

**EXAMINING SECOND
LANGUAGE READING:**

**A CRITICAL REVIEW OF THE SINGAPORE-
CAMBRIDGE GENERAL CERTIFICATE OF
EDUCATION ORDINARY-LEVEL CHINESE
LANGUAGE EXAMINATION**

by

YUN-YEE CHEONG

A thesis submitted in fulfillment of the
requirements for the degree of

Doctor of Philosophy

University College London
Institute of Education

2018

Supervisors:

Professor David Scott, Professor Paul Newton, Ms Katharine Carruthers OBE

I hereby declare that, except where explicit attribution is made, the work presented in this thesis is entirely my own.

© Yun-Yee Cheong 2018

Word count (exclusive of references and appendices):
92,897 words

Abstract

This mixed methods study critically reviews how the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) examines second language reading. The main research question asks, ‘To what degree have the intended measurement objectives of the GCE 1162 reading examination been achieved?’ Four sub research questions address issues of specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters.

Resources drawn on include Singapore Ministry of Education and Singapore Examinations and Assessment Board documents, specifically, examination information booklets, syllabuses, committee reports and annual reviews. Subject matter experts were appointed to analyse the reading comprehension passages and test items from 22 sets of GCE 1162 reading examination papers from 2006 to 2016. Semi-structured interviews were carried out with 22 stakeholders involved in coordination, test design, item construction, marking and reviewing. The interviewees included members of an elite policy group with privileged access to test specifications and procedures. Further interviews were carried out with secondary school Chinese language teachers and students, whose perspectives are seldom considered in validation processes. Opinions were also sought from experts in the field of Chinese as a second language, reading and assessment.

The study begins with an account of the concepts of validity and reading constructs. Chapter 2 discusses the Singapore education and examination system, foregrounding the history of Chinese language education and the bilingual policy introduced in 1966. A methodology chapter follows. Chapters 4 to 8 address separately each of the four sub research questions in which claims, assumptions, supporting evidence and rebuttals are presented. The final chapter, Chapter 9, addresses a posteriori inferences, including scoring, criterion-related components, and washback and impact. A cautious conclusion is drawn, namely that the measurement quality of the GCE 1162 reading examination is at a moderately unsatisfactory level.

Impact statement

This study is among the first to provide an in-depth validation analysis of a national examination in the Singaporean context, specifically, the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162). The primary aim of the study is to evaluate the degree to which the measurement objectives of its reading examination are achieved. In addressing this question, several threats to validity are identified which the Singapore Examinations and Assessment Board (SEAB) and Ministry of Education (MOE) need to help resolve. Rectifying the identified threats to validity will strengthen the measurement quality or validity of the examination.

It is argued that SEAB and MOE should not only clarify the purposes of the GCE 1162 reading examination by presenting the rationale underpinning how examination scores are intended to be interpreted and consequently used but also spearhead extensive theoretical and empirical research on Chinese as a second language reading. Evidence suggests that the GCE 1162 reading examination could encompass more varied dimensions of reading assessment such as multiple text reading for problem-solving, and reading volume and interest and that test designers need to increase the proportion and weightage of higher-order thinking items. Further evidence suggests the need for the reading examination to be more authentic and relevant to Singaporean adolescents. In addition, tangible steps should be taken to ensure that mark schemes are improved and made publicly available. Mark schemes should also state explicitly the principles to which markers must abide in order to facilitate consistent scoring. Together, these findings support the need for a culture of transparency and public access to procedures.

The implications and impact of this study go beyond uncovering the strengths and limitations of the GCE 1162 reading examination by suggesting ways in which to improve measurement quality at the micro-level. The study illuminates perspectives on and understandings of validity and validation, the reading construct and the Singaporean context. While it might not be feasible for a full-scale validation study

such as this to be carried out routinely for all subjects examined by SEAB, the study offers a research foundation and viable frameworks from which smaller-scale and more routine validation studies could be developed.

Analysis of the GCE 1162 reading examination reveals a bi-directional relationship at the meso-level between the validation process and the context in which it is carried out. The specific social, political, cultural and educational environments in which validation occurs inevitably influences its feasibility and meaningfulness. At the macro-level, the study, which draws on a unitary view of validity, demonstrates the adequacy of Weir's (2005) socio-cognitive validity framework and Kane's (2009, 2006) argument-based approach to validation for amassing validity evidence in ways which are feasible.

Acknowledgements

Till, with a sudden sharp hot stink of fox,
It enters the dark hole of the head.
The window is starless still; the clock ticks,
The page is printed.

The thought fox, Ted Hughes

Pursuing a doctoral degree is a test of grit and endurance. I am indebted to my three supervisors at University College London Institute of Education, Professor David Scott, Professor Paul Newton and Confucius Institute Director Katharine Carruthers OBE for their intellectual inspiration and systematic guidance. Their keen sense of academic insight and rigour are qualities that I hope to emulate. My heartfelt appreciation goes to Professor Catherine Snow, too, for her patience and professional advice during my year as a Visiting Fellow with the Harvard Graduate School of Education. Thank you for believing in me.

I would also like to thank the Singapore Ministry of Education and the Tan Kah Kee Foundation for funding my research. Their generous support has enabled me to carry out an in-depth validation study of a national examination in the Singaporean context.

I would further like to thank my wonderful ex-colleagues from the National Institute of Education, Singapore Centre for Chinese Language and Singapore Ministry of Education, especially Professor Chew Cheng Hai, Professor Chin Chee Kuen, Professor Tan Chee Lay, Professor Aw Guat Poh, Professor Ong Yong Peng, Professor Tan Heng Kiat Kelvin, Mrs Catherine Neubronner-Lim, Principal Benedict Keh Chin Chuan and the Master Teachers who fuelled my enthusiasm and interest in the subject of testing and assessment.

I am also grateful to Professor Leong Weng Kee, Professor Zhu Xin Hua, Professor Gordon Stobart, Principal Cheah Chak Mun, Dr Leong See Cheng, Ms Yio Puay Ching and Mr Stuart Shaw who were instrumental in shaping my knowledge of the concepts of validity, the reading construct and the Singaporean context.

My appreciation also goes to my course mates and friends who have enlightened me with their unique insights, especially Ai Lian, Chee Wah, Jenifer, Joo Yeon, Judy, Natalia, Qiuyan, Samantha, Weng Fong, Xin Huan, Xiu-Liu, Yuka, Cher Liek, De Cheng, Kwong Tung, Raymond, Ronald, Sheng Yuan and Wael.

Finally, I would like to thank the interviewees and subject matter experts for participating in my research. A shout out also goes to all my students in the eight years I was a teacher and lecturer. Thank you for having taught me that not everything that matters can be measured, and not everything that is measured matters.

To my family, especially my parents and grandparents
for their unwavering love and support

Table of contents

Abstract	3
Acknowledgments	6
List of figures	14
List of abbreviations	17
Chapter 1: Overview of research	18
1.1 Introduction.....	18
1.2 Background to the problem.....	19
1.3 Validity	22
1.3.1 A gestational period (mid-1800s-1920)	22
1.3.2 A period of crystallization (1921-1951).....	24
1.3.3 The fragmentation of validity (1952-1974)	25
1.3.4 The reunification of validity (1975-1999).....	27
1.3.5 The deconstruction of validity (2000-2012)	33
1.4 Research purpose.....	39
1.5 Conclusion	43
Chapter 2: The reading construct	45
2.1 Introduction	45
2.2 Theories of learning.....	45
2.3 The nature and importance of reading	50
2.4 Purposes and approaches.....	52
2.5 Processes of reading	57
2.5.1 Word recognition	58
2.5.2 Syntactic parsing.....	61
2.5.3 Semantic-proposition encoding	62
2.5.4 Working memory.....	63
2.5.5 Building a mental model.....	64

2.5.6 Forming text and intertextual models.....	65
2.5.7 Metacognitive mechanism.....	67
2.6 Models of reading.....	68
2.7 Reading in a second language.....	78
2.8 Conclusion.....	80
Chapter 3: The Singaporean context	83
3.1 Introduction	83
3.2 Singapore’s education and examination system.....	83
3.2.1 University of Cambridge Local Examinations Syndicate	89
3.2.2 Singapore Examinations and Assessment Board.....	91
3.3 Chinese language assessment issues in Singapore	93
3.3.1 History of Chinese language education in Singapore.....	94
3.3.2 Bilingual policy rationale	95
3.3.3 Limitations and implications of the bilingual policy	101
3.4 Conclusion.....	111
Chapter 4: Methodology	114
4.1 Introduction	114
4.2 Philosophical paradigm.....	114
4.3 Research design.....	116
4.4 Research method: Semi-structured interview.....	119
4.4.1 Pilot study.....	121
4.4.2 Main study.....	123
4.4.3 Data analysis	133
4.5 Research methods: Document analysis and expert judgement.....	136
4.5.1 Pilot study.....	138
4.5.2 Main study.....	141
4.5.3 Data analysis	146
4.6 Ethical concerns.....	147

4.7 Conclusion.....	148
Chapter 5: Specifications and administration.....	151
5.1 Introduction	151
5.2 Interpretive argument	151
5.3 Validity argument.....	152
5.3.1 Purpose.....	152
5.3.2 Construct	161
5.3.3 Administrative structure	168
5.4 Conclusion.....	175
Chapter 6: Test-taker characteristics.....	178
6.1 Introduction	178
6.2 Interpretive argument	179
6.3 Validity argument.....	180
6.4 Conclusion.....	194
Chapter 7: Cognitive parameters	197
7.1 Introduction	197
7.2 Interpretive argument	198
7.3 Validity argument.....	199
7.3.1 Dimensions of reading assessment.....	199
7.3.2 Cognitive demand of items.....	202
7.3.3 Reading approaches.....	219
7.3.4 Item difficulty and discrimination.....	231
7.4 Conclusion.....	234
Chapter 8: Contextual parameters.....	237
8.1 Introduction	237
8.2 Interpretive argument	237

8.3 Validity argument.....	238
8.3.1 Item type.....	238
8.3.2 Mark scheme	248
8.3.3 Discourse mode and text purpose.....	253
8.3.4 Propositional content	258
8.3.5 Readability	264
8.4 Conclusion.....	274
Chapter 9: Concluding remarks	278
9.1 Introduction	278
9.2 Overall validity evaluation of the GCE 1162 reading examination.....	278
9.3 Beyond the a priori inferences	290
9.3.1 Scoring.....	290
9.3.2 Criterion-related.....	293
9.3.3 Washback and impact.....	296
9.4 Directions for future research.....	298
9.5 Key implications and impact.....	299
9.5.1 Practical implications and impact.	300
9.5.2 Theoretical implications and impact.....	304
References	311
Appendix A: GCE 1162 reading examination latest examination format (May 2016 onwards) sample paper	355
Appendix B: GCE 1162 reading examination new examination format (May 2012- November 2015) sample paper.....	368
Appendix C: GCE 1162 reading examination old examination format (May 2006- November 2011) sample paper.....	381
Appendix D: GCE 1162 examination test specifications	392
Appendix E: Letter of invitation to adult interviewees.....	400
Appendix F: Letter of invitation to parent/guardian of student interviewees	403

Appendix G: Interview schedule 406
Appendix H: Excerpts of raw data transcribed and coded using NVivo10 413
Appendix I: Excerpt from Excel evaluation spreadsheet used for expert judgement..... 416

List of figures

Figure 1a: Dominant home language of Primary One Chinese students (1980-2009) (MOE, 2011: 92).....	20
Figure 1b: Timeline of the evolution of validity theory based on Newton and Shaw (2014), Shaw and Crisp (2011) and Brennan (2006).....	23
Figure 1c: Six forms of validity evidence according to Messick (1995a).....	28
Figure 1d: Progressive matrix of validity according to Messick (1989b).....	30
Figure 1e: Adaptation of Weir’s (2005) socio-cognitive framework.....	35
Figure 1f: Claims established around the four a priori inferences in the present study.....	37
Figure 1g: The multi layers of a validation study.....	41
Figure 2a: Reading approaches related to different reading rates and cognitive processes (Grabe & Stoller, 2002).....	55
Figure 2b: Baddeley and Hitch’s model of working memory (Baddeley, 2000: 21).....	64
Figure 2c: Munby’s taxonomy of reading subskills (Munby, 1978).....	71
Figure 2d: Excerpt from the <i>English Language Syllabus 2010</i> , Singapore (CPDD, 2014: 45) ..	73
Figure 2e: Comparison of cognitive levels in reading between Syllabus 2011, Anderson and Krathwohl’s (2001) revised Bloom’s taxonomy and Zhu’s revised Bloom’s taxonomy (2015)....	76
Figure 3a: Outline of Singapore’s education and examination system (Tan, 2012: 47).....	86
Figure 4a: Overview of the mixed methods research design used in this study.....	118
Figure 4b: Interview guide for the pilot study.....	122
Figure 4c: Biographical data of the 22 interviewees in the pilot and main study.....	124
Figure 4d: Final list of categories and codes for tagging qualitative data.....	135
Figure 4e: The eight parameters used for evaluating the 22 sets of GCE 1162 reading examination papers.....	140
Figure 4f: The number of official GCE 1162 reading examination papers, passages and items available from 2006-2016.....	142
Figure 4g: GCE 1162 (and GCE 1160) reading examination formats.....	143

Figure 5a: Purposes attributed to the GCE 1162 reading examination by different stakeholders	156
Figure 5b: Provisional evaluation status of the assumptions underpinning the specifications and administration inference relating to Singapore’s GCE 1162 reading examination.....	176
Figure 6a: Chall’s six stages of reading development (adapted from Chall, 1996)	181
Figure 6b: Provisional evaluation status of the assumptions underpinning the test-taker characteristics inference relating to Singapore’s GCE 1162 reading examination.....	196
Figure 7a: Cognitive level examined by items in the GCE 1162 reading examination (May 2006-May 2016).....	203
Figure 7b: Breakdown by cognitive level in the GCE 1162 reading examination (May 2006-May 2016)	204
Figure 7c: Breakdown by specific cognitive level in the GCE 1162 reading examination (May 2006-May 2016).....	205
Figure 7d: Breakdown of reading items by cognitive level across examination formats	206
Figure 7e: Breakdown of reading scores by cognitive level across examination formats	207
Figure 7f: Reading scores for LOT and HOT items across examination formats.....	208
Figure 7g: Breakdown of reading items by specific cognitive level across examination formats	209
Figure 7h: Breakdown of reading scores by specific cognitive level across examination formats .	210
Figure 7i: Breakdown by reading level in the GCE 1162 reading examination (May 2006-May 2016).....	222
Figure 7j: Breakdown of reading items by reading level across examination formats	223
Figure 7k: Breakdown of reading scores by reading level across examination formats.....	224
Figure 7l: Reading scores for local and global items across examination formats	225
Figure 7m: Breakdown by reading type in the GCE 1162 reading examination (May 2006-May 2016).....	227
Figure 7n: Breakdown of reading items by reading type across examination formats	228
Figure 7o: Breakdown of reading scores by reading type across examination formats	229
Figure 7p: Reading scores for expeditious and careful items across examination formats	230
Figure 7q: Provisional evaluation status of the assumptions underpinning the cognitive parameters inference relating to Singapore’s GCE 1162 reading examination	235

Figure 8a: Breakdown of passages by discourse mode in the GCE 1162 reading examination (May 2006-May 2016).....	254
Figure 8b: Breakdown of passages by discourse mode across examination formats	256
Figure 8c: Frequency of passages of different discourse modes by examination format (May 2006-May 2016).....	257
Figure 8d: Breakdown of passages by topic in the GCE 1162 reading examination (May 2006-May 2016)	260
Figure 8e: Criteria for measuring the literary merit of the GCE 1162 passages	263
Figure 8f: Basic numerical break down of passages in the GCE 1162 reading examination (May 2006-May 2016).....	267
Figure 8g: Numerical breakdown of character and word counts in the GCE 1162 reading examination (May 2006-May 2016).....	268
Figure 8h: Numerical breakdown of character stroke counts in the GCE 1162 reading examination (May 2006-May 2016).....	269
Figure 8i: Breakdown of average sentence length in the GCE 1162 reading examination (May 2006-May 2016).....	272
Figure 8j: Provisional evaluation status of the assumptions underpinning the contextual parameters inference relating to Singapore’s GCE 1162 reading examination	276
Figure 9a: Summary of provisional evaluation status of the assumptions underpinning the specifications and administration inference relating to the GCE 1162 reading examination....	281
Figure 9b: Summary of provisional evaluation status of the assumptions underpinning the test-taker characteristics inference relating to the GCE 1162 reading examination	283
Figure 9c: Summary of provisional evaluation status of the assumptions underpinning the cognitive parameters inference relating to the GCE 1162 reading examination.....	284
Figure 9d: Summary of provisional evaluation status of the assumptions underpinning the contextual parameters inference relating to the GCE 1162 reading examination.....	286
Figure 9e: Summary of evaluation status of the four a priori inferences underpinning the GCE 1162 reading examination.....	289

List of abbreviations

ABV	Argument-based Approach to Validation
AERA	American Educational Research Association
APA	American Psychological Association
CEFR	Common European Framework of Reference for Languages
CL1	Chinese (Mandarin) as a First Language
CL2	Chinese (Mandarin) as a Second Language
CLCPRC	Chinese Language Curriculum and Pedagogy Review Committee
CPDD	Curriculum Planning and Development Division, of the Ministry of Education, Singapore
CRIE-CFL	Chinese Readability Index Explorer for Chinese as a Foreign Language
CTT	Classical Test Theory
DI	Discrimination Index
GCE 1162	Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination
GCE A-Level	Singapore-Cambridge General Certificate of Education Advanced-Level Examination
GCE O-Level	Singapore-Cambridge General Certificate of Education Ordinary-Level Examination
FI	Facility Index
HOT	Higher-order Thinking
HSK	<i>Hanyu Shuiping Kaoshi</i> 汉语水平考试
IA	Interpretive Argument
IB	International Baccalaureate
IGCSE	Cambridge International General Certificate of Secondary Education
IRT	Item Response Theory
L1	First Language
L2	Second Language
LOT	Lower-order Thinking
MCQ	Multiple-choice Question
MOE	Ministry of Education, Singapore
MTLRC	Mother Tongue Languages Review Committee
NCME	National Council on Measurement in Education
NIE	National Institute of Education
OECD	Organization for Economic Co-operation and Development
PAP	People Action's Party
PISA	Programme for International Student Assessment
PSLE	Primary School Leaving Examination
SAQ	Short-answer Question
SEAB	Singapore Examinations and Assessment Board
SME	Subject Matter Expert
Syllabus 2011	Secondary Chinese Language Syllabus 2011
UCLES	University of Cambridge Local Examinations Syndicate
UNESCO	United Nations Educational, Scientific and Cultural Organization
VA	Validity Argument

Chapter 1 Overview of research

1.1 Introduction

This study examines how the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) assesses second language reading ability (see Appendices A, B and C for sample papers and Appendix D for test specifications).¹ Ensuring that a high level of validity for high-stakes national examinations such as the GCE 1162 is upheld should be the fundamental aim for policy makers and test designers. Constructing a validation study to collect and analyse evidence for and against the validity of an examination is, therefore, indisputably necessary (Shaw & Crisp, 2012). This study is essentially a mixed methods validation study, informed by Weir's (2005) socio-cognitive validity framework and Kane's (2006) argument-based approach to validation (ABV). The study is also driven by an in-depth understanding of validity, the reading construct, and the Singaporean context. The focus of the study is on a priori, or before-the-test event, validation components, namely, specifications and administration, and test-taker characteristics, as well as cognitive and contextual parameters. As one of the first detailed validation studies of a national examination in Singapore, this study makes a significant contribution to the field of testing and assessment.

