathy B. satisfaction C. bitterness D. politeness

48. A. embarrassedly B. gratefully C. impatiently D. proudly

49. A. adjusted B. collected C. produced D. covered

50. A. greet B. accompany C. help D. ask

51. A. lifted B. caught C. gave D. filled

52. A. television B. power C. light D. gas

53. A. grew up B. lay down C. broke down D. sat up

54. A. food B. temperature C. battery D. blood

55. A. prove B. expect C. suspect D. accept

Part Three: Reading Comprehension (Question 56-75; 2 marks for each question; total 40 marks)

A

Larry was on another of his underwater expeditions but this time, it was different. He decided to take his daughter along with him. She was only ten years old. This would be her first trip with her father on what he had always been famous for.

Larry first began diving when he was his daughter's age. Similarly, his father had taken him along on one of his expeditions. Since then, he had never looked back. Larry started out by renting diving suits from the small diving shop just along the shore. He had hated them. They were either too big or too small. Then, there was the instructor. He gave him a short lesson before allowing him into the water with his father. He had made an exception. Larry would never have been able to go down without at least five hours of theory and another similar number of hours on practical lessons with a guide. Children his age were not even allowed to dive.

After the first expedition, Larry's later diving adventures only got better and better. There was never a dull moment. In his black and blue suit and with an oxygen tank fastened on his back, Larry dived from boats into the middle of the ocean. Dangerous areas did not prevent him from continuing his search. Sometimes, his was limited zxxk.com to a cage underwater but that did not bother him. At least, he was still able to take photographs of the underwater creatures.

Larry's first expedition without his father was in the Cayman Islands. There were numerous diving spots in the area and Larry was determined to visit all of them. Fortunately for him, a man offered to take him around the different spots for free. Larry didn't even know what the time was how many spots he dived into or how many photographs he had taken. The diving spots afforded such a wide array of fish and sea creatures that Larry saw more than thirty varieties of creatures.

Larry looked at his daughter. She looked as excited as he had been when he was her age. He hoped she would be able to continue the family tradition. Already, she

looked like she was much braver than had been then. This was the key to a successful underwater expedition.

56. In what way was this expedition different for Larry?

A. His daughter had grown up.

B. He had become a famous diver.

C. His father would dive with him.

D. His daughter would dive with him.

57. What can be inferred from Paragraph 2?

A. Larry had some privileges.

B. Larry liked the rented diving suits.

C. Divers had to buy diving equipment.

D. Ten-year-old children were permitted to dive.

58. Why did Larry have to stay in a cage underwater sometimes?

A. To protect himself from danger.

B. To dive into the deep water.

C. To admire the underwater view.

D. To take photo more conveniently.

59. What can be learned from the underlined sentence?

A. Larry didn’t wear a watch.

B. Larry was not good at math.

C. Larry had a poor memory.

D. Larry enjoyed the adventure.

60. What did Larry expect his daughter to do?

A. Become a successful diver.

B. Make a good diving guide.

C. Take a lot of photo underwater.

D. Have longer hours of training.

B

HOLIDAY FUN AT THE POWERHOUSE

500 HARRIS STREET ULTIMO ·TELEPHONE (02)9270111

Join in the holiday fun at the powerhouse this month linked to our new exhibition, Evolution & Revolution: Chinese dress 1700s to now. DON'T FORGET our other special event, the Club Med Circus School which is part of the Circus ! 150 years of circus in Australia exhibition experience!

◆ **Chinese Folk Dancing:** Colorful Chinese dance and musical performances by The Chinese Folk Dancing School of Sydney. Dances include: the Golden and the Chinese drum dance. A feature will be the Qin dynasty Emperor's count dance. Also included is a show of face painting for Beijing opera performances.
Sunday 29 June and Wednesday 2 July in the Turbine Hall, at 11:30 am & 1:30 pm.

◆ **Australian Chinese Children's Arts Theatre:** Well-known children's play experts from Shanghai leas this dynamic youth group. Performances include Chinese fairy tales and plays.
Thursday 3 to Sunday 6 July in the Turbine Hall, at 11:30 am & 1:30 pm.

◆ **Chinese Youth League:** A traditional performing arts group featuring performance highlights such as Red scarf and Spring flower dances, and a musician playing Er Hu.
Sunday 6 to Tuesday 8 July in the Turbine Hall, 11:30 am to 1:30 pm.

◆ **Kids Activity: Make a Paper Horse:** Young children make a paper horse cut-out. (The horse is a frequent theme in Chinese painting, including a kind of advancement.) Suitable for ages 8-12 years.
Saturday 28 June to Tuesday 8 July in the Turbine Hall, 12:30 pm to 1:00 pm.

◆ **Club Med Circus School:** Learn circus skills, including the trapeze, trampolining and magic. Note only for children over 5. There are 40 places available in each 1 hour session and these must be booked at the front desk, level 4, on the day.
Tuesday 1 to Saturday 5 July at 11:30 am & 1:00 pm

Enjoy unlimited free visits and many other benefits by becoming a Family member of the Powerhouse. Our family memberships cover two adults and all children under the age of 16 years at the one address.

Members receive *Powerline*, our monthly magazine, discounts in the shops and restaurants, as well as free admission to the Museum. All this for as little as $50.00 a year! Call (02)9217 0600 for more details.

61. When can you watch the Chinese drum dance?
A. On July 2. B. On July 3.
C. On July 6. D. On July 8.

62. To learn the magic tricks, you can go to ___ .
A. Kids Activity. B. Chinese Youth League.
C. Club Med Circus School. D. Children's Arts Theatre.

63.What is required if you want to enjoy free visits to the Museum?

A. Calling (02)92170600. B. Gaining family membership.

C. Coming for the holiday fun. D. Paying *Powerline* $50.00 a year.

64. What is the main purpose of the text?

A. To attract visitors. B. To present schedules.

C. To report the performances. D. To teach kids Chinese arts.

C

Close your eyes for a minutes and imagine what life would be like if you had a hundred dollars less. Also imagine what it would be like spending the rest of your life with your eyes closed. Imagine having to read this page, not with your eyes but with your finger-tips.