Chapter one begins with a brief overview of the background to this validation study of the GCE 1162 reading examination and introduces the pivotal concepts of validity, the reading construct, and the Singaporean context each of which will be revisited throughout the study. Next is a section that expands on the concept of validity and highlights the two validation frameworks used in the study, namely an adaptation of Weir's (2005) socio-cognitive validity framework and Kane's (2006) ABV. The section is followed by an examination of the key research purposes of this study where the main research question and sub research questions are also set out. The chapter concludes with an outline of the structure of the study. Taken together, the

¹ The subject code for the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination has changed from GCE 1162 to GCE 1160. From the May/June 2016 examination onwards, GCE 1160 has been used. For ease of reference, GCE 1162 and GCE 1160 will be referred to collectively as GCE 1162 in this study.

five sections in Chapter 1 identify current gaps in knowledge and establish the potential contribution of this study to the field of testing and assessment.

1.2 Background to the problem

Several notable academics (Tan, 2016; Qi, 2012; Chin, 2011; Guo, 2011) have pointed out that the language environment in Singapore is complex, due to the use of multiple and very different languages by various ethnic groups. Ethnic Chinese constitute the majority of the population in Singapore at 76.1% (Prime Minister's Office, Singapore Department of Statistics, Ministry of Home Affairs, Immigration and Checkpoints Authority & Ministry of Manpower, 2017) and are well represented in all levels of society, as well as politically and economically. Under the government policy of bilingual education, first adopted in 1966, it is mandatory for most Chinese in Singapore to study Chinese² as a second language (CL2), otherwise known as their mother tongue subject, at primary and secondary levels. Ethnic Singaporean Chinese, however, exhibit a full spectrum of language proficiency, from speaking Chinese as a first language to speaking Chinese as a foreign language. In addition, ethnic Singaporean Chinese are considerably dissimilar from their counterparts in the native Chinese environments of mainland China, Hong Kong and Taiwan, not only in language use and exposure, but also in cultural identity (Tan, 2016).

Not only is the Chinese language environment in Singapore complex, it is also fast evolving. According to the results of a language survey which is obligatory for all incoming Primary One Chinese students, the Singapore Ministry of Education (MOE, 2011) has established that the percentage of Chinese students from English-speaking families has steadily increased since 1980, reaching 59% in 2009 (see Figure 1a). Recent studies imply that home language background and language proficiency are closely related, and the shift of home language from Chinese and Chinese dialects to English could lead to a decline in Chinese language standards (Guo, 2011; Tan, 2011). It should be noted, however, that Chinese language proficiency within the

² Chinese language in this study refers to Mandarin, Standard Chinese or the *Putonghua* equivalent in mainland China and does not include Chinese dialects such as Cantonese or Teochew.

group of Singaporean students from English-speaking families varies hugely, as does their cultural identity. As findings from the Mother Tongue Languages Review Committee Report (MTLRC, 2011) suggest, on the one hand there is a group of students from English-speaking families who use Chinese at home almost as often as English, on the other hand, there are students from English-speaking families who are nearly monolingual. Singapore's diverse language landscape is therefore more nuanced than an English-speaking versus Chinese-speaking family dichotomy as academics such as Tan (2016) and Beardsmore (2003) observe.

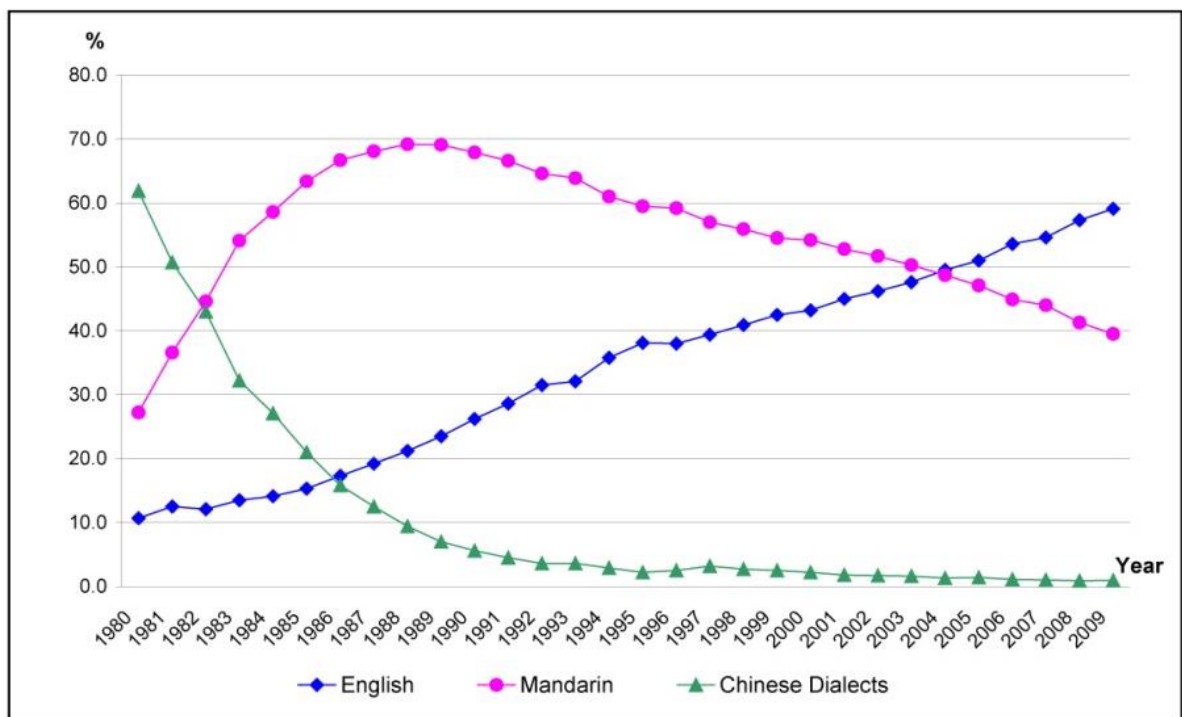


Figure 1a: Dominant home language of Primary One Chinese students (1980-2009) (MOE, 2011: 92)

The unique and changing language landscape in Singapore has brought about major education reforms that will be explored in greater detail in Chapter 3. Whilst there has been considerable effort expended by MOE to revamp the CL2 curriculum and associated pedagogies, assessment, especially national examinations, has been less centre stage. To demonstrate that national CL2 examinations cohere with the characteristics and needs of Singaporean students, validation studies must be carried out regularly. Validation studies are also needed to ensure that the examinations are well grounded in the language ability constructs they are proposing to measure. In

response to the need for validation studies, this study investigates the validity, or measurement quality, of the reading paper of the GCE 1162 examination. Designed with guidelines from the University of Cambridge Local Examinations Syndicate, administered and scored twice a year (May/June and October/November) in Singapore, the GCE 1162 examination has the largest number of test-takers among all Chinese language national examinations at the secondary level. Test-takers are assessed on all four language skills—reading, writing, listening and speaking at the examination. The study scrutinizes the reading paper of the examination for three reasons.

First, the reading paper (Paper 2) carries the highest weightage of 35% of the entire GCE 1162 examination. It is thus essential for policy makers and test designers to possess a solid understanding of the nature, approaches, processes and models of reading comprehension. These aspects of reading comprehension will be examined in Chapter 2 on the reading construct. Second, MOE has in recent years placed great emphasis on nurturing a generation of life-long CL2 readers and learners with higher-order thinking skills (CPDD, 2011). At the same time, the development of 21st century competencies in Singaporean students, a central tenet of Singapore's education policy for the past decade (MOE, 2010a), demands a high level of reading proficiency. Students must be able to read fluently, frequently and broadly to become learners with global awareness, and cross-cultural and creative thinking skills (MOE, 2010a). Given this policy background, reading warrants special attention. Third, the advent and spread of the Internet has changed the way adolescents today read and process information. As Leu, Kinzer, Coiro, Castek and Henry (2013: 1150) assert, 'To have been literate yesterday, in a world defined primarily by relatively static book technologies, does not ensure that one is fully literate today where we encounter new technologies'. It is therefore timely to re-examine the design of the reading paper of the GCE 1162 examination, referred to from this point onward as the GCE 1162 reading examination.

To conclude, the GCE 1162 reading examination is the principal national school-leaving reading examination for the majority of CL2 secondary school students in Singapore. The high-stakes nature of the examination underscores the importance of collecting evidence by which to judge its measurement quality. In this mixed

methods study, documents including examination information booklets, syllabuses, committee reports and annual reviews were examined. Subject matter experts were also appointed to analyse the reading comprehension passages and test items from 22 sets of GCE 1162 reading examination papers from the years 2006 to 2016. Semi-structured interviews were subsequently carried out with 22 stakeholders in roles as varied as those of test-taker, teacher, head teacher, master teacher, curriculum specialist, assessment specialist, academic, item constructor and marker. Validity is at the heart of the study and will be discussed in the next section.

1.3 Validity

Validity, the hallmark of *measurement quality* in testing, is the ‘single most important criterion’ for developing and evaluating a test (Koretz, 2008: 215). The concept of validity is not new—conceptualizations of validity are apparent in assessment literature from around the turn of the twentieth century. Validity is, however, by no means a simple or straightforward concept. Since its inception, notions of validity have evolved significantly and its very nature remains heavily contested within the field of assessment (Newton & Shaw, 2014: 1). Measuring the knowledge and skills that a student has acquired during a course is unlike measuring an objective property such as length or weight—measuring educational achievement is less direct and more resistant to definition. Yet, educational outcomes can be high-stakes in terms of consequences, thus the importance of validity. Before I proceed to evaluate the GCE 1162 reading examination, it is imperative to clarify the concept of validity by tracing its roots and development through the gestational, crystallization, fragmentation, reunification and deconstruction periods (Newton & Shaw, 2014). An understanding of the concept of validity in turn elucidates how a validation study can be framed and conducted.

1.3.1 A gestational period (mid-1800s-1920)

The concept of validity may be better appreciated by tracing its evolution (see Figure 1b). This outline also lays the foundation for later discussions of currently prevalent validation frameworks, in particular, Weir’s (2005) socio-cognitive validity framework and Kane’s (2006) ABV. The history of validity theory begins with the

profound changes accompanying the widespread introduction of written tests in the mid-nineteenth century, especially in Europe and North America. As an American industrialist observed, public examinations were ‘one of the great discoveries of nineteenth century Englishmen’ (Roach, 1971: 3).

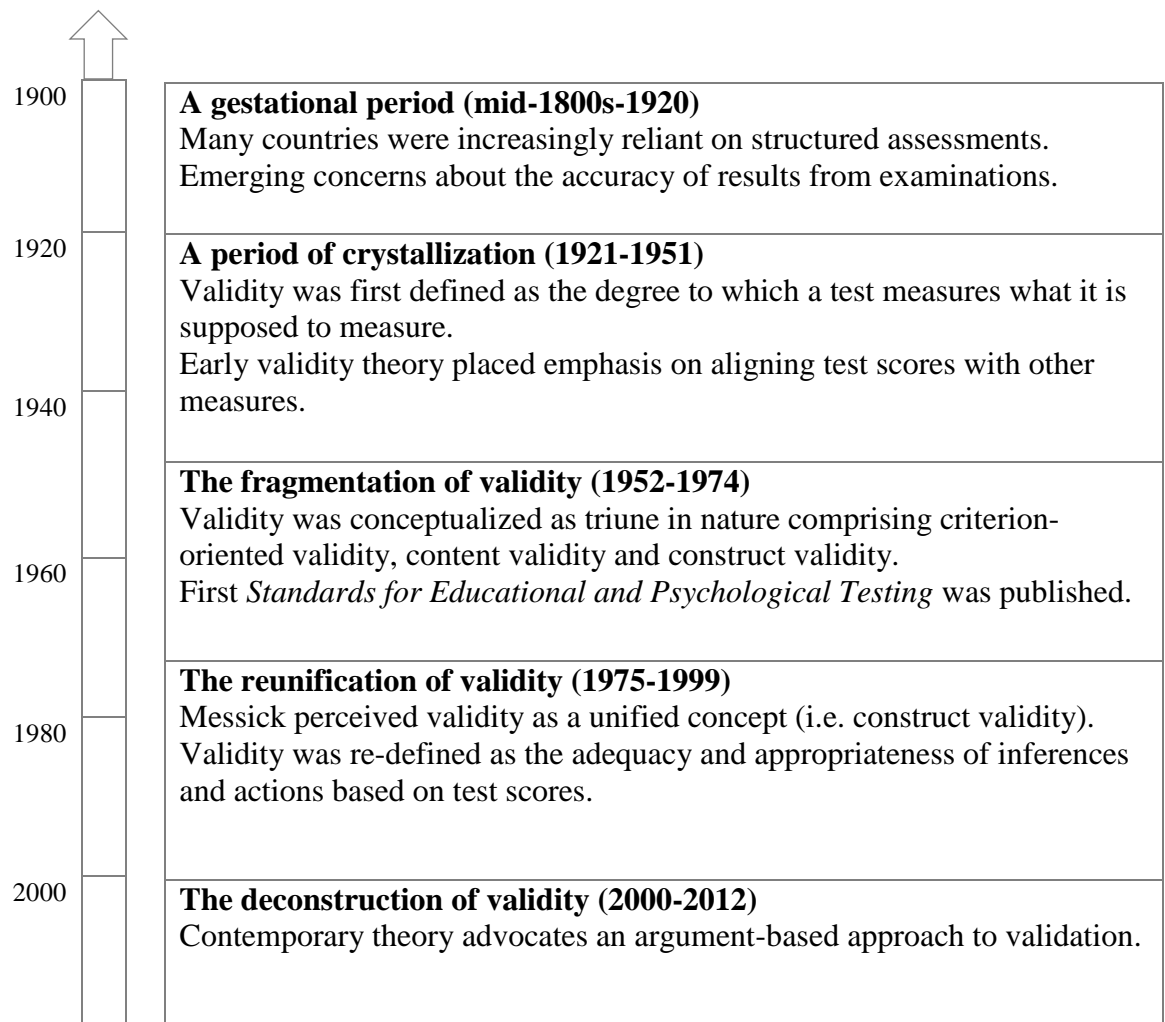


Figure 1b: Timeline of the evolution of validity theory based on Newton and Shaw (2014), Shaw and Crisp (2011) and Brennan (2006)

The rapid spread of public examinations is grounded in a belief that fair and transparent examinations would promote meritocracy and egalitarianism, and would, if universally extended, have a positive influence on society generally. Prior to public examinations, entrance to universities and access to many of the professions depended on a variety of informal routes, where nepotism and favouritism were often prevalent (Black, 2003). The need for social justice and the pressure from an industrial economy created new demands for public examinations. Consequently,

local examinations for schools in England were introduced by the Universities of Oxford and Cambridge. Under the guidance of Mann, schools in the United States of America also began to move from the relatively subjective oral examinations to more standardized and objective written ones. By the end of the nineteenth century, a huge testing and assessment industry had emerged. The burgeoning use of tests naturally led to rising concerns over quality control and regulation which were to become known as the validation process.

1.3.2 A period of crystallization (1921-1951)

One of the most significant characteristics of any quality assessment is validity. In 1921, the North American National Association of Directors of Educational Research provided the first definition of validity as ‘the determination of what a test measures and of how consistently it measures’ (Buckingham et al., 1921: 80). For the next few decades, validity in testing and assessment was generally understood to mean discovering whether a test ‘measures accurately what it purports to measure’ (Kelley, 1927: 14), or uncovering the ‘appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure’ (Henning, 1987: 170). The validity of a test was established through a systematic process known as validation.