With existing medical knowledge and skills, two-thirds of the world's 42 million blind should not have to suffer. Unfortunately, rich countries possess most of this knowledge, while developing countries do not.

ORBIS is an international non-profit organization which operates the world's only flying teaching eye hospital. ORBIS intends to help fight blindness worldwide. Inside a DC-8 aircraft, there is a fully-equipped teaching hospital with television studio and classroom. Doctors are taught the latest techniques of bringing sight back to people there. Project ORBIS also aims at promoting peaceful cooperation among countries.

ORBIS tries to help developing countries by providing training during three-week medical programs. ORBIS has taught sight-saving techniques to over 35,000 doctors and nurses, who continue to cure tens of thousands of blind people every year. ORBIS has conducted 17 plane programs is China so far. For the seven to ten million blind in China ORBIS is planning to do more for them. At the moment an ORBIS is working on a long-term plan to develop a training centre and to provide eye care service to Shanxi Province. ORBIS needs your help to continue their work

and free people from blindness.

For just US$38, you can help one person see; for $380 you can bring sight to 10 people; $1,300 helps teach a doctor new skills; and for $13,000 you can provide a training programme for a group of doctors who can make thousands of blind people see again. Your money can open their eyes to the world. Please help ORBIS improve the quality of life for so many people less fortunate than ourselves.

65.The first paragraph is intended to _____.
A.introduce a new way of reading
B.advise the public to lead a simple life
C.direct the public's attention to the blind
D.Encourage the public to use imagination

66.What do we learn about existing medical knowledge and skills in the world?
A.They are adequate B. They have not been updated.
C. They are not equally distributed D. They have benefited most of the blind

67.ORRIS aims to help the blind by _____.
A. teaching medical students B. training doctors and nurses
C. running flying hospitals globally D. setting up non-profit organization

68.What does the author try to do in the last paragraph?
A. Appeal for donations B. Make an advertisement
C. Promote training programs D. Show sympathy for the blind

69.What can be the best title for the passage?
A.ORRIS in China B. Fighting Blindness
C.ORRIS Flying Hospital D. Sight-seeing Techniques

D

Everyone looks forward to progress, whether in one's personal life or in the general society. Progress indicates a person's ability to change the way he is living at the moment. Progress must lead a better way of doing things. All these, however,

remains true only in so far as people want to accept technology and move forward by finding new and more efficient ways of doing things.

However, at the back of the minds of many people, especially those who miss the "good old days", efficiency comes with a price. When communication becomes efficient, people are able to conduct one another no matter where they are and at whatever time they wish to. The click of a button allows people miles apart to talk or to see each other without even leaving their homes. With the communication gadgets, such as mobile phones and IPads, people often do not take the effect to visit one another personally. A personal visit carries with the additional feature of having to be in the person's presence for as long as the visit lasts. We cannot unnecessarily excuse ourselves or turn the other person off.

With efficiency also comes mass production. Such is the nature of factories and the success of industrialization today. Factories have improved efficiency. Unskilful tasks are left to machines and products are better made and produced with greater accuracy than any human hand could ever have done. However, with the improvements in efficiency also comes the loss of the personal touch when making these products. For example, many handcrafts are now produced in a factory. Although this means that supply is better able to increase demand, now that the supply is quick and efficient, the demand might fall because mass production lowers the quality of the handicraft and it is difficult to find unique designs on each item.

Nevertheless, we must not commit the mistake of analyzing progress only from one point of view. In fact, progress has allowed tradition to keep up. It is only with progress and the invention of new technology that many old products can be brought back to their old state. New technology is required for old products to stay old.

It is people's attitude towards progress that causes the type of influence that technology has on society. Technology is flexible. There is no fixed way of making use of it. Everything depends on people's attitude. The worst effects of progress will fall on those who are unable to rethink their attitudes and views of society. When we accept progress and adapt it to suit our needs, a new "past" is created.

70. According to Paragraph 1, progress can benefit people when they are willing to _____.

A. live a better life B. look for better methods
C. change ways of living D. accept technology and advance steadily

71. The underlined word "gadgets" is closest in meaning to _____.

A. tools B. messages C. barriers D. skills

72. The author explains "efficiency comes with a price" by _____.

A. describing a process B. using examples
C. following time order D. making classification

73. Compared with home-made handicrafts, machine made products _____.

A. lack great accuracy B. lack the personal touch
C. are of high value D. are quite welcome

74. What can be learned about technology from Paragraph 4?

A. It can destroy old traditions. B. It can lead to social progress.
C. It can be used to correct mistakes.
D. It can be used to preserve old products.

75. What can be concluded from the last paragraph?

A. Progress can suit the needs of daily life.
B. People review the past with great regret.
C. Technology should be introduced in a fixed way.
D. People's attitude decides the use of technology.

Volume Two (Non-multiple-choice Questions; total 35 marks)

Please write down your answers on the answer sheet with black gel pens. The answers written in the test paper are invalid.

Part Four: Writing (Two Sections; total 35 marks)

Section One: Reading and Writing (Question 76-80; 2 marks for each question; total 10 marks) Attention on the WORDS requirement for the answers.

【1】A safari park is a park in which wild animals are kept. They are mainly located in east or central Africa. They often occupy a very wild area, with mountains and rivers. To visit the park and look at the animals, people have to drive around in a car for a few of hours because the park is huge. 【2】In south Africa there is a safari park, which contains all sorts of wild animals like lions, elephants, rhinoceroses, zebras, wild pigs, deer and giraffes. 【3】There is a wild road leading through the park, but nobody is permitted to walk on the road. Anyone traveling in the park has to go in a car because wild animals may fiercely attack people. From the car he may see almost every types of African wildlife. Some of these are getting scarce because people kill them for various reasons. For example, rhinoceroses are killed for their horns, which are used in traditional Chinese medicines for colds and headaches. Perhaps they will be seen only in museums and books one day. 【4】 Travels may purchase food for the animals. They can feed them when they tour the park. Of course, they should not feed them in a close distance because the wild animals may attack people. In addition, they should only give proper food to the animals. 【5】A traveller may carry a gun with him in his journey. The gun is given to him by the government. However, it is not used for hunting. In fact, a seal is fixed to it. The traveller may fire at a wild beast to defend himself in case he is attacked. However, he has to prove to the government that he has been attacked and that he has not fired at a harmless animal.

76. List one of the reasons why travellers have to stay in a car during their visit. （no more than 6 words）

_____

_

77. Why are rhinoceroses getting scarce? （no more than 7words）

_____

_

78．What warnings are given to travellers when they feed the animals? （no more than 15words）

_____
_

79．How does the government know the gun is fired? （no more than 10 words）

_____
_

80．What is the passage mainly about? （no more than 6 words）

_____
_

Section Two: Writing.

Assume yourself as a graduate student of Xingguang High School, Li Hua. You are invited to talk about your experiences in a freshman welcoming events called '*What to learn in senior high school*?' Please write a speech according to the following instructions.

1, learn how to learn: methods, habits and so forth;

2, learn how to be behave：honestly, friendly;

3, learn… (Please add extra information)

Attention: 1, write around 120 words

2, both the beginning and the ending have been provided to you. There is no need to copy them on the answers sheet.

Good morning, everyone! It is my honor to be here to share with you my opinions on learn in senior high school.

_____
_____

_____

_____

Thank you!

Reference:

Jiangxi Provincial Education Examination Authority (2014). *The 2014 Jiangxi National Matriculation English Test*. Retrieved June 9, 2014, from http://www.jxeea.cn/ksydgsj/2014/06/2014060908530634.html

Appendix Two:

Questionnaire Used for Pilot Study (*English Version)*

Part One:

Please put an "**x**" in the box ☐ that is in front of the information that suits you. .

No.	Questions	Choice Items	
1	What is your gender?	☐ Male	☐ Female
2	What class types are you in?	☐ Class of Liberal Arts	☐ Class of Sciences
3	Have you taken the NMET before?	☐ Yes	☐ No

No.	Questions	Choice Items			
3	How long have you been taught English?	☐ 9 years or more	☐ 7 to 8 years	☐ 5 to 6 years	☐ 4 years or less
4	What is your average score of monthly NMET mock exams?	☐ 150-120	☐119-90	☐ 89-60	☐ Below 60
5	Since freshmen year of high school, your achieved score in English tests are _____:	☐ Generally increased	☐ Averagely the same with dramatically fluctuation	☐ Averagely the same with slightly fluctuation	☐ Generally decreased
6	How many hours have your spent normally in English learning each week (classroom hours excluded)?	☐ 20 or more	☐ 15 to 20	☐ 10 to 15	☐ 10 or less
7	What is the primary reason you study English?	☐ To communicate with English-speaking people	☐ To fulfill parents' expectations	☐ To achieve higher score in NMET	☐ English is an easy subject to study

Part Two:

Following are a number of statements with which some people agree and others are not. I would like you to present your opinions of each statement by putting an "**X**" in the box where the column links with the extent of your agreement or disagreement.

For example:

Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
English is the easiest subject in high school.	**X**			

Please avoid answering like this:

Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
English is the easiest subject in high school.		**X**		

Please write '**X**' down, as the example shows, to indicate your degree of disagreement or agreement. Thank you.

No.	Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
1	NMET is the dominant topic in English classes.				
2	Almost all the assigned exercises are NMET-related.				
3	English teachers mention NMET frequently during classes.				
4	Textbooks for English subject are essentially the NMET preparation books.				
5	Contents that irrelevant to NMET are seldom mentioned in the English classes.				
6	NMET scores take too much percentage in university admission decisions.				

7	I practice multiple-choice tasks all the time.				
8	Memorizing the high-frequency words is enough for passing the NMET.				
9	Focusing on practicing Reading exercises is an effective way to improve the overall score.				
10	Sometimes, test-taking strategies would be more important than actual language abilities.				
11	Both my English learning methods and study schedules will be modified constantly according the countdown of NMET.				
12	Comparing reading and listening, hours spent in doing writing exercises are less.				
13	The main English study activity I usually do is practicing NMET stimulated exercises.				
14	The loads of NMET stimulated exercises you have done determine your NMET score.				
15	In the listening section of the NMET, some of the task requires no comprehension. I could select the correct answers by test-taking strategies.				
16	In the Reading section of the NMET, some of the task requires no comprehension. I could select the correct answers by test-taking strategies.				
17	Memorizing the written models is an effective way to achieve high score in writing section.				
18	Teachers have explained the NMET rationale and requirement to use in class.				
19	I am fully aware of the NMET rationale and requirements.				
20	‘Good English learners’ does not equal ‘high score in the NMET’.				
21	‘High score in the NNET’ does not equal ‘Good English learners’.				
22	NMET is a valid test inferring my language ability.				
23	Studying for NMET is directly related to my future needs in university.				
24	NMET motivates me to work harder.				
25	Preparing NMET helps me develop confidence.				
26	I felt under stressed preparing for NMET.				

27	Sometimes, I felt I want to give it up.				
28	I prefer that the NMET could be taken more than once a year.				
29	Generally speaking, NMET positively influences my study and me.				
30	Generally speaking, NMET negatively influences my study and me.				

Part Three:

1. The statements in the Part two, are they clear and understandable?
   If not, please write down your reasons.

   Yes No_____

2. The statements in the Part two, are they easy to interpret?
   If yes, please write down the number of the statement.

   Yes_____ No

3. The statements in the Part two, are there any sensitive topics?
   If yes, please write down the number of the statement.

   Yes _____ No

4. If you have any suggestions for this survey, please feel free to write them down on the following lines: _____

_____

_____

_____

_____

Thanks for your kind participation!