The focus of the concept of validity was on the test itself; and validation, or an investigation into validity, was based primarily on empirical evidence provided through correlation. From an operational perspective, Bingham (1937: 214) stated that test scores should be correlated with ‘some other objective measure of that which the test is used to measure’. This view was shared by several renowned measurement theorists at the time, including Cureton (1951), Gulliksen (1950) and Guilford (1946). Validation was to address the question of how well a test estimated the criterion, which could be defined in terms of ‘performances of the actual task’ (Cureton, 1951: 623). A test was deemed valid for any criterion for which it provided accurate estimates.

The judgements of subject matter experts were often used to decide the criterion against which to measure the accuracy of the results. The main limitation of this method was the difficulty in obtaining an adequate criterion. In some cases, for

example achievement tests, it may be difficult to implement a criterion that is clearly better than the test itself; and in other cases, for example tests measuring intelligence and creativity, it may be difficult to even conceptualize a satisfactory criterion (Guion, 1998; Cronbach, 1980b, 1971). As Ebel (1961: 642) commented, ‘The ease with which test developers can be induced to accept as criterion measures quantitative data having the slightest appearance of relevance to the trait being measured is one of the scandals of psychometry’.

1.3.3 The fragmentation of validity (1952-1974)

The 1950s began with the publication of an early proposal that led to the first *Standards for Educational and Psychological Testing* by the American Psychological Association (APA, 1952). Produced with the intention of promoting the sound and ethical use of tests, it quickly gained recognition as a basis for evaluating the quality of test practices. By the time the *Standards for Educational and Psychological Testing* was revised in 1966 by the APA, American Educational Research Association (AERA) and National Council on Measurement in Education (NCME), validity was conceptualized as being triune in nature, comprising criterion-oriented, content and construct validity (APA, AERA & NCME, 1966). Criterion-oriented validity concerns the relationship between a particular test and other measures or criterion. It can be divided into concurrent validity, which compares the test with an existing similar measure, and predictive validity, which assesses whether the test foretells later performance on a related criterion. Content validity refers to the extent to which the content of a test is appropriate or representative of the domain that is to be tested. Construct validity defines how well a test relates to its underlying theoretical concepts. The continuing impact of criterion-oriented, content and construct validity was such that Guion (1980) referred to them as the holy trinity, and validity theory and validation practice became fragmented along these axes. It was not uncommon for validity researchers to assume that they were allowed to ‘pick and choose’ from the three, selecting only the type that would best support interpretations.

Of these three types of validity, construct validity warrants special attention. When it was first introduced by Cronbach and Meehl (1955), the criterion-oriented model was already well-developed and the content model was often applied to measures of

academic achievement. Models for the validation of measures of theoretical attributes were, however, lacking. It was therefore suggested that construct validity be used 'whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined' (Cronbach & Meehl, 1955: 282) and 'for which there is no adequate criterion' (Cronbach & Meehl, 1955: 299). Rather than simply being an alternative, construct validity was a much more fundamental concern. Cronbach and Meehl contended that even if a test was initially validated using criterion-oriented or content evidence, construct-related evidence would still be desirable in the evaluation process. Soon after, Loevinger (1957) further accentuated the role of construct validity by proposing that it is the whole of validity from a systematic, scientific point of view.

To understand construct validity, it is useful to first define what a construct is. A construct must be defined in such a way that it becomes measurable and that it can be related to other constructs that are different. The early validity theorists were positivistic in their outlook and they assumed that there is a 'psychologically real construct' that has independent existence in the test taker and that the test scores represent the degree of presence or absence of this very real property (Cronbach & Meehl, 1955: 284). Constructs are positioned in a nomological network with other constructs and variables. Meaning is then created by measuring the variables and testing how these relate to the constructs that they define in terms of a theory that establishes relationships among constructs. Only propositions that can be verified relative to empirical evidence are regarded as 'real' or 'true'. The underlying philosophical assumptions have been heavily criticized and Cronbach (1989) was to later refer to his earlier position as 'pretentious'. However, there are important elements of early positivist works that continue to influence validity research, particularly the argument that construct validity lies at the centre of assessment and that at the heart of any validation study is the investigation of the intended meaning and interpretation of test scores.

1.3.4 The reunification of validity (1975-1999)

Cronbach and Meehl's (1955) exposition of construct validity helped to pave the way for the next phase in the history of validity theory. Newton and Shaw (2014) described the period between 1975 and 1999 as the *Messick years* as Messick was to become the most prominent validity theorist of this time. In a seminal treatise on validity in the third edition of *Educational Measurement*, Messick asserted (1989b: 20) that:

Traditional ways of cutting and combining evidence of validity, as we have seen, have led to three major categories of evidence: content-related, criterion-related and construct-related. However, because content- and criterion-related evidence contribute to score meaning, they have come to be recognised as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence.

By the early 1970s, fragmented, separatist views of validity have become deeply entrenched (Newton & Shaw, 2014). In challenging the 'unholy trinity' of validity, Messick (1996, 1994, 1989b) perceived validity as a unitary concept, with construct validity being the overarching concern. It is essential to stress that validity conceptualized as a unified view did not in any way diminish content- or criterion-related sources of evidence but instead subsumed them under construct validity in an attempt to build a robust argument for validity. In other words, validity is a unified but multi-faceted concept and Messick argued that evidence from all relevant validity aspects should be collected prior to judging the interpretations and uses of test takers' results.

As a result of this unified perspective, validation is seen as 'a lengthy, even endless process' (Cronbach, 1989: 151), requiring the continuous monitoring and updating of related information. Six forms of interdependent and complementary forms of validity evidence were subsequently highlighted in Messick's works in the 1990s (1996, 1995a), namely content, substantive, structural, generalizability, external and consequential (see Figure 1c). In effect, these six aspects conjointly function as

general validity criteria or standards for all educational and psychological measurement.

Form of validity evidence	Explanation
Content	The relevance, representativeness and technical quality of assessment content.
Substantive	Theoretical rationale for the observed consistencies in test responses.
Structural	Fidelity of the scoring structure to the construct domain.
Generalizability	The degree to which scores and their interpretations can be generalized across populations, settings and tasks.
External	Evidence of criterion validity and utility; convergent and discriminant validity.
Consequential	The implications of score interpretations such as bias, fairness and distributive justice.

Figure 1c: Six forms of validity evidence according to Messick (1995a)

Another fundamental tenet of Messick's validity theory is that validity is not a property or characteristic of a test. Rather, validity concerns 'the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment' (Messick, 1989b: 13). In short, validity is a feature of the inferences made based on test scores and the uses to which a test is put. It is not a test that is validated, rather, it is a principle for making inferences.

Messick has been criticized for shifting validity to test score inferences, mainly because it seems natural and instinctive to consider validity to be a feature of a test. Borsboom, Mellenbergh and van Heerden (2004), for example, claim that there is no reason why it should be concluded that the term 'validity' can only be applied to test score interpretation. Instead, they suggest that validity is a property of measuring

instruments or tests, that encodes whether they are sensitive to variation in a targeted attribute. Fulcher and Davidson (2007: 279) also question, ‘if a test is typically used for the same inferential decisions, over and over again, and if there is no evidence that it is being used for the wrong decisions, could we not speak to the validity of that particular test—as a characteristic of it?’ Despite these concerns, Messick’s view of validity as interpretation remains basically unchallenged. I share Shaw and Newton’s (2012: 4) perception that ‘there is no argument [to counter Messick’s notion of validity] since validity is conditional: conditional upon the observance of administration guidelines; conditional upon the group being assessed; conditional upon the context within which assessment occurs’. Since validity is conditional, a test cannot be valid by virtue of its features alone as Messick posited.

Messick’s concern with score interpretations and uses necessarily raises the issue of test consequences and washback effect. The much quoted progressive matrix (see Figure 1d) was introduced to illustrate how the evaluation of scientific and ethical questions could be integrated within a common validity framework, based on a foundation of construct validity (Messick, 1995, 1989b). Logical distinctions are made between empirical evidence for construct validation which forms the evidential basis, and functional impacts on social systems and values, including unintended negative effects, which form the consequential basis. Messick similarly distinguishes arguments for construct validation based on analyses of test interpretation and test use. Together, these aspects form Messick’s four-faceted progressive matrix.

Ultimately, Messick proposes that the social consequences of test use must be appraised. Aspects such as washback and ethics are part of validity, as are administration procedures and the test environment. Messick also argues that all test constructs and score interpretations involve questions of values. All test result interpretation inevitably reflects the values to which the assessor adheres. Messick’s progressive matrix is a major step in aiding understanding of validity and serves as a theoretical foundation for future research on validity and validation (see Figure 1d).

	Test interpretation	Test Use
Evidential basis	Construct validity	Construct validity + Relevance/Utility
Consequential basis	Construct validity + Value implications	Construct validity + Relevance/Utility + Value implications + Social consequences

Figure 1d: Progressive matrix of validity according to Messick (1989b)

Some validity researchers dispute whether social consequences of test use is a dimension of validity (e.g. Davies & Elder, 2005; Popham 1997; Bellack & Herson, 1984). Popham (1997: 9) writes that ‘cluttering the concept of validity with social consequences will lead to confusion, not clarity’. A clearer and simpler definition of validity, according to Popham, would be better understood and used by educational practitioners. Energy could then be channelled into investigating the accuracy of score-based inferences. Popham (1997: 10) concludes that Messick’s matrix ‘did, indeed, cut and combine evidence so that social consequences became a key facet of validity. It’s just that the price to be paid for doing so is far too high’. Popham then deflects some of the difficulties in employing Messick’s matrix by proposing that test-use consequences be addressed separately from validity. In the same vein, Sackett (1998) implies that adding a social dimension to validity could cause ambiguity in validation practice. To maintain the integrity of validation investigation, especially for high-stakes examinations, means that there is little room for the ambiguous. By including social consequences, Sackett (1998: 121) claims that ‘the concept of validity loses its stature as the most important consideration in test development and use’.

Nevertheless, Messick’s notion of validity has become the accepted paradigm in testing and assessment which can be traced in the evolution of the *Standards for Educational and Psychological Testing*. In an early edition of *Standards for Educational and Psychological Testing* (APA et al., 1966), three types of validity, namely content, criterion-oriented and construct validity were described. A later

edition retained the same categorizations but claimed that the three types of validity were closely related (AERA, APA & NCME, 1974). Only a decade later, the categories of validity were abandoned and the unitary interpretation made explicit (AERA, APA & NCME, 1985: 9):

Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the score. The inferences regarding specific uses of a test are validated, not the test itself.

At the turn of the century, the guidelines reiterated the unitary nature of validity (AERA, APA & NCME, 1999: 11):

These sources of evidence may illuminate different aspects of validity but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose. Like the 1985 Standards, this edition refers to types of validity evidence, rather than distinct types of validity.

While Messick's approach remains dominant in validity theory, there have been further developments within the field of testing and assessment that need consideration. According to Shepard (1993: 408), Messick 'offers an integrated but faceted conception of validity that starts with a traditional investigation of test score meaning and then adds test relevance, values and social consequences. Orchestrated this way, typical validity investigations never get to consequences'. In the same vein, McNamara (2006) argues that the integrative nature of construct validation is not

easily understood and its demands are often perceived to be too complex to be implemented. To further complicate matters, Messick does not regard categories in his progressive matrix as watertight but considers the boundaries to be blurred. The complexity of Messick's concept and progressive matrix of validity created the need for validation frameworks to be developed that focus more on practical applications. Changes to validation frameworks will be elaborated next.

In conclusion, although Messick's views are not without contention, this study draws heavily on many of his arguments, the most pertinent nine are summarized below.

Messick's first argument is that all validity is understood to be construct validity. Throughout the study, validity refers to Messick's unified concept of validity unless otherwise specified.

A second argument is that validity is perceived not as a property of tests but a trait of scores. This study may sometimes refer to the validity of GCE 1162 reading examination, but only for convenience and should not be interpreted otherwise.

A third argument is that validity is conceptualized as a unified but multi-faceted concept. In other words, there are distinguishable aspects of validity evidence that can be collected and evaluated. In this study, validity evidence is amassed from four a priori aspects—specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters.

A fourth argument is that this study acknowledges that validity comprises a social dimension as Messick argues. There is a short discussion on the washback and impact of the GCE 1162 reading examination in Chapter 9, however, a full exploration is beyond the scope of this study.

A fifth argument which relates to the social dimension in Messick's conceptualization of validity is taken a step further in this study. Explicit in the study is the recognition that validation processes can be shaped and influenced by the social context in which they are carried out.

Another argument is that the most significant contribution Messick made, arguably, was to reclaim measurement as the focus of validation in all contexts (Newton & Shaw, 2014: 21). The study adopts the working definition of validity as measurement quality and it is by this yardstick that the GCE 1162 reading examination is judged. This is not to trivialize the fact that validity has a social consequential facet.

A seventh argument is that validation is a scientific inquiry which ‘embraces all of the experimental, statistical and philosophical means by which hypotheses and scientific theories are evaluated’ (Messick, 1989a: 6). Hence, this study triangulates qualitative and quantitative data gathered from interviews, document analysis and expert judgement to compensate for the limitations of individual data sets and to exploit their respective strengths.

An eighth argument is that validity is a matter of degree—it is not an all or nothing concept. The overall findings of this study provide an indication of the extent to which evidence supports the measurement quality of the GCE 1162 reading examination.

The last argument is that validation is a never-ending process. Evolving understanding of validity, reading and the Singaporean context necessitates routine and regular validation studies, especially when high-stakes examinations are involved.

1.3.5 The deconstruction of validity (2000-2012)

During the 1990s, understandings of validity were heavily influenced by the theoretical work of Messick in the United States of America. Messick’s (1989b) chapter on validity in *Educational Measurement* was the most cited and authoritative reference in the field of testing and assessment during that decade. Consensus regarding Messick’s unified conception of validity was widespread throughout this decade (Kane, 2001; Shepard, 1993; Cronbach, 1989), with one researcher suggesting that it brought about ‘close to universal consensus among validity theorists’ (Moss, 1995: 6). Yet at the same time, many academics realized that Messick’s unitary conceptualization of validity did not provide enough guidance as to how to engage in validation studies. To bridge the gap between Messick’s validity theory and

validation practice, several academics have since offered validation frameworks that focus more on practical applications. Examples include Newton and Shaw's (2014) neo-Messickian framework, Bachman and Palmer's (2010) assessment use argument framework, Kane's (2006) argument-based approach to validation (ABV) and Weir's (2005) socio-cognitive validity framework. Among these four frameworks, Weir's socio-cognitive framework has been extensively used across a range of validation studies, including studies conducted by Cambridge Assessment for its suite of English as a second language examinations (e.g. Geranpayeh & Taylor, 2013; Taylor, 2011; Khalifa & Weir, 2009; Shaw & Weir, 2007). Likewise, Kane's ABV is 'increasingly gaining credibility as an alternative approach for thinking about validity' (Shaw & Crisp, 2012: 6). Numerous influential academic works have been shaped directly by the ABV, including Shaw and Crisp's (2015, 2012) evaluation of general qualifications in England. Weir's socio-cognitive validity framework and Kane's ABV, are hence selected to structure the present study and are elaborated below.

Weir's socio-cognitive framework is based on a unitary concept of validity but visualizes the validation process within a temporal frame thereby distinguishing the various components of validity evidence that can be collected. The framework is socio-cognitive because taking a language examination, specifically a reading examination, involves test-takers in cognitive processes. At the same time, an examination is a social phenomenon, influencing and influenced by the contexts in which it is designed and administered. Weir's framework comprises both a priori (before-the-test event) validation components of test-taker characteristics, cognitive parameters and contextual parameters and a posteriori (after-the-test event) scoring, criterion-related, and washback and impact components. The six components are arranged in temporal sequence according to the stages in an examination cycle. The present study concentrates on the a priori components in Weir's socio-cognitive framework. Further, a specifications and administration component, which is implicit and subsumed under the cognitive and contextual parameters, is added to the adapted framework. The adapted Weir's socio-cognitive framework used in this study is represented in Figure 1e.