Questionnaire Used for Pilot Study *(Chinese Version)*

第一部分:

请在符合你情况的格子 □中画 "X":

No.	**问题**	**选项**	
1	你的性别?	□男	□女
2	你在文科班还是理科班?	□ 文科班	□ 理科班
3	请问你以前参加过高考么?	□ 参加过	□ 没有参加过

No.	**问题**	**选项**			
1	你学习英语多少年了?	□9 年或者更多	□7 到 8 年	□5 到 6 年	□4 年或更少
2	你在英语模考中的平均成绩属于下面那一个范围?	□ 150-120	□ 120-90	□ 90-60	□ 低于 60
3	你平常每周有多少小时花在英语学习上(课堂时间除外)?	□ 20 或更多	□ 15 to 20	□ 10 to 15	□ 10 或更少
4	从高一开始,我的英语月考成绩_____:	□总体为上升趋势	□总体保持平均,中间有较大的浮动	□总体保持平均,中间有较小的浮动	□总体为下降趋势
5	你为什么学习英语?(请选择你认为最重要的一项)	□为了同外国人交流	□为了满足父母亲的期望	□为了在高考中取得高分	□因为英语相对而言比较简单,容易学

第二部分:

以下是一系列意见性的陈述,请通过在相应格子中画"X" 的方式来表达你同意或不同意的程度.

例如:

意见 观点	非常不赞同	不赞同	赞同	非常赞同
英语是最简单的学科。	X			

请避免如此填写:

意见 观点	非常不赞同	不赞同	赞同	非常赞同
英语是最简单的学科。		X		

以下为正式问卷内容,请认真填写:

题号	意见 观点	非常不赞同	不赞同	赞同	非常赞同
1	高考相关的英语语言知识是英语课上的主导课题。				
2	几乎所有的英语家庭练习题是高考相关的。				
3	英语课上，“高考”是频繁提起的词。				
4	英语教科书本质上是高考备考书。				
5	同高考英语不相关的内容，极少在课堂上被提及。				
6	我觉得高考英语分数不应该在大学入学				

	中占据那么大的决定性。				
7	我经常做单项选择题。				
8	背高考高频词就足够应付高考了。				
9	对于阅读能力的重点突破，可以在短期内提高分数。				
10	考试技巧有时比语言能力更重要。				
11	我用在英语上的学习方法是会随着高考的倒计时变化而变化的。				
12	我花在练习写作上的时间常常少于我花在听力和阅读上的时间。				
13	我的主要学习英语的方法是做高考题。				
14	做多少题决定了考多少分。				
15	在做听力题目的时候，我可以选对答案，即使我没有完全听懂听力材料。				
16	在做阅读题的时候，我可以选对答案，即使我没有完全看懂阅读材料。				
17	熟记模板是一种有效的方法提高写作部分的得分。				
18	英语课上，老师详细的向我们解释了高考的考纲和要求。				
19	我已经完全了解了高考英语的考试要求。				
20	“英语好”不等于“考得高”。				
21	“考得高“不等于”英语好“。				
22	英语高考能够准确的检测我的英语语言水平。				
23	针对英语高考的复习能够满足我未来对于英语语言的需求。				

24	高考英语激励我学的更加努力。				
25	准备高考英语让我更加自信。				
26	准备高考英语让我感到压力。				
27	有时，在备考中，我欲放弃。				
28	我希望以后高考英语能够一年多考。				
29	总体而言，高考英语积极地影响了我对于英语语言的学习。				
30	总体而言，高考英语消极地影响了我对于英语语言的学习。				

第三部分:

1. 第二部分的陈述是否容易理解，如果否，请写明原因：
   是　　否_____

2. 第二部分的陈述是否有表达不清或是有歧义，如果有，请标明题号：
   是 _____　否

3. 第二部分的陈述是否涉及敏感话题，如果有，请标明题号：
   是_____　否

4. 如果你对本次问卷调查有意见和看法，请填写在下面横线上：

感谢你的参与！

Appendix Three：

Revised Questionnaire (*English Version*)

Part One:

Please put an "**x**" in the box ☐ that is in front of the information that suits you. .

No.	Questions	Choice Items	
1	What is your gender?	☐ Male	☐ Female
2	What class types are you in?	☐ Class of Liberal Arts	☐ Class of Sciences
3	Have you taken the NMET before?	☐ Yes	☐ No

No.	Questions	Choice Items			
4	How long have you been taught English?	☐ 9 years or more	☐ 7 to 8 years	☐ 5 to 6 years	☐ 4 years or less
5	What is your average score in the monthly NMET mock exams?	☐150-120	☐119-90	☐ 89-60	☐ Below 60
6	Since freshmen year of high school, your scores in the monthly English exams were _____:	☐ Generally increased	☐ Average the same with significant fluctuation	☐ Average the same with slight fluctuation	☐ Generally decreased
7	How many hours have your spent normally in English learning each week (class-hours excluded)?	☐ 20 or more	☐ 15 to 20	☐ 10 to 15	☐ 10 or less
8	Which one do you think is the most important reason for studying English?	☐ To communicate with English-speaking people.	☐ To fulfil parents' expectations.	☐ To achieve higher score in the NMET.	☐ English is a relatively easy subject to learn.

Part Two:

Following are a number of statements with which some people agree and others are not. I would like you to present your opinions of each statement by putting an "**X**" in the box where the column links with the extent of your agreement or disagreement.

For example:

Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
English is the easiest subject in high school.	**X**			

Please avoid answering like this:

Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
English is the easiest subject in high school.		**X**		

Please write '**X**' down, as the example shows, to indicate your degree of disagreement or agreement. Thank you.

No.	Choices / Statements	Strongly disagree	Disagree	Agree	Strongly agree
1	NMET is the dominant topic in English classes.				
2	Almost all the assigned exercises are NMET-related.				
3	English teachers mention NMET frequently during classes.				
4	Textbooks for English subject are essentially the NMET preparation books.				
5	Content that is irrelevant to NMET is seldom mentioned in the English classes.				
6	NMET scores take too much percentage in university admission				

	decisions.				
7	I practice multiple-choice tasks all the time.				
8	Memorizing the high-frequency words is enough for passing the NMET.				
9	Focusing on practicing reading exercises is an effective way to improve the overall score.				
10	Sometimes, test-taking strategies are more important than actual language abilities.				
11	My English learning methods will be modified constantly according the countdown to the NMET.				
12	Comparing reading and listening, hours spent in doing writing exercises are less.				
13	The main English study activity I usually do is practicing NMET stimulated exercises.				
14	The NMET should have speaking section.				
15	Sometimes, I could choose the correct answer without comprehending the listening materials.				
16	Sometimes, I could choose the correct answer without comprehending the reading materials.				
17	Memorizing the written models is an effective way to achieve high scores in the *writing* section.				
18	Teachers have explained the NMET rationale and requirement to use in class.				
19	I am fully aware of the NMET rationale and requirements.				
20	Being ‘Good English learners’ does not equate achieving ‘high score in the NMET’.				
21	‘High score in the NNET’ does not equate being achieving ‘Good English learners’.				
22	NMET is a valid test inferring my language ability.				
23	Studying for NMET is directly related to my future needs in				

	university.				
24	NMET motivates me to work harder.				
25	Preparing NMET helps me develop confidence.				
26	I felt under stressed preparing for NMET.				
27	Sometimes, I felt I wanted to give up English learning.				
28	I would prefer that the NMET could be taken more than once a year.				
29	Generally speaking, NMET positively influences my studying and me.				
30	Generally speaking, NMET negatively influences my studying and me.				

Thanks for your kind participation!

Questionnaire Used for Pilot Study (*Chinese Version*)

第一部分:

请在符合你情况的格子 □画 "X":

No.	问题	选项	
1	你的性别?	□男	□女
2	你在文科班还是理科班?	□ 文科班	□ 理科班
3	请问你以前参加过高考么?	□ 参加过	□ 没有参加过

No.	问题	选项			
1	你学习英语多少年了?	☐9 年或者更多	☐7 到 8 年	☐5 到 6 年	☐4 年或更少
2	你在英语模考中的平均成绩属于下面那一个范围?	☐ 150-120	☐ 119-90	☐ 89-60	☐ 低于 60
3	你平常每周有多少小时花在英语学习上(课堂时间除外)?	☐ 20 或更多	☐ 15 to 20	☐ 10 to 15	☐ 10 或更少
4	从高一开始,我的英语月考成绩_____:	☐总体为上升趋势	☐总体保持平均,中间有较大的浮动	☐总体保持平均,中间有较小的浮动	☐总体为下降趋势
5	你为什么学习英语?(请选择你认为最重要的一项)	☐为了同外国人交流	☐为了满足父母亲的期望	☐为了在高考中取得高分	☐因为英语相对而言比较简单,容易学

第二部分:

以下是一系列意见性的陈述,请通过在相应格子中画"X" 的方式来表达你同意或不同意的程度.

例如:

意见 / 观点	非常不赞同	不赞同	赞同	非常赞同
英语是最简单的学科。	**X**			

请避免如此填写:

意见 / 观点	非常不赞同	不赞同	赞同	非常赞同
英语是最简单的学科。		**X**		

以下为正式问卷内容,请认真填写:

题号	意见 观点	非常不赞同	不赞同	赞同	非常赞同
1	高考相关的英语语言知识是英语课上的主导课题。				
2	几乎所有的英语家庭练习题是高考相关的。				
3	英语课上，“高考”是频繁提起的词。				
4	英语教科书本质上是高考备考书。				
5	同高考英语不相关的内容，极少在课堂上被提及。				
6	我觉得高考英语分数不应该在大学入学中占据那么大的决定性。				
7	我经常做单项选择题。				
8	背高考高频词就足够应付高考了。				
9	对于阅读能力的重点突破，可以在短期内提高分数。				
10	考试技巧有时比语言能力更重要。				
11	我的英语学习计划是随着高考倒计时而变化的。				
12	我花在练习写作上的时间常常少于我花在听力和阅读上的时间。				
13	我的主要学习英语的方法是做高考题。				
14	高考应该有口语考试的。				
15	有时，我可以在没有完全听懂听力材料的情况下选对答案。				
16	有时，我可以在没有完全看懂阅读材料的情况下选对答案。				
17	熟记模板是一种有效的方法提高写作				

	部分的得分。				
18	英语课上，老师详细的向我们解释了高考的考纲和要求。				
19	我已经完全了解了高考英语的考试要求。				
20	“英语好”不等于“考得高”。				
21	“考得高“不等于”英语好“。				
22	英语高考能够准确的检测我的英语语言水平。				
23	针对英语高考的复习能够满足我未来对于英语语言的需求。				
24	高考英语激励我学的更加努力。				
25	准备高考英语让我更加自信。				
26	准备高考英语让我感到压力。				
27	有时，因为备考让我想放弃学习英语。				
28	我希望以后高考英语能够一年多考。				
29	总体而言，高考英语积极地影响了我对于英语语言的学习。				
30	总体而言，高考英语消极地影响了我对于英语语言的学习。				

感谢你的参与！

Appendix Four:

The Information Letter (*English Version*)

Department of Linguistics
Faculty of Human Sciences
MACQUARIE UNIVERSITY NSW 2109
Phone: +61 (0)45 251 8533

Email: jun.wang9@students.mq.edu.au

Chief Investigator's / Supervisor's Name:

Jill Murray

Chief Investigator's / Supervisor's Title:

Dr.

Participant Information and Consent Form

Name of Project: A Washback Investigation of the National Matriculation English Test in China from the Perspectives of Test-takers' Attitudes and Behaviors

You are invited to in a study of *Washback* investigation of the National Matriculation English Test (NMET). The purpose of the study is to investigate the *washback* effects of the NMET on test-takers' **attitudes** and **behaviours ("behaviours" refer to "test preparation activity")**. The notion of "*Washback*" is defined as the influence that a test has on the teaching and learning. The term "*washback*" stems from the common fact that examinations often come at the end of a course, and its impact appears to function in a backward direction, hence "*washback*". The notion of '*washback*' could be investigated from different perspectives, which depends on the overall purpose of the study and contextual condition of the research sites. In this study, we are particularly interested in three aspects: how NMET affects your attitudes towards English learning, whether it motivates you positively, and how this test affects your use of language learning activities.

This study is being conducted by Miss Jun Wang, a current student in the Linguistics Department of the Faculty of Human Sciences in Macquarie University (Tel: +61 (0)45 251 8533; Email: jun.wang9@students.mq.edu.au). This survey being conducted is to reach the requirements of Master of Research under the supervision of Dr. Jill Murray (Tel: +61 (0) 2 9850 9605; Email: jill.murray@mq.edu.au; Department of Linguistics, Faculty of Human Sciences).