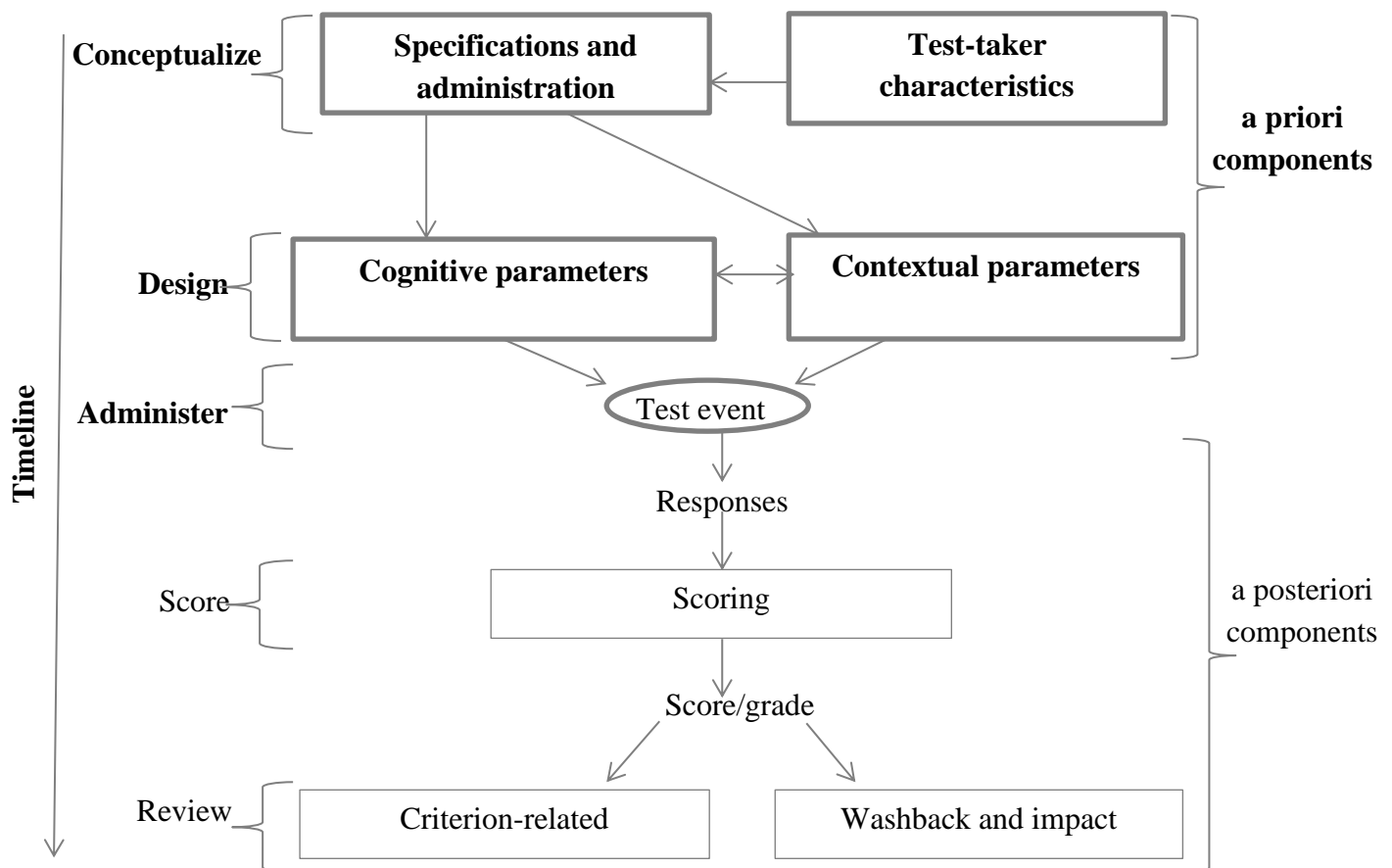


Figure 1e: Adaptation of Weir’s (2005) socio-cognitive framework (components in bold typeface indicate those investigated in this study)

Figure 1e demonstrates how the different validation components fit together temporally as well as conceptually. The timeline runs from top to bottom, beginning when the examination is conceptualized and designed, then administered, and finally when the examination is scored and reviewed, with the arrows indicating the principal direction or directions of influence between components. The timeline in Figure 1e is representative of a complete examination cycle which naturally can have an impact on the next cycle. Mapping the sequence of the validation process onto the order of an examination cycle is helpful as it provides a delineation of ‘*what* should be happening in terms of validation and just as importantly *when*’ (Weir, 2005: 43, original emphases).

The specifications and administration component relates to the intended purposes, constructs and administrative procedures of an examination. The more clearly the

purposes, constructs and administrative procedures are articulated, the more plausible it is for an examination to be constructed and delivered effectively. Next, an examination needs to be conceptualized with the characteristics of its test-takers in mind. Test-taker characteristics that can be explored include gender, age, prior knowledge, interests, needs and socio-economic and cultural background. At the test design stage, the cognitive parameters address the cognitive requirements of an examination. Items, or questions, and tasks must both adequately represent the skills or knowledge an examination is measuring and indicate broader competence beyond the examination. Cognitive processing is mediated by the contextual parameters or the characteristics of the items, texts and mark scheme of an examination. The examination contextually should also be as authentic as possible in order to approximate real-world situations. In Chapters 5 to 8, these four a priori components in the adapted Weir's socio-cognitive framework will be considered in detail in relation to the GCE 1162 reading examination. Although an extensive investigation of the a posteriori components in Weir's socio-cognitive framework is beyond the scope of this study, a brief overview of the scoring, criterion-related, and washback and impact validation components is provided in Chapter 9. These three a posteriori inferences are associated with evidence generated after an examination has been administered, documenting reliability, comparability and consequential effects respectively.

Whilst the adapted Weir's socio-cognitive framework has its merits, for the purposes of this study synthesis with another framework was found necessary to guide the collection of evidence within each of the four a priori components. The argument-based approach to validation framework by Kane (2006) calls for validation to be perceived as assembling a cogent and persuasive argument. The argument in Kane's ABV refers to the way that the validation studies are planned, conducted and reported. Extending ideas from Toulmin (2003), Cronbach (1988) and House (1977), Kane shaped and refined the ABV framework in the 1990s. It was not until a decade later, however, that the ABV framework took root in the field of testing and assessment.

Kane's ABV comprises two primary elements—an interpretative argument (IA) and a validity argument (VA). As Toulmin (2003) has stated, a substantial argument is an attempt to justify an inference or claim. It does not do this by appeal to some reality

beyond the evidence, but it does assume that an inference may be justified according to some level of probability that can be agreed upon by subjecting the argument to criticism and testing. This position is consonant with Kane’s conceptions of IA and VA. In this study, the IA and VAs are established around four inferences, namely the four a priori components in an adaptation of Weir’s socio-cognitive framework. The first step in an IA is to state the claim, which is a statement about one aspect of validity where inferences can be made from the evidence amassed. The four claims in this study are shown in Figure 1f. For a claim to hold true, the assumptions associated with each claim must be substantiated. Assumptions connect the evidence in the VA to the claim in the IA. An example of an assumption for the test-taker characteristics component is that the GCE 1162 reading examination is supported by knowledge of adolescence and adolescent literacy. Assumptions for each claim are laid out at the beginning of Chapters 5 to 8.

Inference	Claim
Specifications and administration	The intended purposes, constructs and administrative procedures of the examination are clearly and sufficiently articulated.
Test-taker characteristics	The characteristics and needs of Singaporean test-takers are taken into careful consideration.
Cognitive parameters	The cognitive requirements of the examination are appropriate and the reading constructs sampled indicate broader competence beyond the examination.
Contextual parameters	The characteristics of test items and passages in examination are appropriate and fair.

Figure 1f: Claims established around the four a priori inferences in the present study

Based on the assumptions, VAs consisting of supporting evidence and rebuttals are constructed. An overall validation evaluation of the examination is subsequently

formed premised upon the strength of the ABV. Kane (2009: 39) states clearly that ‘if the proposed interpretation of test scores is limited, as it is for some observable attributes, the requirements for validation can be very modest. If the proposed interpretations are more ambitious, as they are for traits and theoretical constructs, more evidence and more kinds of evidence are required for validation’. A distinction is made here between observable attributes and theoretical constructs. Observable attributes describe in simple terms how well people perform a task or how they respond to a stimulus. They are, therefore, theory-neutral and ‘can be interpreted without employing the theory currently under investigation. A universe of tasks can be specified without an appeal to cognitive theories of performance for these tasks’ (Kane, 2001: 334). On the other hand, theoretical constructs such as reading are much more ambitious and theory-laden than observable attributes. Satisfactory construct definition would therefore be essential. Formulating an argument when the underpinning constructs are complex is also a much more intensive task.

In sum, looking at the validation process as constructing an argument has a number of advantages. The first advantage, ‘is the guidance [an ABV] provides in allocating research effort and in deciding on the kinds of validity evidence that are needed’ (Kane, 2001: 331). The kinds of validity evidence that are most relevant are those in direct connection with the assumptions in the IA, particularly those assumptions that are the most problematic. For Kane, the ABV is only as strong as its weakest link, thus making potential threats to validity the primary focus of this study. Typically, weak links become obvious in the development of an IA although sometimes it is revealed in the validation process. As Chapelle, Enright and Jamieson (2008) rightly summarizes, an ABV ensures that validation studies are both scientifically sound and logistically manageable. The second advantage is that the ABV is highly tolerant. The ABV ‘does not have to be associated with formal theories’ (Kane, 1992a: 534) and can be used for any type of assessment. Likewise, there is no preference for any particular kind of evidence, allowing for a mixed methods research design. Another advantage is that as each argument must be developed in relation to specific claims, policy makers and test designers are compelled to acknowledge that no examination can be used in any or every situation. Last, the term argument emphasizes the existence of an audience to be persuaded and the need to consider and evaluate

competing evidence, hence promoting a validation study that incorporates both official and unofficial narratives from multiple and diverse stakeholders.

This section has traced the development of validity, foregrounding insights and frameworks pertinent to the study. As evidenced by the rich discussion on how validity has evolved through the periods of gestation, crystallization, fragmentation, reunification and deconstruction, the nature of validity is complex and not without contention. While the debate on validity continues, the consensus is that validity is the paramount concept in the field of testing and assessment (Newton, 2017a; Newton & Baird, 2016). It is no coincidence that the theme in 2015 of one of the largest and most influential testing and assessment conferences, the *Annual Conference of the International Association for Educational Assessment*, was *The Three Most Important Considerations in Testing: Validity, Validity, Validity*. Clearly, the concept of validity is central to any validation study and it is with this understanding that the main research question and sub research questions of this study are articulated in the next section.

1.4 Research purpose

Data obtained from semi-structured interviews, document analysis and expert judgement are triangulated in this embedded mixed methods study to address the main research question:

To what degree have the intended measurement objectives of the GCE 1162 reading examination been achieved?

The main research question in essence concerns validity, defined for the purposes of this study as measurement quality. Four critical sub research questions focusing on each of the four a priori inferences and their accompanying claims are as follows:

1. Are the intended purposes, constructs and administrative procedures of the examination clearly and sufficiently articulated? (Specifications and administration)

2. Are the characteristics and needs of Singaporean test-takers taken into consideration? (Test-taker characteristics)
3. Are the cognitive requirements of the GCE 1162 reading examination appropriate and do the reading constructs sampled indicate broader competence beyond the examination? (Cognitive parameters)
4. Are the characteristics of the test items and passages appropriate and fair? (Contextual parameters)

These research questions which define the scope of the study are informed by the research purposes identified below. Personal reasons sparked my interest in testing and assessment research, and deserve to be mentioned first, even though, according to Maxell (1996: 15), they might ‘bear little relationship to the “official” reasons for doing the study’. As a Chinese language teacher in Singapore, I have taught students with vastly different language proficiency levels. When working at the front line of education, I was constantly motivated to understand the needs, interests and challenges my students had in reading and how these were reflected in the design of the CL2 national examinations. Subsequently posted to the National Institute of Education and the Singapore Centre for Chinese Language, I had an invaluable opportunity to work closely with various stakeholders in education, especially teachers. It was then that I realized how many teachers were unaware of the theories, constructs, processes and procedures that underpin Singapore’s national examinations.

To strengthen the assessment literacy and competency of teachers in Singapore, the Singapore Examinations and Assessment Board (SEAB) has in recent years conducted more than 50 courses and workshops, training over 3,000 teachers (SEAB, 2015c, 2013). Although SEAB’s perceptive interventions were laudable, the courses and workshops covered mainly assessment practice in classroom settings. Opportunities to encourage robust discussion and research pertaining to the national examinations were not provided. As was the case at SEAB’s inception in 2004, much of the work that takes place behind the scenes at SEAB remains fairly obscure to teachers and other users of Singapore’s national examinations. There has also been little conscious effort to publicly release validation studies of national examinations, including the GCE 1162 reading examination. The lack of transparency, a recurring concern in this study, hinders research into Singapore’s national examinations and as Spolsky (2000: 537) rightfully warns, ‘testing is an important but potentially

dangerous component of language teaching. It deserves better understanding than most language teachers have of it and it demands more careful use than most teaching experts seem ready to acknowledge'. Funded by the Singapore Ministry of Education and the Tan Kah Kee Foundation, the present study addresses this gap in testing and assessment which is among the first to provide an in-depth validation study of a national examination in the Singaporean context.

At the micro-level (see Figure 1g), this study concentrates primarily on the measurement quality of the GCE 1162 reading examination paper. Put differently, the study seeks to determine whether there is adequate evidence to support the claims made in the ABV. Supporting evidence and threats to validity for the four inferences are explored in depth from Chapters 5 to 8, and suggestions for ameliorating these threats are brought together in the final chapter, Chapter 9. Hence, research findings have the potential to communicate the strengths and weaknesses of the GCE 1162 reading examination to policy makers and test designers, and provide suggestions for improving the conceptualization, design and administration of the examination.

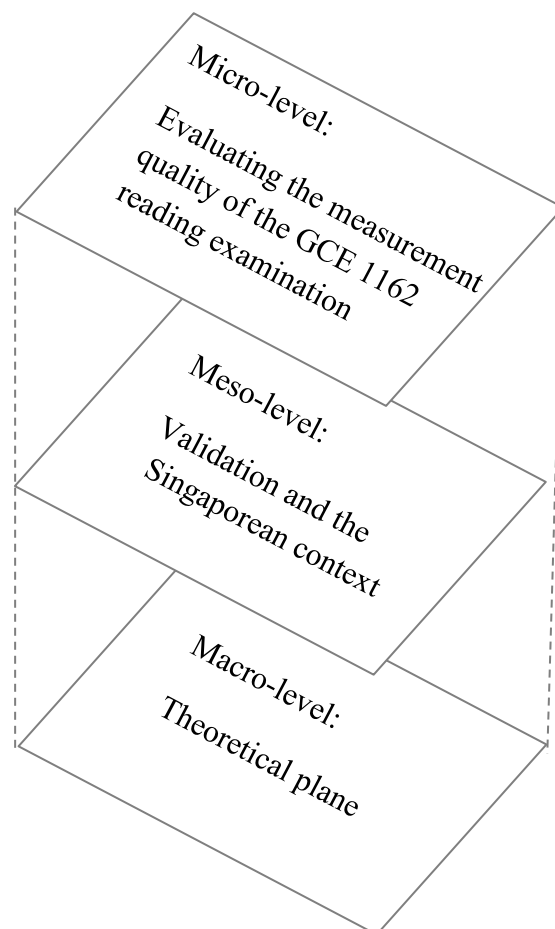


Figure 1g: The multi layers of a validation study

At the meso-level, this study hypothesizes that the broader context in which a validation study is undertaken influences and shapes the process of validation itself (see Figure 1g). Close attention will be paid, therefore, to Singapore's education and examination system and policies that unavoidably facilitate or inflect validation. In addition, the study advocates transparency, and the dissemination and documenting of information pertaining to national examinations. Policy-makers and test designers in Singapore, I argue, are likely to be faced with increasing demands from stakeholders and the general public, as reported in the press in recent years, to demonstrate that there are robust validation processes in place for the national examinations.³

The micro- and meso-levels this validation study will be subsequently superimposed onto a theoretical plane or the macro-level (see Figure 1g). Most of the influential validation frameworks, including Weir's (2005) socio-cognitive framework and Kane's (2006) ABV were conceived in Europe and the United States of America, calling their generalizability and transferability to the Singaporean context into question. This study offers preliminary insights into whether Weir's socio-cognitive framework and Kane's ABV can provide adequate and applicable bases for amassing evidence in feasible ways, as well as raising relevant questions and structuring findings in a cohesive manner that are relevant to Singapore. Although the GCE 1162 reading examination in question focuses on CL2 reading proficiency, outcomes of this validation study may contribute to a better understanding of validity and validation in general, as well as the reading construct and the Singaporean context. As one of the first comprehensive investigations of a national examination in Singapore, this study offers a research foundation upon which other validation studies in Singapore could be built.

³ An example of a press article pertaining to testing and assessment is a letter titled *Primary Six Leaving Examination Minefields to Avoid* (Ho: 2014) published in the most widely read local newspaper, *The Straits Times*. The letter, written by a lead examiner in SEAB, sparked concerns over the quality of Singapore's national examinations. Ho wrote that examiners 'have come across marking guidelines with specimen marked scripts which, besides including spelling, grammar, punctuation or usage errors, also listed controversial mark ranges. This baffled many experienced markers and confused relatively new ones'. SEAB (2014b) quickly responded with an official media reply, which unfortunately contained little detail. Furthermore, Chinese language education and examinations are issues close to the heart of the Chinese community in Singapore. There have been numerous press articles expressing anxiety over falling standards in general, examples include Sin and Ng (2018), Kong (2017), Chen (2013) and Pan (2010).

1.5 Conclusion

Testing plays an integral role in contemporary educational practice and its use is ubiquitous across all formal education settings. Examinations in Singapore have enjoyed lengthy institutionalized traditions and have always been a social leveller and cornerstone of meritocracy. In a speech delivered at the 2014 *Annual Conference of the International Association for Educational Assessment*, the Chief Executive of SEAB, Tan Lay Choo, envisaged the need to review Singapore's current assessment systems, theories and practices to ensure their relevance in the 21st century. Discourse around assessment in Singapore has been dominated by issues of bias and fairness, assessment load and balance of assessment forms. Although various assumptions regarding validity are embedded in these matters, informed conversations about validity and concerted effort devoted to validation have assumed too low a profile. I argue that a new focus on validity will be a step towards increased recognition of validation studies.

This study has begun by setting forth the background to the problem of assessing the measurement quality of the GCE 1162 reading examination, which led to an account of the core concept of validity, the research questions and research purposes. Following this introductory chapter, Chapter 2 examines the reading construct, giving prominence to the nature, purposes, approaches, processes and models of reading relevant to the study. This is accompanied by an exploration in Chapter 3 of the Singapore education and examination system, foregrounding the history of Chinese language education and the bilingual policy introduced in 1966. Chapter 4 concerns methodology, where the philosophical paradigm, research design and research methods are described. Each sub research question, consistent with the adapted Weir's socio-cognitive framework, forms the core of a separate chapter. Chapter 5 addresses specifications and administration, Chapter 6 examines test-taker characteristics. Chapter 7 is concerned with cognitive parameters while Chapter 8 explores contextual parameters. Within these four chapters, the claim, assumptions, supporting evidence and rebuttals are presented in accordance with Kane's ABV. The final chapter, Chapter 9, briefly touches upon the a posteriori validation components, namely those relating to scoring, criterion-related, and washback and impact. Strands of validity evidence are also drawn together and a cautious

conclusion about the measurement quality of the GCE 1162 reading examination is made. Potential threats to validity are summarized and where relevant, improvements to the conceptualization, design and administration of the examination are suggested. Contributions to the understanding of validity and validation, the reading construct and the Singaporean context are also highlighted.

Chapter 2 The reading construct

2.1 Introduction

To assess a construct, that is the attribute or characteristic we wish to test, we need to first understand what the construct is. In order to devise an assessment procedure for reading, we must surely appeal, if only intuitively, to some concept of what it means to read texts and comprehend them. Extensive research on the theory and practice of reading comprehension should be the basis of an assessment development process. Unfortunately reading comprehension research that serves as the foundation of reading examinations has rarely been explicitly and sufficiently defined by test developers. Not surprisingly, the conceptual framework for reading underpinning the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) remains largely unclear. This chapter provides a critical review of research on reading comprehension and postulates the possible implications these findings have for the GCE 1162 examination.

This chapter begins with a theoretical consideration of learning. It functions as a framing device, serving to position the ensuing discussion on reading constructs in its epistemic locales. Next, it introduces the concept of reading and the importance of reading literacy in modern societies such as Singapore. The following section provides a more detailed definition of reading by examining its purposes, approaches and processes. It also explores the various reading models which are currently influential. It then focuses specifically on first language (L1) and second language (L2) relations in L2 reading. Last, this chapter zooms in on reading assessment and highlights how the aspects investigated in the previous sections might affect the way that the GCE 1162 examination is evaluated and reconceptualized.

2.2 Theories of learning

Learning to read and reading to learn are both part of the larger concept of learning, a complicated human experience of acquiring and modifying knowledge, skills, strategies, beliefs and attitudes. Although the precise nature of learning is contentious, most educational professionals accept the general definition of learning as ‘an

enduring change in behaviour, or in the capacity to behave in a given fashion, which results from practice or other forms of experience' (Schunk, 2012: 3)

There are three components that are inherent in this definition. First, it postulates that a human entity has a relationship, both inward and outward with an environment. Learning occurs when people become capable of doing something differently by internalising what was previously external to them in the environment. As learning can only be inferentially assessed, we have to evaluate the extent of learning based on what people display or exteriorize. Second, learning needs to be understood within a socio-historical setting. What is learned is first generated in the society and outside the individual; absolute knowledge untouched by the external world does not exist. People take in the external world and alter it according to subjective, innate principles. Hence, learning is 'both externally and internally mediated, and the form taken is determined by whether the process is cognitive, affective, meta-cognitive, conative or expressive' (Scott & Hargreaves, 2015: 3). Third, learning reflects a change in behaviour potential; it does not automatically lead to a change in behaviour. Learners must be sufficiently motivated to translate learning into behaviour. Likewise, behavioural changes brought about by learning are not always permanent.

A curriculum can be broadly understood as an intended programme of learning (Scott, 2015). It indicates a collection of learning activities and tasks that should happen in the programme and the circumstances in which they can take place. It comprises three sets of standards: curriculum, pedagogic and evaluation. Together, they define what students are expected to know and be able to do at specific stages of their education, how their learning can be scaffolded and how they will be assessed. There are several dimensions to a curriculum, such as the educational, the total and 'hidden' curriculum, the planned versus the received curriculum and the formal versus the informal curriculum (Kelly, 2009). It is not my intention in this study to flesh out these distinctions. Rather, I am concerned with the notions of knowledge and learning and how they affect the curriculum in general and assessment in particular.

The concepts of knowledge and learning are closely intertwined. It is obvious that however we conceive of curriculum and assessment, learning of some kind is central

to it; likewise, any assertions or statements we make about learning are necessarily related to what is (to be) learned. Research in cognitive sciences has shown separate but interconnected forms of knowledge, commonly categorized as declarative knowledge, procedural knowledge and knowledge by acquaintance. Declarative knowledge refers to the factual knowledge a person has or the *knowing-that*. In other words, it can be thought of as the who, what, when and where of information. Thus, a student who is taught that the word 药 (medicine) is pronounced as *yào*, or that 药 is a short story written by *Lu Xun* (鲁迅), the father of modern Chinese literature, can be said to be acquiring declarative knowledge. In contrast to this, procedural knowledge is about *knowing-how*. This is a person's ability to perform actions to complete a task. The student who knows how to read short stories in Chinese, and interprets the themes and characters can be said to have demonstrated procedural knowledge. Last, knowledge by acquaintance is an immediate or unmediated awareness of some propositional truth as Russell (1997) posits. A person is acquainted with an object when they become directly aware of the object itself. For example, a person who regularly reads the literary works of *Lu Xun* and other modern Chinese writers could develop an aesthetic appreciation and judgement of modern Chinese literature. Such knowledge is not simply a matter of learning facts about modern Chinese literature or becoming proficient in some of the procedures for critiquing it. Rather, it is a 'relatedness' or experience with the object which underpins knowledge by acquaintance (Heidegger, 2004).

Many policies within education in Singapore, particularly those pertaining to curriculum and assessment, draw upon these Western philosophical conceptions of knowledge. One such example is the 21st century competency discourse that prioritises procedural knowledge, foregrounded in the last decade in Singapore through buzz terms like 'critical thinking skills, 'collaboration skills' and 'cross-cultural skills'. The Chinese as a second language (CL2) reading curriculum has also adopted a skills-based approach, with the various components of literacy instruction being compartmentalized and sequentially introduced in separate units. The skills taught are thought to be transferable and capable of utilization beyond the context in which the initial acquisition occurs. This emphasis on procedural knowledge has turned the focus away from an explicit learning of declarative knowledge such as

grammar rules and vocabulary. The whole language approach that immerses students in a text rich environment, in the belief that they will learn to read by acquaintance, is also no longer in vogue.

Following this brief discussion of knowledge and the curriculum, a robust curriculum and assessment system would need to take into consideration all three aspects of knowledge. In Heidegger's (2004) analogy of a cabinet-maker's apprentice, the apprentice needs to understand the structure of different cabinets, gain facility in the use of tools and also be attentive to the way the material shows itself, which can be construed as declarative knowledge, procedural knowledge and knowledge by acquaintance respectively. Williams and Standish (2015) point out that there is not always a clear delineation between realms of knowledge for they overlap and interrelate with each other in multiple ways. Recognizing the complexity of knowledge, by implication, opens up the variety of ways in which we can learn. Our views on what is 'useful knowledge', 'powerful knowledge' and knowledge that is 'empowering' lead us towards different understandings of learning (Moore, 2015), which in turn are functionally inseparable from how we define an assessment construct in general and a reading construct in particular.

There are many different kinds of learning theory, each reflecting a deliberate emphasis on a slice of the multidimensional nature of knowledge and knowing. Learning 'has traditionally been the province of psychological theories' (Wenger, 2009: 216). These include behaviourist theories proposing that behaviour can be conditioned (e.g. Skinner, 1974; Watson, 1913; Pavlov, 1897), cognitive theories focusing on how the human mind processes information (e.g. Wenger 1998; Hutchins 1995; Anderson 1983) and constructivist theories contending that learners build their own mental structure while interacting with the environment (e.g. Bruner, 1996; Papert 1980; Piaget 1954). Some learning theories have moved away from an exclusively psychological approach, seeking to bridge the gap between the individual and social reality as demonstrated in the three following examples. First, there are activity theories that consider how a wide range of factors, such as the community, objects and artefacts, work together to impact on learning (e.g. Engeström, 2008). Second, there are socialization theories that are heavily centred upon the development of self-awareness and identity through internalising the norms of a

social group (e.g. Parsons, 1962). Third, there are organizational theories that study the ways in which an individual acquires knowledge within an organization and also how knowledge is created, retained and transferred within the organization itself (e.g. Argyris, 1999).

I shall expand on the constructivist explanation of learning which is central to the content and structure of this chapter. Broadly conceived, constructivism refers to the philosophical viewpoint that knowledge is formed inside individuals and within communities, and that bodies of knowledge are also human constructions (Phillips, 1995). There are, however, many representations of constructivism and no one version should be assumed to be more correct than any other (Simpson, 2002). The divergence amongst constructivist theories emerges from different responses to two major issues. First, although the basic premise of constructivism is that knowledge is not ‘discovered’ and passively absorbed by learners, constructivists differ in the extent to which they ascribe the construction of knowledge entirely to individuals. Some theorists perceive reality as endogenous, in which new knowledge develops out of earlier knowledge primarily in an individual’s mind through cognitive stages (e.g. Piaget, 1969, 1964); whereas others view reality as exogenous, believing that mental structures come to reflect an external reality via teaching and experiences (e.g. Bandura, 1986, 1977). Second, at one end of the spectrum, constructivists are concerned only with the individual learner; at the other end are those constructivists who focus on how human knowledge in general is constructed (Bruning, Schraw & Norby, 2011).

Many combinations of convictions can be held and still considered constructivist. For the purposes of this study, I shall adopt a dialectical perspective of constructivism, which holds that constructions are not always bound to the external world nor do they reside solely in the mind of individuals. This approach also takes into account both individual cognition and social influences on learners as readers. Hence, attention is not only directed to the cognitive processes that operate within the reader but also to the broader society which shapes and constrains these processes.

In this section, I have examined in broad strokes, concepts of knowledge and learning. I argued that an understanding of learning and knowledge is crucial to the

formulating and planning of a curriculum and, by implication, its assessment constructs. In defining the reading construct, a moderate social constructivist approach has been drawn on, as illustrated above, which foregrounds the following three issues. First, the construct of reading comes about as a result of agreements reached in society by influential individuals and institutions, that is, power arrangements in society determine what aspects of reading are given more prominence in summative examinations. Second, reading examinations are designed not only to include individual skills and the knowledge necessary to accomplish reading activities but with an understanding and appreciation of these reading activities as part of a meaningful social practice to achieve a larger communicative goal. Third, it follows that the GCE 1162 reading examination can promote increased engagement and motivation by providing an authentic context and purpose. Keeping these issues in mind, I now offer an extended discussion of the reading construct.

2.3 The nature and importance of reading

Reading is a complex process that has been understood and explained in numerous ways. Those who need to test reading clearly need to develop some idea of what reading is and yet this is an enormous task. Any review, therefore, of the nature of reading, will inevitably be selective, rather than exhaustive.

A meaningful way to begin this section is to provide an initial definition of reading. Reading can be defined as the ability to draw meaning from the printed page and interpret this information appropriately (Grabe & Stoller, 2002). It involves an interaction between the reader and the text (Kucer, 2001; Rosenblatt, 1978), whereby each reader is unique in that they possess certain traits or characteristics that are distinctly applied to each text and situation (Fletcher, 1994). In other words, readers themselves affect the reading process and product. The state of the reader's knowledge, including their world and cultural knowledge, metalinguistic and metacognitive abilities and understanding of the language, genre and subject matter, is one of the most important reader variables. Ever since Bartlett's (1932) research on schema theory, it has been clear that what readers know affects what they understand. Schemata are seen as interlocking mental structures representing readers' knowledge. When readers process text, they integrate new information from the text

into pre-existing schemata. Their schemata influence how they recognize information as well as how they organize and store it. Readers also vary in their motivation to read, their interests and the strategies they use. Just as every reader is different, no two texts are identical. The content, language, structure, readability, intention and even specific word choices of a text influence the interaction where comprehension takes place.

Reading is something that many of us take for granted. Yet, it is fascinating when we remember that we were never born to read. ‘Human beings invented reading only a few thousand years ago. And with this invention, we rearranged the very organization of our brain, which in turn expanded the ways we were able to think, which altered the intellectual evolution of our species’ (Wolf, 2007: 3). It is remarkable, according to the United Nations Educational, Scientific and Cultural Organization (UNESCO), that nearly 83% of the world’s adult population can read to some extent (UNESCO, 2016), with many being able to read in more than one language.

According to the United States of America National Reading Panel Report, reading is one of the most important skills necessary for a happy, productive and successful life (National Institute of Child Health and Human Development, 2001). Without it, opportunities for improving one’s life are limited. The Programme for International Student Assessment (PISA), a worldwide study by the Organization for Economic Co-operation and Development (OECD), also established the importance of reading skills (OECD, 2016). As in the past, the ability to read is inextricably linked with life possibilities and the social trajectories of individuals and groups. The OECD maintains that those with inadequate reading competency have little hope of fully participating in complex societies where people are increasingly required to take on additional responsibility for different aspects of their lives: from active citizenship, to planning their careers, to nurturing and guiding their children. Hence, a major goal for many inter-governmental organizations, non-governmental organizations and educational institutions around the world is to promote greater literacy and we often

hear of efforts to eradicate illiteracy altogether.¹ Reading is often seen as an integral part of our daily routine and an essential first step in cultivating life-long learning.

The *Secondary Chinese Language Syllabus 2011* (Syllabus 2011) places great emphasis on reading literacy according to the Curriculum Planning and Development Division, of the Ministry of Education (MOE), Singapore (CPDD, 2011). As society and technology evolve, CPDD argues, so does literacy. Being able to read in the twenty-first century involves a wide range of abilities and competencies. The literacies today—from reading online newspapers to participating in virtual classrooms—are multiple, dynamic and malleable. The advent of the Internet calls for more effective reading skills and strategies to select, interpret and evaluate the large quantities of information made available to us. Although English is still the first language in Singapore’s schools, Syllabus 2011 endorses the claim that being able to read in an additional language has a positive impact on cognition, shaping the way adolescents think and solve problems. Chinese reading literacy skills, as Syllabus 2011 spotlights, matter not just for individuals but for Singapore’s economy as a whole. In modern societies, human capital—the sum of what the individuals in an economy know and can do—is often seen as one of the most important forms of capital. Despite the fact that CL2 reading literacy in Singapore is seldom viewed as a prerequisite for achievement either in other subject areas at school or at the workplace, CPDD envisages that enabling more adolescents to read in the Chinese language will naturally benefit Singapore, opening the ways to more business opportunities in the future with the economic giant, China. I now proceed to sketch an exploratory map of key issues in reading and indicate their connections to assessment and testing.

2.4 Purposes and approaches

The rudimentary definition of reading in Section 2.3, though useful, is insufficient for analytical purpose. To capture the complexity of reading entails, at the very least, a discussion on the following: Why do students read? What processes are used by

¹ One such effort is the United Nations’ publication *Literacy for Life: Shaping Future Agendas Resolution* (UNESCO, 2014) which calls for ‘intensified efforts from countries and development partners to promote literacy for children, youth and adults, regardless of gender, ethnicity, socio-economic status, and other conditions’.

fluent readers when they read? What are the more widely accepted models of reading? How do these aspects work together to build a general notion of reading?

Students nowadays often face myriad texts, some that they consciously intend to read and others that they just seem to encounter or pick up. Some texts exist in print whilst others are in electronic formats. There is also a wide range of non-continuous texts, such as charts, graphs, forms and information sheets. Students in Singapore's secondary schools are often required to read in formal academic settings, engaging in the synthesizing and evaluation of information which can be demanding. In casual settings, students may read for relaxation or to gain information. In Singapore where Chinese is one of the four official languages, there is an abundance of texts in the Chinese language in the everyday environment—newspapers, advertisements, flyers and signage, to name a few. The founder of modern Chinese linguistics, Shu-Xiang Lv perceives students to be reading in many more ways than they are aware of, 'Reading could occur everywhere. Whether they are poring over a newspaper, browsing through an instruction manual or simply looking at a sign...reading skills allow students to become engaged with the world around them' (Lv, 1987: 90). In summary, reading takes place for varied reasons, all of which have implications for test designs.

The Progress in International Reading Literacy Studies, a large-scale international comparative study of reading literacy in fourth grade students (ages 9 to 10) conducted by the International Association for the Evaluation of Educational Achievement, indicates two broad reading purposes: for literary experience and to acquire and use information (Mullis & Martin, 2015). The early reading of most young students, the study postulates, centres on these two purposes. Reading for literary experience is mainly reading for interest or pleasure and is often accomplished through reading fiction. On the other hand, reading to acquire and use information is performed mainly for learning and is generally associated with informative articles and instructional texts. These purposes are not mutually exclusive, for example, biographies may be read for both literary and informational purposes.

PISA (OECD, 2016; Thomson, Hillman & De Bortoli, 2013) identifies four main reading purposes for adolescents. The first purpose is reading for personal fulfilment and communication, for example, reading fiction, electronic mail and diary-style blogs. The second purpose of reading involves social interaction in public spaces, for example, reading public notices, government blogs and news websites. The third purpose is to learn in educational settings. Educational reading is normally prescribed by teachers and involves acquiring information as part of a larger learning task. Printed textbooks and interactive learning software are common examples of material generated for this kind of reading. The fourth purpose of reading is for occupational reasons. In order to prepare adolescents for the workforce, training them to read for work is essential. This type of reading is also referred to as ‘reading to do’ and typical tasks include searching for a job online or following workplace directions. PISA states that the purposes of reading do not function in isolation from one another and may overlap.