If you decide to participate, you will be asked to complete a four-page anonymous questionnaire. There is no open-ended question in the questionnaire, and you do not need to

produce any textual words as answers. Please select the choice-item that is most agreeable to you. The questionnaire will take you about 5 to 15 minutes to complete. Please complete it during the *Self-study Hour* or you can take it home, complete it there, and hand in the next morning. There is no need to leave your personal information on the questionnaire. It is only your opinions that we are interested in.

Any information or details gathered in the course of the study are confidential, except as required by law. No individual will be identified in any publication of the results. Only three persons have access to the data: Dr. Jill Murray (Chief investigator/ Supervisor), Miss Jun Wang (Associate Investigator) and Ms. Liqing Song (Research Assistant). A summary of the results of the data can be made available to you on request. Please contact the co-investigator by her Email (jun.wang9@students.mq.edu.au). She will send the data to you individually.

Participation in this study is entirely voluntary. You are not obliged to participate and if you decide to participate, you are free to withdraw at any time without having to give a reason and without consequence. If you have any questions, feel free to contact any one of the research members mentioned above. The local research assistant's email is ytslq@126.com, and her telephone is +86(0)138-0701-6563.

Investigator's Name: JUN WANG

(Block letters)

Investigator's Signature:

Date:

The ethical aspects of this study have been approved by the Macquarie University Human Research Ethics Committee. If you have any complaints or reservations about any ethical aspects of your participation in this research, you may contact the Committee through the Director, Research Ethics (telephone：＋61 (02) 9850 7854; email: ethics@mq.edu.au). Any complaint you make will be treated in confidence and investigated, and you will be informed of the outcome.

The Information Letter (*Chinese Version*)

语言学系
人类科学院
麦考瑞大学　NSW 2109
电话: +61 (0)45 251 8533

邮箱: jun.wang9@students.mq.edu.au

首席研究员 / 指导老师姓名：

Ｊ．莫瑞

首席研究员 / 指导老师头衔:

博士

研究项目信息及参加同意书

项目名称：高考英语“反驳效应”在高三学生备考阶段英语学习的态度和行为上的反应。

你好！欢迎参加有关高考英语“反拨效应”的问卷调查。我们调查的目的在于辨认高考英语对于高三学生英语学习态度和学习行为（这里“行为”即“学习方法”）

的影响。“反拨效应”是一个语言学研究领域的专业词汇，特指语言考试对于其相应语言教育的影响。“反拨”这个词来源一个普遍的现象，那就是大部分的考试都被安排在某一段课程的结尾，那么这个考试对于教学的影响是以反方向作用的，因此被称为“反拨效应”。“反拨效应”在实践研究中可以从多种角度切入，不通的角度得出的研究成果固然不同。作为一名备考生，你的意见和看法对于本次研究非常重要，我这一次的“反拨效应”研究便是从你，一个备考者，的态度和行为为切入点的。具体来看，主要包括三个部分：高考英语是否影响你对于学习英语的态度，高考英语是否带动你学习英语的积极性，最后，高考英语是如何影响你对于英语学习方法的选择。

我是本次调查的协作调查员，王隽（电话：+61 （0）45 251 8533；邮箱：jun.wang9@students.mq.edu.au），现就读于悉尼麦考瑞大学人类科学院语言学系。这次问卷调查是我硕士毕业论文中的实践数据收集部分。我的导师，即首席研究员，是与我同院同系的莫瑞博士（电话：+61 （0）2 9850 9605；邮箱：jill.murray@mq.edu.au）。

参加本次研究，主要需要你填写一份四页左右的调查问卷。问卷中，所有的问题都为封闭式问题，即不需要任何文字性的回答，题型类似单项选择题。各问题都包含四个意见类选项，只需选择你最为认同的一项即可。整个填写过程大约需要五到十五分钟。请在下课时间填写，或是回家填写第二日上交。这次问卷时全匿名问卷，不需要留下任何个人身份信息，我们只是对于你的意见和看法感兴趣。

同时，我向你保证，问卷上的所有信息将会被处于最严格的保密制度。除了法律上的要求，任何除了研究员，即莫瑞博士，王隽和宋女士，之外的人都将不被允许接触或是阅读问卷信息，所以请放心填写。如果你对本次研究的结果感兴趣，请通过邮件联系研究员（jun.wang9@students.mq.edu.au）。 我会非常乐意将结果发送给你。

是否参加此次问卷是完全自愿的，你有完全的自由选择填写或是不填写。如果没有填写，你不用承担任何的责任或是后果。你有任何疑问，可是随时通过上面的联系方式咨询我们。研究助理，宋女士，的联系方式是：电话：+86 （0）138 0701 6563，邮箱：ytslq@126.com。

辅助研究员姓名：WANG JUN

辅助研究员签名：

日期：

又及： 本次研究中可能涉及的“人权道德”问题已由麦考瑞大学人类研究道德委员会审查并且通过。如果你在参加过程中有任何不满或是抱怨，你可以直接联系委员会的委员提出投诉（电话： +61 （02）9850 7854；邮箱：ethics@mq.edu.au）。任何投诉都将会被严格的保密和审查，处理的结果也会保证通知到个人。

Appendix Five:

The Codebook for SPSS Data Entering

Codebook for Part One:

No. & Variable	SPSS Variable Name	Coding Instructions
1, Identification Number	ID	Number assigned to each case
2, Gender	Gender	1= Male
		2= Female
3, Types of Class	clssinfr	1= Class of Liberal Arts
		2= Class of Sciences
4, Test-taking Experiences	NMET-tking	1= Yes
		2= Not
5, English-learning Experiences	Yr of Enlrning	4= 9 years or more
		3=7 to 8 years
		2= 5 to 6 years
		1= 4 or less than 4 years
6, Average scores on the monthly NMET mock tests	NMETmck	4= 150-120
		3=119-90
		2= 89-60
		1= less than 60
7, The changes of 3-years English monthly tests scores	3yr-mnthtst	4= Generally increased
		3= Averagely the same with dramatically fluctuation
		2= Averagely the same with slightly fluctuation
		1= Generally decreased
8, Hours spent on English learning each week	Hrs in Enlrning	4= 20 or more
		3=15-20
		2= 10-15
		1= less than 10
9, Reasons of learning English	rsnsEnlrning	4= To communicate with English-speaking people.
		3= To fulfill parents' expectations.
		2= To achieve higher score in NMET.
		1= English is a relatively easy subject to study.

Codebook for Part Two

No. & Variables	SPSS Variable Name	Coding Instruction
Item No.	Item No.	4= strongly agree
		3= agree
		2= disagree
		1= strongly disagree

Appendix Six:

The Correlation Matrix from the PCA's Suitability Examination

Correlation Matrix								
	item1	item2	item3	item4	item5	item6	item7	item8
item1	1.000	**.484**	.260	**.305**	.191	.013	.242	.148
item2	**.484**	1.000	**.396**	**.331**	.296	.000	.187	.208
item3	.260	**.396**	1.000	**.537**	**.501**	.097	.171	.253
item4	**.305**	**.331**	**.537**	1.000	**.419**	.123	-.019	.236
item5	.191	.296	**.501**	**.419**	1.000	.178	.035	**.305**
item6	.013	.000	.097	.123	.178	1.000	.023	.203
item7	.242	.187	.171	-.019	.035	.023	1.000	.139
item8	.148	.208	.253	.236	**.305**	.203	.139	1.000
item9	.020	.056	.127	.099	.149	.003	.059	.078
item10	.066	.125	.280	.223	.238	.299	.104	.235
item11	.221	.216	.245	.167	.124	.180	.144	.235
item12	.163	.171	.156	.141	.153	.076	.043	.210
item13	.258	.274	**.332**	.249	.291	.129	.199	.257
item14	.042	.096	.072	.045	-.061	-.154	-.009	-.123
item15	-.033	.096	.045	.063	.075	.212	-.067	.050
item16	.017	.175	.169	.124	.132	.167	.105	.123
item17	.210	.267	.207	.247	.110	.170	.152	.294
Reitem18	-.214	-.257	-.278	**-.333**	-.114	-.094	-.135	-.098
Reitem19	-.192	-.145	-.184	-.243	-.142	-.094	-.100	-.066
item20	-.042	.015	.126	.119	.108	.162	-.024	.010
item21	-.015	.054	.143	.065	.146	.132	-.032	.107
Reitem22	-.080	.015	.102	.051	.107	.219	-.101	.080

Reitem23	-.101	-.016	.043	-.008	.111	.227	-.198	-.009
Reitem24	-.064	-.003	.097	.054	.238	**.361**	.021	.152
Reitem25	-.068	-.039	.099	.070	.140	**.437**	.003	.113
item26	-.008	.011	.170	.069	.192	**.519**	.045	.260
item27	-.041	-.022	.161	.160	.153	**.517**	.052	.270
item28	.066	.002	.044	.032	-.083	.056	.039	-.036
Reitem29	-.103	-.026	.142	.162	.188	.391	-.064	.168
item30	-.023	.031	.180	.082	.272	.535	-.036	.213

Correlation Matrix								
	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16
item1	.020	.066	.221	.163	.258	.042	-.033	.017
item2	.056	.125	.216	.171	.274	.096	.096	.175
item3	.127	.280	.245	.156	**.332**	.072	.045	.169
item4	.099	.223	.167	.141	.249	.045	.063	.124
item5	.149	.238	.124	.153	.291	-.061	.075	.132
item6	.003	.299	.180	.076	.129	-.154	.212	.167
item7	.059	.104	.144	.043	.199	-.009	-.067	.105
item8	.078	.235	.235	.210	.257	-.123	.050	.123
item9	1.000	.142	.204	.189	.191	.087	.121	.227
item10	.142	1.000	.233	.122	.198	-.139	.146	.134
item11	.204	.233	1.000	.247	**.368**	.127	.169	.220
item12	.189	.122	.247	1.000	.258	-.026	.219	**.309**
item13	.191	.198	**.368**	.258	1.000	.046	.131	.175
item14	.087	-.139	.127	-.026	.046	1.000	.079	.091
item15	.121	.146	.169	.219	.131	.079	1.000	**.625**
item16	.227	.134	.220	**.309**	.175	.091	**.625**	1.000
item17	.134	.218	**.323**	.114	.247	-.037	.227	.291
Reitem 18	-.158	-.080	-.221	-.210	**-.334**	-.110	-.164	-.228
Reitem 19	-.159	-.086	-.177	-.041	-.233	-.090	-.151	-.208
item20	.174	.073	.082	.090	.114	.136	.268	.270
item21	.064	.055	.109	.132	.082	.079	.240	**.305**
Reitem 22	.069	.013	.041	-.036	.049	.007	.115	.088
Reitem 23	-.039	.020	.011	-.105	-.014	.018	.136	.078
Reitem 24	.007	.145	-.062	.028	-.032	-.215	.032	.045
Reitem	.009	.135	.028	-.030	-.004	-.184	.106	.090

25								
item26	.068	.200	.106	.065	.160	-.110	.150	.185
item27	.055	.250	.140	.144	.192	-.139	.100	.104
item28	.008	-.026	-.055	.037	.201	.170	.099	.132
Reitem 29	.055	.160	.049	.021	.067	-.114	.101	.080
item30	.076	**.314**	.105	.062	.129	-.165	.121	.129