Successful reading, as Linderholm and van den Broek (2002: 778) discern, ‘includes the ability to adjust processing in such a way that learning goals, as a function of reading purpose, are met’. Likewise, Weir, Hawkey, Green and Devi (2012: 214) observe that ‘the multiple reading models that are now acknowledged in second language literature suggest that reading for different purposes may engage quite different cognitive processes or constellations of processes on the part of the reader’. A range of studies in educational psychology and discourse processing have demonstrated that varying reading purposes lead to significant differences in comprehension processing (e.g. Grabe, 2009; Linderholm & van den Broek, 2002). It is therefore necessary for test developers to identify and decide the relative importance, and weight, of different reading purposes in order for reading comprehension tasks in the assessment to be evenly distributed. Examining the test specifications and syllabus for the GCE 1162 reading examination reveals a less than satisfactory explanation of the overarching reading purposes. There is only a brief mention of reading for the appreciation of literary texts and a list of discourse modes, namely narrative, expository, argumentative, informational and functional texts, that students are required to comprehend (CPDD, 2011). Such a lack of clarity raises problems for a validation study of the GCE 1162 reading examination, a theme that I will enlarge on in chapters 5 and 9.

When students read for varying purposes, and with different goals and levels of motivation, they employ, both consciously and subconsciously, a plurality of approaches or mental strategies to negotiate their way into, around and between texts. Different reading approaches also tend to impose differing levels of demand on the reader so as to establish an acceptable level of understanding and detail required by a specific approach (Linderholm, Virtue, Tzeng & van den Broek, 2004). We therefore need to account for the various approaches when we consider any definition of reading.

Carver (1997, 1990) classifies reading approaches into five gears, namely, scanning (Gear 5), skimming (Gear 4), rauding (Gear 3), learning (Gear 2) and memorizing (Gear 1). Each gear relates to a different reading rate, with Gear 5 being the fastest and Gear 1 the slowest. Subsequently, Grabe and Stoller (2002) distinguish seven main reading approaches (see Figure 2a), with 1 and 2 being carried out at high reading speeds (450 to 600 words per minute) and 3 to 6 at lower speeds (200 words per minute and below). Approach 7, reading for general comprehension, is the most basic of the above approaches; it operates at about 300 words per minute for fluent readers.

1. Reading to search for simple information
2. Reading to skim quickly
3. Reading to learn from texts
4. Reading to integrate information
5. Reading to write (or search for information needed for writing)
6. Reading to critique texts
7. Reading for general comprehension

Figure 2a: Reading approaches related to different reading rates and cognitive processes (Grabe & Stoller, 2002)

In reading to search, readers typically scan the text for a specific piece of information or a specific word, for instance, searching through a telephone directory to find a telephone number or an address. Similarly, reading to skim, which involves sampling segments of a text for a general understanding, is a common purpose for reading and

a useful strategy in its own right. An example of this is flipping through a newspaper to get the gist of articles. Reading to learn typically occurs in academic and professional contexts in which readers need to acquire a considerable amount of information from a text, recognize and build rhetorical frames to organize the information and link it to an existing knowledge base. When readers read to integrate information, write and critique texts, they often need to make decisions about the relative importance of complementary or conflicting information and the likely restructuring of a rhetorical framework to accommodate information from various sources. These three purposes are representative of common scholastic tasks. Reading for general comprehension underlies and supports the other approaches for reading. Its demands for processing efficiency can be high, requiring readers to possess reading automaticity and strong skills in coordinating many processes within limited time constraints.

Khalifa and Weir (2009) collapse Grabe and Stoller's seven approaches into two dimensions and four categories, namely reading levels, subcategorized as local or global, and reading types, subcategorized as careful or expeditious. The reader is viewed as a goal setter who selects the appropriate approach to reading as determined by their reading purpose. Local reading occurs at the levels of decoding (word recognition, lexical access and syntactic parsing) and establishing propositional meaning (at the clause and sentence level). Retrieving explicit information at the sentence level is also deemed local reading. Global reading, in contrast, refers to comprehension beyond the sentence level. It concerns the macro-structure level of the text, for example the relationships between ideas and the author's style and intention. Reading approaches can then be further characterized as either careful or expeditious. Careful reading refers to processing a text thoroughly with the intention to extract complete meanings from presented material. Conversely, expeditious reading is quick, selective and efficient reading to access desired information in a text (scanning, skimming and search reading). Both careful and expeditious reading can take place at either a local or a global level, that is, within or beyond the sentence right up to the level of the complete text or texts. We will return to these approaches in more detail when enquiring into the cognitive parameters of the GCE 1162 reading examination in Chapter 7.

There are certainly other ways of classifying the purposes of, and approaches to, reading aside from those listed above (e.g. Linderholm & van den Broek, 2002; Alderson, 2000; Lorch, Kluzewitz & Lorch, 1995). The intention here is not to list every possibility, rather, it is to make the point that reading occurs under various circumstances. Purposes and approaches are inextricably linked to the processes of reading, which we will review below.

2.5 Processes of reading

When we are cognizant of the complexity of reading, its multiple purposes and its many approaches, it becomes clear that the processes that operate when we read must also be complex. The reading process refers to the active and multifaceted interaction between a reader and the text. During this process, clearly, many things are happening. When the reader looks at the text, their brain tries to recognize the words and derive their meaning. A fluent reader will also construct or reconstruct meaning from what they read, connecting what is read to their prior knowledge and experience. Words may be transformed into images; feelings may be evoked. The reader may also be passing judgement: Is the text interesting, useful, entertaining, challenging? Does the reader need to adjust their expectations or strategies to better understand the text? The reader may be fully absorbed in reading or consciously reflecting on the ease or difficulties experienced when reading.

This scenario serves to remind us of the difficulty in trying to understand how humans think and what the mind does in its efforts to process texts and make meaning. Reading comprehension, by its very nature, is normally silent, internal and private. As Pearson (2009: 4) remarks, ‘we are seldom privy to the “aha!” that occurs when there is a “meeting of the minds” between author and reader’.

It is useful to think of the reading process as assembling different building blocks. At the fundamental level are the lower-level processes, including word recognition, syntactic parsing and semantic-proposition encoding, cemented together by working memory. Labelling these components as lower-level ‘does not mean that they are simple or undemanding’, rather they ‘have the potential to become strongly automatized’ (Grabe, 2009: 21) and this automaticity is a prerequisite for fluent

reading. Building upon these are the higher-order processes, including mental model building (interpreting the text within the reader's global knowledge), text model formation (constructing the discourse structure of a single text) and intertextual model representation (constructing an organized representation across multiple texts). The locus of this processing activity is a metacognitive mechanism that regulates comprehension and learning. Together, lower-level and higher level components form a persuasive explanation for how we read.

Grabe (2009) contends that it is sometimes easy to criticize component approaches to the reading process. By dissecting the reading process into discrete units, it may be argued that we lose sight of the big picture—the dynamic and complex interaction that occurs between the reader and the text. Academics currently acknowledge that the reading process is not a simple aggregation of its component parts; however, there has been no credible research to date that proves that the component parts do not operate together to generate reading comprehension (Grabe, 2009; Khalifa & Weir, 2009). Ignoring the component parts in this study (or, any meaningful discussion of the reading process) is, therefore, tantamount to losing invaluable insights into reading comprehension.

2.5.1 Word recognition

In the first volume of the influential *Handbook of Reading Research*, Gough (1984: 225) stressed that 'word recognition is the foundation of the reading process'. In subsequent volumes of the handbook, academics such as Roberts, Christo and Shefelbine (2010) and Stanovich (1991) cited this quotation from Gough and affirmed the centrality of word recognition to the total reading process. Research has shown that word recognition serves as a major predictor of later reading abilities (Grabe, 2009; Perfetti, 1999; Stanovich, 1991). That is, while an adolescent may recognize a sufficiently wide range of vocabulary yet possesses poor comprehension abilities, the converse virtually never happens. Rapid and automatic word recognition frees up cognitive load, enabling a reader to focus on other aspects of the text, such as structure and style, rather than what word the print represents (Birch, 2007). Fluent native readers of Chinese can read a text comfortably at 260-300 characters per minute (Fraser, 2007). Syllabus 2011 does not specify the reading rate that CL2

students are expected to attain; however, if we refer to the test specifications of a comparable examination, the *Hanyu Shuiping Kaoshi* (HSK 汉语水平考试),² an advanced CL2 learner at Levels 5 and 6, which are approximately equivalent to C1 and C2 Levels for the Common European Framework of Reference for Languages (CEFR),³ should be able to read at 200-250 characters per minute (The Office of Chinese Language Council International, 2016).⁴

For fluent word recognition to occur, a reader must match the graphic form of words in a text very rapidly by drawing on memorized orthographic, phonological and semantic constituents of the words (Perfetti, 2007). In the case of the less experienced Singapore CL2 reader, the corresponding process is often complicated by a limited sight vocabulary in the Chinese language and by the fact that word recognition in Chinese, which is a logographic writing system, is distinctively different from that in their first language, English, which is an alphabetic writing system. In Chinese orthography, more than 80% of modern Chinese characters (字) are compound characters (Shu & Anderson, 1999), that is, they are made up of recurring structural patterns known as radicals (偏旁部首). Radicals are defined as the ‘smallest, meaningful orthographic units that play semantic or phonetic roles in compound characters’ (Shen & Ke, 2007: 99). The typical Chinese character has two parts to it: a semantic radical that provides a visual cue to the meaning of the character, and a phonetic radical that serves as a clue to its pronunciation. For

² The Hanyu Shuiping Kaoshi (HSK 汉语水平考试), translated as the Chinese Proficiency Test, is a standardized test administered by the Office of Chinese Language Council International, an agency of the Ministry of Education, China. The HSK, which measures the Chinese proficiency level of non-native speakers such as heritage learners and Chinese as a foreign language learners, comprises six levels, namely Levels 1 and 2 (Beginner); Levels 3 and 4 (Intermediate) and Levels 5 and 6 (Advanced). Details can be obtained from the HSK official website (The Office of Chinese Language Council International, 2014).

³ The CEFR, developed by the Council of Europe, provides a detailed description of the achievements of learners of foreign languages across Europe and increasingly worldwide (Verhelst, Van Avermaet, Takala, Figueras & North, 2009). Divided into three broad bands and six main levels, the CEFR forms a common basis for the elaboration of language assessment, curriculum, teaching and learning. The three broad bands are A (Basic user), B (Independent user) and C (Proficient user) which can each be divided into two levels, namely A1 (Breakthrough) and A2 (Waystage); B1 (Threshold) and B2 (Vantage); C1 (Effective operational proficiency) and C2 (Mastery). More information is available at the CEFR official website (Council of Europe, 2018).

⁴ On the issue of external criteria, specifically HSK and CEFR, please see Chapter 9 for a brief discussion.

example, the character 韵, /yùn/, is composed of the radicals 音 and 匀. There are more than 200 semantic radicals and 1000 phonetic radicals (Shu & Anderson, 1999) which can tax a CL2 reader. Moreover, in some characters, the radicals may not necessarily reflect the meaning and pronunciation of the character, partly because Chinese writing, with origins dating back to the *Shang* dynasty (1600-1046 BC), and Chinese phonology have evolved substantially over several thousands of years.

Chinese characters, with rare exceptions, are monosyllabic. A morpheme (语素), the smallest combination of meaning and phonetic sound in the Chinese language, can be made up of one or more characters. To complicate matters, some morphemes can be used independently as words (词) while others can only form words in combination with other morphemes. The prior instance is known as free morpheme (自由/成词语素, examples include 我、好、水、蝴蝶) while the latter includes the half-bound morpheme (半自由语素, the position of this type of morpheme in a word is not fixed; examples include 民、丽、观、伟) and bound morpheme (不自由语素, this type of morpheme occupies a fixed position in a word; examples include 阿、第、们、子). Thus, being able to identify a morpheme or a character does not guarantee recognition of a word. For struggling CL2 readers, the absence of word demarcation—in written English, words are clearly divided by whitespaces—could be yet another hurdle. Less fluent readers often have to make use of contextual information as support for word recognition, though this further slows down their reading speed.

As word forms are being processed visually, the potential matches in a reader's lexicon are being activated. This act of retrieval is termed lexical access. The ease of accessing a lexical entry depends primarily on the context, that is, what other entries the reader has just accessed, and on the entry's resting level of activation. Words that are frequently accessed by the reader have higher resting levels. Unsuitable words are suppressed and a selection is made. In the light of this, test designers have to ensure that the proportion of frequent words and words with less frequent coverage in a reading examination is appropriate to the test taker's level of language proficiency.

2.5.2 Syntactic parsing

In linguistics, syntax is the set of rules, principles and processes that govern the structure of sentences in a given language. The term also refers to the study of such principles and processes (Chomsky, 2002). Syntax may be ‘taken as synonymous with “grammar” and therefore covers not only word order, but also word form (morphology) and structural elements (determiners, prepositions, auxiliary verbs etc.)’ (Khalifa & Weir, 2009: 49). Syntactic parsing is essentially analysing a sentence in terms of its grammatical constituents; not unlike word recognition, fluency in syntactic parsing is critical to reading comprehension. Once a reader has identified and accessed the meaning of the words, they would have to group words into larger units at the phrase and sentence levels to construe the meaning of the text.

The more complex and ambiguous syntactic structures naturally increase the reading processing time and difficulty in comprehension. Consider the V + N1 + *de* (Auxiliary) + N2 sentence, a classic syntactically ambiguous construction in Chinese:

V	+	N1	+	AUX	+	N2
咬		猎人		的		狗
<i>yao</i>		<i>lieren</i>		<i>de</i>		<i>gou</i>
bite		hunter		AUX		dog

The sentence can be understood as either *To bite the hunter's dog* (with *dog* and *hunter* being the object and modifier respectively) or *The dog that bites the hunter* (with *hunter* and *dog* being the object and subject respectively). Unlike English, Chinese is an uninflected language. On the one hand, the lack of inflection frees learners from the need of handling a more complex verb system. On the other hand, an absence of explicit markers or changes in the form of words to express different grammatical categories, such as tense, gender, aspect and voice, may result in ambiguities as in the dog and hunter example. Syntactic ambiguity may also result when there is more than one possible way of word segmentation:

A) 已/结婚/的/和/尚未/结婚/的/青年/都/要/参加
 yi/ jiehun/ de/ **he/ shangwei/** jiehun/ de/ qingnian/ dou/ yao/ canjia
 already/married/ AUX/ **and/ yet/** married/ youths/ all/must/ attend

B) 已/结婚/的/和尚/未/结婚/的/青年/都/要/参加
 yi/ jiehun/ de/ **he shang/ wei/** jiehun/ de/ qingnian/ dou/ yao/ canjia
 already/married/ AUX/**monks /yet/** married/ youths/ all/must/ attend

As mentioned, there is no spacing between either adjacent characters or words in Chinese. A and B are the exact same sentence. In A, we interpret the three characters, *he* (和), *shang* (尚) and *wei* (未) as the words *he* (和, and) and *shangwei* (尚未, yet); the sentence thus means *All married and unmarried youths must attend*. However, when we segment the three characters into the words *heshang* (和尚, monks) and *wei* (未, yet), as in B, a different meaning is produced (*All married monks and unmarried youths must attend*). It should be evident that competence in the syntax of a language is crucial for deriving meaning from the text.

2.5.3 Semantic-proposition encoding

As words are being processed during reading and structural groups of words are parsed, information is extracted at the same time from the words and structures to form units of meaning also known as semantic propositions (Kintsch, 1998, 1974.). Words will be encoded differently depending on the semantic propositions and their linkages which the reader has constructed to date. Words, then, are encoded in their contextually appropriate sense. This is referred to as semantic-proposition encoding. For example, in the sentences below, the same word *da* (打) is encoded in three ways by a proficient reader:

A) 他 在 屋子 里 打 人
 ta zai wuzi li **da** ren
 he PREP house PREP **beat** someone
 He is **beating someone** inside the house.

B) 他 在 屋子 里 打 电话
ta zai wuzi li da dianhua
he PREP house PREP **call** telephone
He is **making a phone call** inside the house.

C) 他 在 屋子 里 打 毛线
ta zai wuzi li da maoxian
he PREP house PREP **knit** wool
He is **knitting** inside the house.

Clearly, a reader's mental dictionary must contain at least three entries under *da* (打).⁵ As semantic-proposition encoding occurs, units of meaning are activated and built. The linkages between the units reflect the relational aspect of concept knowledge, for example, that *shen* (深, deep) is the opposite of *qian* (浅, shallow) and that the words *gou* (狗, dog) and *mao* (猫, cat) can be grouped under the category of animals. As Perfetti and Curtis (1986: 26) point out, 'it is this knowledge which constrains how a reader can interpret a text'.

2.5.4 Working memory

Atkinson and Shiffrin (1968) proposed the multi-store model of working memory by likening memory to an information processing model comprising a series of stores, namely sensory memory, short-term memory and long-term memory. The model was well-received at that time and initiated a new line of research in this area. Building on the multi-store model, Baddeley and Hitch (1974) introduced the concept of working memory to address the limitations of the short-term memory component. Like short-term memory, working memory is a temporary store; however, it is not unitary and consists of various sub-systems for different types of inputs received (see Figure 2b).

⁵ There are, in fact, 25 entries under the verb *da* (打) in the *Xiandai Hanyu Cidian* (《现代汉语词典》).

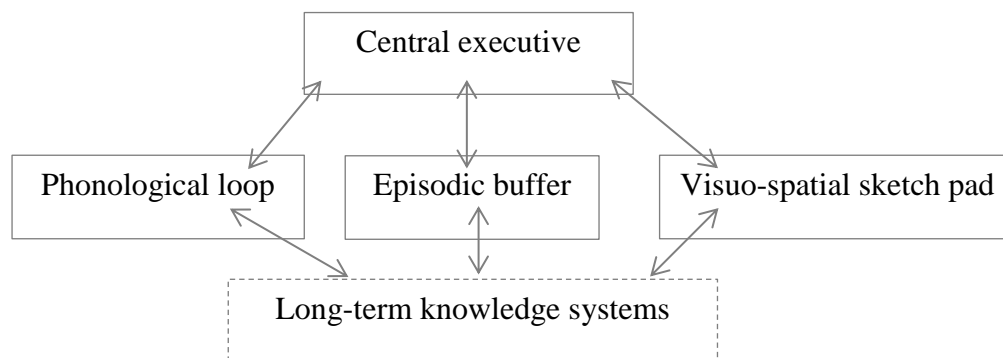


Figure 2b: Baddeley and Hitch’s model of working memory (Baddeley, 2000: 21)

The role of working memory in lower-level processing for reading is relatively direct and well established. During the process of reading, the overarching central executive is responsible for channelling the reader’s attention to relevant information while preventing distractions through suppression. The central executive also monitors and co-ordinates the phonological loop and episodic buffer—sub-systems which are activated when reading. The phonological loop deals with written material, converting printed words into an articulatory (spoken) code before storing them; and the episodic buffer functions as a link between working and long-term memory. The central executive together with its sub-systems support word recognition, store and combine words that have been activated and aid syntactic and semantic processing to build a coherent representation of the text.

2.5.5 Building a mental model

The process of reading must be described not only at the local micro-level but also at the global macro-level. Accumulated research over the past few decades (e.g. Zwaan & Rapp, 2006; Perfetti, 1999) indicate that higher-order comprehension processing mainly involves the following component abilities: a mental model, text model, intertextual model and a metacognitive mechanism.

The mental model or situation model is the reader’s interpretation of a text. The construction of meaning is not based entirely on the text; rather, ‘readers not only process a text at a propositional level, but also construct a mental model that is analogous in structure to the events, situations or layouts described by the text’

(McNamara, Miller & Bransford, 1991: 493). When we read, information obtained from the text is constantly being integrated into a mental representation of the text so far. Often, this is preceded by inferencing, whereby readers have to go beyond the literal meaning of what is written to predict or deduce ideas that are implicit.

The mental representation is non-static and is continuously updated with new information from the text. When there is a shift or break from the reader's expectations, effortful adjustments have to be made to the representation, usually resulting in a longer reading time. This process ensures that incoming information 'contributes to the developing representation of the text in a way that is consistent, meaningful and relevant [... It] entails an ability to identify main ideas, to relate them to previous ideas, distinguish between major and minor propositions and to impose a hierarchical structure on the information in the text' (Field, 2004: 241).

As the reader continues to read, a macro mental model is built up. It has been demonstrated that a reader's mental model is personal and hugely dependent on their purposes, approaches, attitudes and interests. The reader's background knowledge and experience, stored in their long-term memory, also exert a powerful influence, selecting and reinforcing what is remembered from the text and determining its relative importance. Test designers should therefore be mindful that test-takers have varied mental models and may not interpret the text in the same way as intended. Comprehension passages and items that may penalize test-takers who lack certain background knowledge and experience should also be avoided to ensure fairness and equity.

2.5.6 Forming text and intertextual models

If the mental model accounts for how the reader interprets a text, then the text model calls for understanding what the text itself is trying to signal. The RAND Reading Study Group posits that while reading is defined by three components: the reader, the text and the activity (which includes the purposes, approaches and consequences associated with the act of reading), it is a phenomenon that occurs within a larger socio-cultural context (Snow, 2002). Research confirms that these three components, individually and collectively, affect a reader's comprehension of what is being read

(e.g. Duke, 2005; Alexander & Fox, 2004). Reading comprehension, therefore, entails not only the *reader*, who is doing the comprehending, but also the *text* which is to be comprehended. In addition to surface codes (the literal wording of texts), texts often contain a hierarchy of propositions which gives rise to a text base. Development of a reasonably accurate and complete text model of comprehension would seem to involve understanding at both levels. To put it differently, a reader needs to tease out, in a network of propositions, the ideas and concepts that are central to the text and those that are of secondary importance. The reader would also have to decipher the relationships between these units of information, and the author's designs and intentions. When these text base elements are interwoven with the reader's mental model, a discourse-level representation is created for the text as a whole.

Reading comprehension at the higher processing level, hence, does not occur simply by extracting an author's meaning from the text. During reading, the reader is also actively interpreting information in the light of their knowledge and experience. Variables such as different purposes and approaches for reading, and different types of texts being read will determine whether the mental model or text model plays the more active role. For example, when a reader scans a poster to obtain information about an event, their understanding should match the information intended by the designer of the poster. An event poster is not meant to be regarded in ten dissimilar ways by ten different readers. In contradistinction, an avant-garde literary text studied during a Chinese literature class privileges the mental model of reader interpretation. Avant-garde literature with its narrative gaps, experimental techniques and vague expressions lends itself to numerous interpretations. Readers are also expected to analyse and critically evaluate texts in a literature classroom.

From the reading processes outlined above, we can gather that texts can be difficult or easy, depending on several factors. The vocabulary load, lexical density and syntactic complexity of texts are strong predictors of text difficulty at the lower processing level. At the higher processing level, the structure and content of texts have a critical bearing on reading comprehension. When a text's propositions are largely unaligned with the reader's knowledge and experience, the text may prove too challenging for optimal comprehension to occur. Furthermore, compelling

evidence from relatively recent research (e.g. Britt & Sommer, 2004; Stromso & Braten, 2002) suggests that collating and synthesizing information from more than one text places added demands on the reader and may require an intertextual model. As Unaldi (2010: 37) aptly points out, ‘since texts are not normally written to be read in conjunction with other texts, they lack explicit links to facilitate integration of information across texts. The demands on the reader to form a macrostructure are higher than when reading a single text with intra-textual coherence’. In today’s information age, students are increasingly required in class to process multiple, complex texts to form a cogent intertextual representation. The sheer volume of electronic texts, often with hyperlinks and multimedia, also means that students have to compare and connect information across texts in their daily routines. Ensuring the quality and relevance of the GCE 1162 reading examination would surely require more knowledge about the intertextual model.

2.5.7 Metacognitive mechanism

We have looked at the various processes happening at the object level, the level at which ‘one’s thinking’ occurs (Nelson & Narens, 1990). At the meta level, higher-order cognition about cognition or ‘thinking about one’s thinking’ takes place. As we read, our metacognitive mechanism directs our attention to planning, monitoring, evaluating and repairing breakdowns in comprehension. Two aspects are often considered to be involved: metacognitive knowledge and metacognitive regulation. Metacognitive knowledge includes the reader’s knowledge of his reading ability, their strengths and weaknesses as a reader, the requirements of various reading tasks and the strategies to use to facilitate better understanding of texts. Metacognitive regulation refers instead to how the reader regulates their reading. For example, if the reader is satisfied with their comprehension level, they will continue reading; if not, they may decide to check a dictionary or draw a mind map to address the problem. These actions reroute the reader’s cognitive processes or related behaviours based on the feedback received from their metacognitive mechanism. A long-standing tenet in reading research is that skilled readers are conscious of their metacognitive mechanism. They have better control of their reading, respond to reading purposes appropriately, and apply and revise strategies when necessary. Less successful and struggling readers, unfortunately, do not show such sophistication in metacognition.

The reading processes and approaches that test-takers select when taking the GCE 1162 reading examination will be revisited in Chapter 7. In the following section, attention is turned to models of reading. Some of the more dominant reading models are worthy of consideration in the endeavour to better define reading comprehension.

2.6 Models of reading

Research advances in various aspects of reading are commonly assembled to form models of reading. Models typically make further predictions about reading beyond accounting for existing research findings. A brief review of reading models could begin with the popular bottom-up, top-down and interactive metaphorical models.

A bottom-up reading model focuses on a single direction, the part to whole processing of a text. The reader is assumed to be involved in a somewhat mechanical process where they decode the text which has been previously encoded by the author, starting from ‘the smallest linguistic unit, gradually compiling the smaller units to decipher and comprehend higher units (e.g. sentence syntax)’ (Dechant, 1991: 23). The reader’s knowledge and higher-order processing strategies are of little significance. Bottom-up models owe much to the work of LaBerge and Samuels (1974), Gough (1972) and Flesch (1955). A top-down model, in contrast, stresses the centrality of the reader. Readers activate what they consider to be relevant general and domain specific knowledge (or schema) and map incoming information from the text onto it. Comprehension is expectation based and the reader’s knowledge is actively used to predict text meaning. In this view, reading is a matter of bringing meaning to texts and not extracting meaning from them (McCormick, 1988). Proponents of top-down models include Smith (2004) and Goodman (1985, 1969).

It is now generally accepted by academics that a pure bottom-up or top-down reading model is inadequate as readers are not passive decoders of meaning, as assumed by bottom-up models, nor is reading carried out in a serial order as it is envisioned by both models. Top-down models are also unable to provide sufficient explanation as to how prior knowledge is activated and subsequently used in comprehending texts. More adequate models, known as interactive models, were proposed as a result (Kintsch, 2004; McCormick, 1988; Stanovich, 1980, Rumelhart, 1977). These

interactive models combine useful elements from both bottom-up and top-down models, portraying reading as a bi-directional cognitive process in which every component can interact with any other component, be it 'higher up' or 'lower down'. Even in contexts where one processing direction is preferred (for example, a young novice L1 reader is likely to favour the bottom-up direction), the other is also operative typically. Thus, although it may be useful for test designers to ask themselves when evaluating reading comprehension items whether bottom-up or top-down reading give test-takers a better chance of getting this item right, it is highly unlikely that any test item would involve only one or the other reading direction (Alderson, 2000).

Bottom-up, top-down and interactive models, as mentioned earlier, are metaphorical rather than truly scientific models. Although they represent the most common way to discuss reading, they are informal generalizations that stem from reading comprehension research from the 1970s to the present (Hedgcock & Ferris, 2009). These three types of model provide a useful foundation for studying the many scientific models of reading available today which are more grounded in empirical evidence, descriptive and psychologically plausible. A full volume could readily be written about each of these reading models; however, for the rest of this section we are only going to cover subskill models of reading which are of direct relevance to the design and analysis of the GCE 1162 reading examination.

Subskill models of reading represent the view that the reading activity is composed of a number of distinguishable and hierarchical subskills. A reading subskill can be defined as a cognitive ability which a person is able to use when interacting with written texts (Hudson, 1998). These models evolved in the 1970s in large part to meet the practical needs of testers and teachers (Alderson, 2000). Users of test data often require specific and reliable information about a test-taker's reading ability. In an attempt to satisfy that need, testers have to minimize the occurrence of construct irrelevance variance. A test specification designed with an awareness of the construct of reading per se is therefore extremely useful. Likewise, teachers argue that students would benefit from more focused and structured practice in relation to reading skills or strategies rather than general language lessons. There was also a growing demand

towards the end of the last century for communicatively oriented pedagogical syllabuses and curricula with ‘bite size’ teaching and learning chunks.

The different subskill models of reading aim to break down reading into constituent subskills which the skilled reader is argued to have. Davis (1968) posits that there are eight interactive subskills: recalling word meanings, drawing inferences about the meaning of a word in context, finding answers to questions when information is stated explicitly, weaving together ideas in the content, drawing inferences from the content, recognizing a writer’s purpose, attitude, tone and mood, identifying a writer’s technique and following the structure of a passage. Heaton (1988) defines fourteen subskills of reading and the New York City Board of Education identifies thirty-six (Lunzer & Gardner, 1979). Kintsch and Yarbrough (1982) describe two levels of subskills: micro-processes (local, phrase-by-phrase understanding) and macro-processes (global understanding). Hughes (1989) expands that to four levels by adding grammatical and lexical abilities and low-level operations. Williams and Moran (1989: 224) conclude that while these researchers ‘may disagree on the emphasis to be devoted to any particular skill, there seems to be substantial agreement on the importance of such skills as guessing the meaning of unknown words, identifying anaphoric reference, identifying the main idea and inference’.

It is necessary to note that in L2 education, Munby’s taxonomy of subskills is often used in the designing of syllabuses and materials as well as language tests (Alderson, 2000). Munby (1978) distinguishes the following reading subskills (see Figure 2c):

1.	Recognizing the script of a language
2.	Deducing the meaning and use of unfamiliar lexical terms
3.	Understanding explicitly stated information
4.	Understanding information when not explicitly stated
5.	Understanding conceptual meaning
6.	Understanding the communicative value of sentences
7.	Understanding relations within the sentence
8.	Understanding relations between parts of text through lexical cohesion devices
9.	Understanding cohesion between parts of a text through grammatical cohesion devices
10.	Interpreting text by going outside it
11.	Recognizing indicators in discourse
12.	Identifying the main point or important information in discourse
13.	Distinguishing the main idea from supporting details
14.	Extracting salient details to summarize (the text, an idea)
15.	Extracting relevant points from a text selectively
16.	Using basic reference skills
17.	Skimming
18.	Scanning to locate specifically required information
19.	Transcoding information to diagrammatic display

Figure 2c: Munby’s taxonomy of reading subskills (Munby, 1978)

As Pearson (2009) maintains, Munby’s taxonomy together with other more widely recognized subskill reading models remain influential today, particularly in the teaching and assessing of L2 reading. In more pedagogically oriented discussions, several academics and educators in China (Wei, 2012; Zeng & Wan, 2012; Zhou, 2003; Xia, 2001; Yang & Yang, 2001) have also reiterated the significance of reading subskills such as scanning, surveying for general meaning, activating background knowledge and recognizing story structure in the CL2 classroom.

The subskill approach to reading curriculum design and teaching is not new in Singapore. It has been adopted since the 1980s and the *English Language Syllabus*

2010 (CPDD, 2010) includes detailed documentation of reading subskills that students are expected to attain at each key stage. Figure 2d is an excerpt from the syllabus. Examining the list of reading subskills in the *English Language Syllabus 2010* could offer a fresh perspective on Chinese curriculum planning and design. Syllabus 2011, in comparison, only lists six cognitive aspects that the curriculum aims to develop without further elaboration. These subskills are: remember, classify, infer, create, evaluate and analyse. The connection between this list and the revised Bloom's taxonomy (Anderson & Krathwohl, 2001) is clear although it remains uncertain why the subskill of applying has been omitted and the sequence of subskills changed.

Reading and viewing			Secondary				
Focus areas	Learning outcomes	Skills, strategies, attitudes and behaviour	1N	1E/ 2N	2E/ 3N	3E/ 4N	4E/ 5N
<p>Reading and viewing of different types of rich texts</p> <p>...and text type-specific comprehension skills and strategies,...</p> <p>(continued)</p>	<p>LO4:</p> <p>Apply close and critical reading and viewing to a variety of literary selections and informational/functional texts, from print and non-print sources, for learning in the literary/content areas and to understand how lexical and grammatical items are used in context</p> <p>(continued)</p>	Reading and viewing informational/functional texts					
		Layout					
		<ul style="list-style-type: none"> Identify typographical and visual features (e.g. captions, font types/sizes, text layout, illustrations) 					
		<ul style="list-style-type: none"> Identify text features (e.g. titles/headlines, main and sub-headings, captions/labels for visuals) 					
		<ul style="list-style-type: none"> Recognize the organizational patterns in a text (e.g. comparison-contrast, problem-solutions) 					
		Text Response					
		<ul style="list-style-type: none"> Make predictions about the content of a text using, e.g. <ul style="list-style-type: none"> prior knowledge typographical and visual features text features organizational patterns organizational structure (e.g. in an exposition, thesis statement – justification – restatement of thesis) 					
		<ul style="list-style-type: none"> Explain whether predications about the content of a text are acceptable or should be modified and why 					
		<ul style="list-style-type: none"> Restate the gist/main idea and key details 					
		<ul style="list-style-type: none"> Examine the arguments for or against an issue, including the quality of the arguments 					
		<ul style="list-style-type: none"> Identify and interpret the evidence in arguments, e.g. <ul style="list-style-type: none"> facts reasons appeal to an authority 					

Figure 2d: Excerpt from the *English Language Syllabus 2010*, Singapore (CPDD, 2014: 45)

The revised Bloom's taxonomy represents a continuum of increasing cognitive complexity: remember, understand, apply, analyse, evaluate and create. Nineteen specific cognitive processes such as classifying and inferring are further subsumed under these six categories (see Figure 2e). Neither the revised Bloom's taxonomy (Anderson & Krathwohl, 2001) nor the original Bloom's taxonomy (Bloom, Englehart, Furst, Hill & Krathwohl, 1956) is meant specifically to explain reading comprehension—it was developed as a taxonomy of educational objectives that can be applied to a wide range of disciplines. Nonetheless, researchers and educators have subsequently used Bloom's taxonomic frame to unpack various infrastructures for reading comprehension (e.g. Irwin, 1986; Herber, 1978; Pearson & Johnson, 1978; Clymer, 1968). Zhu (2015), an academic based in Hong Kong, has expanded the taxonomy after examining the Chinese language curriculum and assessment in Singapore, Hong Kong and mainland China. Tailored specifically to assist educators in analysing and developing reading comprehension items, Zhu's revised Bloom's taxonomy comprises six levels of comprehension. The levels are sequenced in increasing cognitive demand: recall, explain, organize, deduce, critique and create. Description of each level is provided in *Assessment for Learning: Reading* (Zhu, 2015) (see Figure 2e for summary), and advice for item writers is also given. For example, to set a recall level item that is intended to be less challenging, item writers may require test-takers to locate information explicitly stated in the text. If an item of higher difficulty is required, item writers can, for instance, design an evaluation level item which assesses the test-takers' ability to judge the strengths and weaknesses of a character.