Correlation Matrix								
	item17	Item 20	Item 21	Reitem 22	Reitem 23	Reitem 24	Reitem 25	item26
item1	.210	-.042	-.015	-.080	-.101	-.064	-.068	-.008
item2	.267	.015	.054	.015	-.016	-.003	-.039	.011
item3	.207	.126	.143	.102	.043	.097	.099	.170
item4	.247	.119	.065	.051	-.008	.054	.070	.069
item5	.110	.108	.146	.107	.111	.238	.140	.192
item6	.170	.162	.132	.219	.227	**.361**	**.437**	**.519**
item7	.152	-.024	-.032	-.101	-.198	.021	.003	.045
item8	.294	.010	.107	.080	-.009	.152	.113	.260
item9	.134	.174	.064	.069	-.039	.007	.009	.068
Item 10	.218	.073	.055	.013	.020	.145	.135	.200
Item 11	**.323**	.082	.109	.041	.011	-.062	.028	.106
Item 12	.114	.090	.132	-.036	-.105	.028	-.030	.065
Item 13	.247	.114	.082	.049	-.014	-.032	-.004	.160
Item 14	-.037	.136	.079	.007	.018	-.215	-.184	-.110
Item 15	.227	.268	.240	.115	.136	.032	.106	.150
Item 16	.291	.270	.305	.088	.078	.045	.090	.185
Item 17	1.000	.074	.120	-.067	-.098	-.047	-.068	.087
Reitem18	-.343	-.208	-.109	.044	-.036	.044	-.032	-.085
Reitem19	-.195	-.088	-.096	.032	.020	-.006	.082	-.002
Item20	.074	1.000	**.548**	.182	.146	.048	.094	.260
Item21	.120	**.548**	1.000	.083	.031	.060	.050	.221
Reitem22	-.067	.182	.083	1.000	**.448**	**.325**	**.407**	.174
Reitem23	-.098	.146	.031	**.448**	1.000	**.364**	**.358**	.159

Reitem24	-.047	.048	.060	**.325**	**.364**	1.000	**.658**	.296
Reitem25	-.068	.094	.050	**.407**	**.358**	**.658**	1.000	**.413**
item26	.087	.260	.221	.174	.159	.296	**.413**	1.000
item27	.073	.130	.132	.179	.110	**.353**	**.411**	.515
item28	.130	.115	.053	.021	-.059	-.175	-.126	.099
Reitem29	-.008	.049	.000	**.302**	**.307**	**.484**	**.571**	.416
item30	.066	.140	.211	.238	.223	**.462**	**.469**	.577

Correlation Matrix						
	item27	item28	item30	Reitem29	Reitem18	Reitem19
item1	-.041	.066	-.023	-.103	-.214	-.192
item2	-.022	.002	.031	-.026	-.257	-.145
item3	.161	.044	.180	.142	-.278	-.184
item4	.160	.032	.082	.162	**-.333**	-.243
item5	.153	-.083	.272	.188	-.114	-.142
item6	**.517**	.056	**.535**	**.391**	-.094	-.094
item7	.052	.039	-.036	-.064	-.135	-.100
item8	.270	-.036	.213	.168	-.098	-.066
item9	.055	.008	.076	.055	-.158	-.159
item10	.250	-.026	.314	.160	-.080	-.086
item11	.140	-.055	.105	.049	-.221	-.177
item12	.144	.037	.062	.021	-.210	-.041
item13	.192	.201	.129	.067	**-.334**	-.233
item14	-.139	.170	-.165	-.114	-.110	-.090
item15	.100	.099	.121	.101	-.164	-.151
item16	.104	.132	.129	.080	-.228	-.208
item17	.073	.130	.066	-.008	**-.343**	-.195
Reitem18	-.126	-.031	-.027	-.004	1.000	**.416**
Reitem19	-.013	-.060	-.069	.003	**.416**	1.000
item20	.130	.115	.140	.049	-.208	-.088
item21	.132	.053	.211	.000	-.109	-.096
Reitem22	.179	.021	.238	**.302**	.044	.032
Reitem23	.110	-.059	.223	**.307**	-.036	.020
Reitem24	**.353**	-.175	**.462**	**.484**	.044	-.006
Reitem25	**.411**	-.126	**.469**	**.571**	-.032	.082
item26	**.515**	.099	**.577**	**.416**	-.085	-.002
item27	1.000	.027	**.575**	.392	-.126	-.013
item28	.027	1.000	-.021	-.066	-.031	-.060
Reitem29	**.392**	-.066	**.552**	1.000	-.004	.003
item30	**.575**	-.021	1.000	**.552**	-.027	-.069

Appendix Seven： Ethics Approval and Amendment

Dear Dr Murray,

Re: "A Washback Investigation of the National Matriculation English Test in China from the Perspective of Test-preparing Candidates' Attitudes and Behavior"(5201400385)

Thank you for your recent correspondence. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee and approval has been granted, effective 12th May 2014. This email constitutes ethical approval only.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:

Dr Jill Murray

Miss Jun Wang

Ms Liqing Song

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 12th May 2015

Progress Report 2 Due: 12th May 2016

Progress Report 3 Due: 12th May 2017

Progress Report 4 Due: 12th May 2018

Final Report Due: 12th May 2019

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Sub-Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms

5. Please notify the Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:

http://www.mq.edu.au/policy

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of approval to an external organisation as evidence that you have approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of ethics approval.

Yours sincerely,

Dr Simon Boag

Acting Chair

Faculty of Human Sciences

Human Research Ethics Sub-Committee

Faculty of Human Sciences - Ethics

Research Office

Level 3, Research HUB, Building C5C

Macquarie University

NSW 2109

Ph: +61 2 9850 4197

Fax: +61 2 9850 4465

Email: fhs.ethics@mq.edu.au

http://www.research.mq.edu.au/

Dear Dr Murray,

RE: 'A Washback Investigation of the National Matriculation English Test in China from the Perspective of Test-preparing Candidates' Attitudes and Behavior ' (Ref: 5201400385)

Thank you for your recent correspondence regarding the amendment request.

The amendment has been reviewed and we are pleased to advise you that the amendment has been approved.

This approval applies to the following amendment:

1. Change in Data Storage - To store the electronic version of the collected questionnaires instead of the hard copies. The PDF and SPSS files are stored securely in a password-protected personal computer, as stated in Section 6.

Please accept this email as formal notification that the amendment has been approved. Please do not hesitate to contact us in case of any further queries.

All the best with your research.

Kind regards,

FHS Ethics

****

Faculty of Human Sciences - Ethics

Research Office

Level 3, Research HUB, Building C5C

Macquarie University

NSW 2109

Ph: +61 2 9850 4197

Fax: +61 2 9850 4465

Email: fhs.ethics@mq.edu.au

http://www.research.mq.edu.au/