The cognitive levels in Zhu's revised Bloom's taxonomy and Syllabus 2011 are mapped onto the revised version of Bloom's taxonomy and displayed in tabular format below to capture the similarities, differences and relationships between them (see Figure 2e). Example items from GCE 1162 reading examination papers are provided, where possible, to illustrate the demands of each cognitive level.⁶ For ease of reference, expert judges shall continue to use the six levels listed in the 2001 revised Bloom's taxonomy when evaluating sample items for this research, while

⁶ When reviewing the 22 sets of GCE O-Level Chinese Language Examination papers from the past decade, expert judges were unable to identify any items from the create level of cognitive complexity.

taking into account Zhu's useful descriptors (see Figure 2e).⁷ Clear specification of terms and appropriate methodology are essential for expert judges to reach a closer agreement on what subskills are being tested. Although it is not possible to link with absolute accuracy every test item to a specific subskill, data obtained through such an analysis offer a step towards preventing construct underrepresentation. Test developers should attempt to design the overall spread of items in an examination in such a way as to cover all subskills that are commensurate with the target level of difficulty (Khalifa & Weir, 2009).

⁷ I have retained the apply category from Anderson and Krathwohl's (2001) revised Bloom's taxonomy, even though it is neither listed in Syllabus 2011 nor Zhu's (2015) revised Bloom's taxonomy. The main reason for this retention being the greater emphasis in recent years on authenticity and relevance across all subjects in Singapore (MOE, 2010a). Reading comprehension items from the apply category tend to be more authentic and relevant as they often require students to connect real life experiences with the text.

Secondary Chinese Language Syllabus 2011	2001 revised Bloom's taxonomy	Specific cognitive processes	Zhu's revised Bloom's taxonomy for reading comprehension item analysis	Descriptors	Example item	
记忆 Remember	Lower-order thinking skills	Remember	<ul style="list-style-type: none"> ▪ Recognizing ▪ Recalling 	复述 Recall	<u>Literal comprehension</u> Recognize or locate information and ideas explicitly stated in the text.	(根据原文第二段) 网上注册号码有什么作用? (According to paragraph 2), what purpose does an online identification number serve? (Q15, May 2016)
比较分类 Classify		Understand	<ul style="list-style-type: none"> ▪ Interpreting ▪ Exemplifying ▪ Classifying ▪ Summarizing ▪ Inferring ▪ Comparing ▪ Explaining 	解释 Explain	<u>Reorganization/explanation/basic inferential comprehension</u> Paraphrase or summarize information explicitly stated in the text;	试解释(这句话)在文中的意思: 还没起步, 便先被心理的阴影绊倒了。 Explain the meaning of the following sentence: Letting the fear of what could happen makes nothing happen. (Q27a, November 2013)
推测 Infer				重整 Organize	explain the meaning of words and sentences; basic inference of, for example, main ideas and causal effects.	
		Apply	<ul style="list-style-type: none"> ▪ Executing ▪ Implementing 		<u>Application</u> Relate personal experiences to the text.	“年轻一辈缺乏的就是多走几步路的勇气和精神”。你同意这种看法吗? 为什么? 试举生活中的例子加以说明。 ‘The millennials often lack the courage and resilience to persevere until the end’. Do

						you agree or disagree with the above statement? Why? <i>Use specific examples from your personal experience to support your opinion.</i> (Q30, November 2015, emphases added)
创造 Create	Higher-order thinking skills	Analyse	<ul style="list-style-type: none"> ▪ Differentiating ▪ Organizing ▪ Attributing 	伸展 Deduce	<u>Deduction</u> Conjecture and form hypotheses; deduce e.g. character traits, author's intentions and literal meanings from author's figurative uses of language.	作者写第二段的用意是什么? What is the author's intention of writing paragraph 2? (Q22, May 2013)
思考评价 Evaluate		Evaluate	<ul style="list-style-type: none"> ▪ Checking ▪ Critiquing 	评鉴 Critique	<u>Evaluation</u> Form judgements of, for example, central ideas, characters; articulate emotional and aesthetic responses to the text.	“年轻一辈缺乏的就是多走几步路的勇气和精神”。你同意这种看法吗？为什么？试举生活中的例子加以说明。 ‘The millennials often lack the courage and resilience to persevere until the end’. <i>Do you agree or disagree with the above statement? Why? Use specific examples from your personal experience to support your opinion.</i> (Q30, November 2015, emphases added)
分析排列 Analyse		Create	<ul style="list-style-type: none"> ▪ Generating ▪ Planning ▪ Producing 	创意 Create	<u>Creation</u> Propose solutions and alternatives; modify plot and ending.	如果你是老师，你会怎么说以取得更好的效果？ If you were the teacher, how would you provide criticism more effectively?

Figure 2e: Comparison of cognitive levels in reading between Syllabus 2011, Anderson and Krathwohl's (2001) revised Bloom's taxonomy and Zhu's revised Bloom's taxonomy (2015)

2.7 Reading in a second language

To this point in the chapter, the purposes, approaches, processes and models of reading have been outlined. Reading is a complex multifaceted construct. Therefore, it cannot be fully understood unless it is dissected into its major operations and components, and each studied in turn. The GCE 1162 reading examination is a standardized national examination to measure the reading proficiency of CL2 students in Singapore. Assessing reading in a second language inevitably brings us to the question of the nature of reading in L2. The acquisition of literacy is a lengthy, deliberate and effortful process, particularly in L2. L2 reading, unlike L1 reading, involves two languages; put another way, L2 reading is an ability that combines L2 and L1 reading resources into a dual-language processing system. The dual-language involvement ‘implies continual interactions between the two languages as well as incessant adjustments in accommodating the disparate demands each language imposes. For this reason, L2 reading is cross-linguistic and, thus, inherently more complex than L1 reading’ (Koda, 2007: 1).

We will now identify the differences between L1 and L2 reading, before drawing attention to the relationship between the two. Differences between L1 and L2 reading are more pronounced in weak readers who wrestle with word recognition; however, there are noticeable differences even with advanced L2 readers who have been learning the language for many years. First, as we will see in Chapter 3 on the Singapore context, CL2 reading in Singapore encompasses a wide range of learners, of varying family language backgrounds, and with disparate language proficiency levels. Their CL2 literacy experiences and the linguistic resources available to them also differ considerably.

Second, students often have different motivations for reading in CL2 compared to reading in English, their first language. Although most Chinese students are *required* to read in CL2, there are substantially fewer students who *want* to or *need* to read in Chinese either because of genuine interest or to fulfil their academic goals and future aspirations. With English being the medium of instruction for most subjects in Singapore, the kinds of Chinese texts students are exposed to in schools will necessarily be very limited. Seldom is there an expectation that CL2 reading texts are

resources for acquiring new and challenging information or building academic skills and expertise. Reading in CL2, for many students, becomes a classroom practice solely for developing language skills and inheriting Chinese culture.

Third, one would be hard pressed to argue that most Singapore CL2 readers can achieve the same reading speed, fluency and automaticity as their CL1 peers in mainland China or Taiwan, even after ten to eleven years of mandatory Chinese classes. Many Singapore CL2 readers started without the much-needed oral language foundation. L2 words are likely to be connected to L1 words first, instead of being linked directly to concepts, though this will change as readers become more competent. It often takes several years for L2 readers to build sufficient active vocabulary and even longer before they develop strong implicit knowledge of the syntax, nuances, appropriate register and level of formality for specific types of texts in their L2.

Fourth, L2 reading differs markedly from L1 reading simply because it involves the interaction between two languages in virtually all its operations. Such cross-language relationship is referred to as language transfer, a key theoretical concept in L2 reading research. Traditionally, language transfer is regarded as the state resulting from a reader's falling back on their L1 linguistic knowledge and rules when there is an insufficient grasp of L2 (Odlin, 1989; Krashen, 1983). The transfer is understood as either having a facilitating effect (positive transfer) or inhibiting effect (negative transfer) on L2 reading. Studies in more recent years (e.g. Koda, 2007; August & Shanahan, 2006; Riches & Genesee, 2006) have, however, adopted broader definitions of transfer. The L1 learning experience is now seen as having reserves of knowledge, skills and abilities that are potentially available to a reader comprehending texts in L2 and research efforts have been invested in identifying these resources.

To explain the relationship between L1 and L2 learning, a number of hypotheses have been conceptualized. An influential pioneer theory is the Developmental Interdependence Hypothesis proposed by Cummins (1979). Cummins (2000: 173) asserts that 'proficiency transfers across languages such that students who have developed literacy in their first language will tend to make stronger progress in

acquiring literacy in their second language'. Underlying this notion is the belief that there is a common proficiency that supports all language learning and that learning to read need only be accomplished once. A number of scholars have since questioned whether a linguistic threshold exists which must be crossed before L1 reading ability can be transferred to L2 reading contexts (Bernhardt, 2005; Alderson, 2000, 1984; Clarke, 1980). These contentions led to Alderson's (1984) famous question: 'Is second language reading a language problem or a reading problem?' Persuasive evidence from a number of studies confirms that poor L2 reading is not due primarily to inadequate L1 reading (Koda, 2007). Additionally, little is known about how and when reading skills, shaped in one language, get transferred and become functional in another. This is especially so when the language distance between L1 and L2 is substantial, as in the case of English and Chinese. The implication for L2 reading assessment is that unsatisfactory performance in L2 reading is likely to be due to insufficient L2 proficiency and that readers stand to benefit most when remedial action pays attention to the linguistic problem rather than to any supposed L1 reading deficit (Alderson, 2000).

2.8 Conclusion

Having explored the concepts of assessment and reading, we now turn to describe how the reading construct and its various components can be, and have been, operationalized under test conditions. The testing of reading comprehension has been a source of dissatisfaction throughout its history (Snow, 2002). Academics (e.g. Alderson, 2000; Grabe, 2000) have pointed out a disjunction between research into reading, as reviewed in earlier sections of this chapter, and the actual design of reading examinations. The complexity of the reading construct it seems is often inadequately captured in reading examinations (e.g. Keenan, Betjemann & Olson, 2008; Magliano, Millis, Ozuru & McNamara, 2007; Valencia & Pearson, 1987; Johnston, 1984). Indeed, some argue that although our understanding of reading has advanced significantly over the past few decades, this seems to have little bearing on how reading is being summatively assessed. Many test developers, as interviewees in this study speculate, prefer to maintain the status quo, opting for the straightforward pen and paper test comprising passages and comprehension items on content, main ideas and vocabulary. Traditional approaches as such are popular in Singapore

because they not only provide strong reliability and at least arguable validity but are also economical and easy to administer, score and scale. The higher the stakes an examination carries the likelier test developers are to await the outcome of the evolving state of research on reading and see what consensus eventually emerges.

Whilst it may not be prudent for test developers to adopt all new developments without due consideration, they cannot afford to ignore trends and outcomes in reading research. It is not uncommon to hear anecdotally of reading examinations that bear little or no resemblance to those encountered in good instruction or the world beyond the classroom; decontextualized reading passages and trivial questions; and items that allow little opportunity for test-takers to demonstrate higher cognitive abilities or to make personal connections with reading. Essentially, test developers need to critically review their understanding of the reading construct periodically to make certain that the results from a reading examination have high validity and can be extrapolated to real-world reading (Alderson, 2000). At this juncture, we should also note that the relationship between reading research and assessment practice is two-way, rather than one-way. Much of the data gathered from reading assessment instruments could inform our understanding of the reading construct.

While a brief description of relevant parameters for validity evidence collection has been given here, an in-depth exploration is provided in Chapters 5 to 9. A reading examination development process generally follows the procedures of determining the purposes of the examination; defining the construct; ascertaining test-takers' needs; developing test specifications, item-writer guidelines and administrative procedures; selecting suitable texts; drafting items and tasks through a process of constructing, editing and revising until they are considered ready for piloting. Piloting is then carried out with suitable samples of test-takers. Items are marked and analysed both qualitatively by expert judges and statistically using Classical Test Theory and/or Item Response Theory methods. Once the actual examination has been administered, responses are marked and scores are calibrated and reported. Feedback collected from test-takers and stakeholders are channelled to policy makers, test developers, markers and other relevant parties in order to correct or compensate for any weaknesses identified. Testing standards such as *The Standards for Educational and Psychological Testing* (American Educational Research Association,

American Psychological Association & National Council on Measurement in Education, 2014) and *The Educational Testing Service Standards for Quality and Fairness* (Educational Testing Service, 2014) strongly advocate transparency to the greatest extent practicable in the entire process; and these are the yardsticks against which we evaluate the administrative structure of the GCE 1162 reading examination.

The quality of a reading examination paper is at the very least a function of both the items designed and passages chosen. In Chapter 7, I will first examine the cognitive processes and reading approaches activated by selected items in the GCE 1162 reading examination. Next, in Chapter 8, I will address the contextual parameters that are likely to influence test performance in reading. The item type and mark scheme are both given consideration. With regard to passages, there is a need to examine the discourse mode, text purpose, propositional content and readability. It is clearly possible to have passages of appropriate difficulty but items that are cognitively unchallenging; or items measuring a suitable range of reading subskills but passages that are unappealing and limited to certain topics. All these factors will come into play when I assemble cognitive and contextual validity evidence to provide a coherent account of the GCE 1162 reading examination, but for now an account of the Singapore education and assessment landscape is necessary.

Chapter 3 The Singaporean context

3.1 Introduction

Any discussion of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) is incomplete without consideration of its local education context in relation to the wider political, economic and social background. Black (2003: 7) highlights these inextricable ties, claiming that the ‘methods and procedures used in assessment and testing in any country can only be understood in the light of that country’s historical development, in relation both to its education system and broader social factors’. The approach used in this study will, first, outline the history of Singapore’s education system and the relationship between the Singapore Examinations and Assessment Board (SEAB) and the University of Cambridge Local Examinations Syndicate (UCLES). The chapter will then focus on Chinese language education in Singapore, defining key terms pertinent to this study, such as ‘mother tongue’ and ‘bilingual policy’, to bring out the main issues that impact assessment.

3.2 Singapore’s education and examination system

A sovereign state since 1965, Singapore is one of Asia's great success stories. In a short span of 50 years, Singapore has evolved into a first-world nation that fares extremely well in several categories of global competitiveness and effectiveness. With a land area of slightly over 700 square kilometres and limited natural resources, Singapore attributes much of its economic success to its high quality education system. As a nation steeped in Asian values, especially Confucian ideology, educational achievement has always been held in high regard. Singapore’s first and longest-serving Prime Minister, Mr Lee Kuan Yew, put it succinctly when he said that ‘one of the great strengths in our society is the strong support for education. It springs from the conviction of our people that our children’s future depends on education’ (Lee, 1978). As a result, Singapore’s students regularly rank among the top scorers in international assessments, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study. In a recent report by the Organization for Economic Cooperation and

Development (OECD), Singapore emerged as one of the strong performers and successful reformers in education (OECD, 2011).

Meritocracy is heralded as a fundamental ideology in Singapore and a founding policy in the education system (Mauzy & Milne, 2002; Lee, 2000). Although the word ‘meritocracy’ was coined and used in a pejorative sense in British politician and sociologist Michael Young’s (1958) satirical novel *The Rise of the Meritocracy*, it has since evolved into ‘a positive ideal against which we measure the justice of our institutions’ (Allen, 2011: 367). The very essence of meritocracy today lies in allowing everyone to progress in various fields based on their ability and effort rather than on class privilege and wealth. As early as 1959, when Singapore attained full internal self-government, the government sought to ensure that the education system would benefit the lowest common denominator of the society, allowing every child to have a fair chance at success. Concerted efforts were made to eradicate illiteracy and to equip students with the necessary skills and knowledge needed for an expanding economy. Education became a matter of right, instead of a privilege enjoyed only by elite groups. The 1960s and 1970s saw a movement towards a national education and examination system, culminating in the *Report on the Ministry of Education 1978*, (Goh, 1979) often referred to as the *Goh Keng Swee Report*.¹ To address the high attrition rates in Singapore’s education system² and to provide an opportunity for less able students to develop at a pace slower than that of more able students (Goh, 1979), a system of ability-based streaming was introduced.

The recommendations in this report have ‘far-reaching ramifications on Singapore’s education system up till today’ (Tan, Chow & Goh, 2008: 112) as students continue to be streamed according to their ability. This system espouses meritocratic advancement pathways, serving to maximize the differing capacities of students.

¹ Dr Goh Keng Swee, the late Deputy Prime Minister of Singapore (1973-1984), led the study team which completed the *Report on the Ministry of Education 1978*.

² According to the *Report on the Ministry of Education 1978*, attrition rates were 29% and 36% at the primary and secondary levels respectively, which were very much higher in comparison with education systems such as those in Taiwan, Japan, the United Kingdom and France. By 2000, the overall proportion of each primary one cohort that did not complete secondary education had fallen to 4% and has been less than 1% in the past five years (MOE, 2014b).

Singapore's Ministry of Education (MOE) is a national-level jurisdiction which controls the development and operation of most national schools. It also directs the formulation and implementation of education policies (MOE, 2015a). In the system of formal schooling following the national curriculum, students typically go through six years of primary education, followed by four to five years of secondary education. Primary education has been compulsory since 2003; while secondary education is not mandatory, the 'completion of ten to eleven years of general education is virtually universal' (MOE, 2010a: 1). Students then advance to post-secondary education of two to three years along an academic, applied-oriented or vocational pathway, before one-quarter of each cohort (approximately 13,000 students per cohort) continue to pursue a university degree. National examinations namely the Primary School Leaving Examination (PSLE), the Singapore-Cambridge General Certificate of Education Ordinary-Level Examination (GCE O-Level) and the Singapore-Cambridge General Certificate of Education Advanced-Level Examination (GCE A-Level) are conducted at the end of Grades 6, 10 (or 11) and 12, respectively.

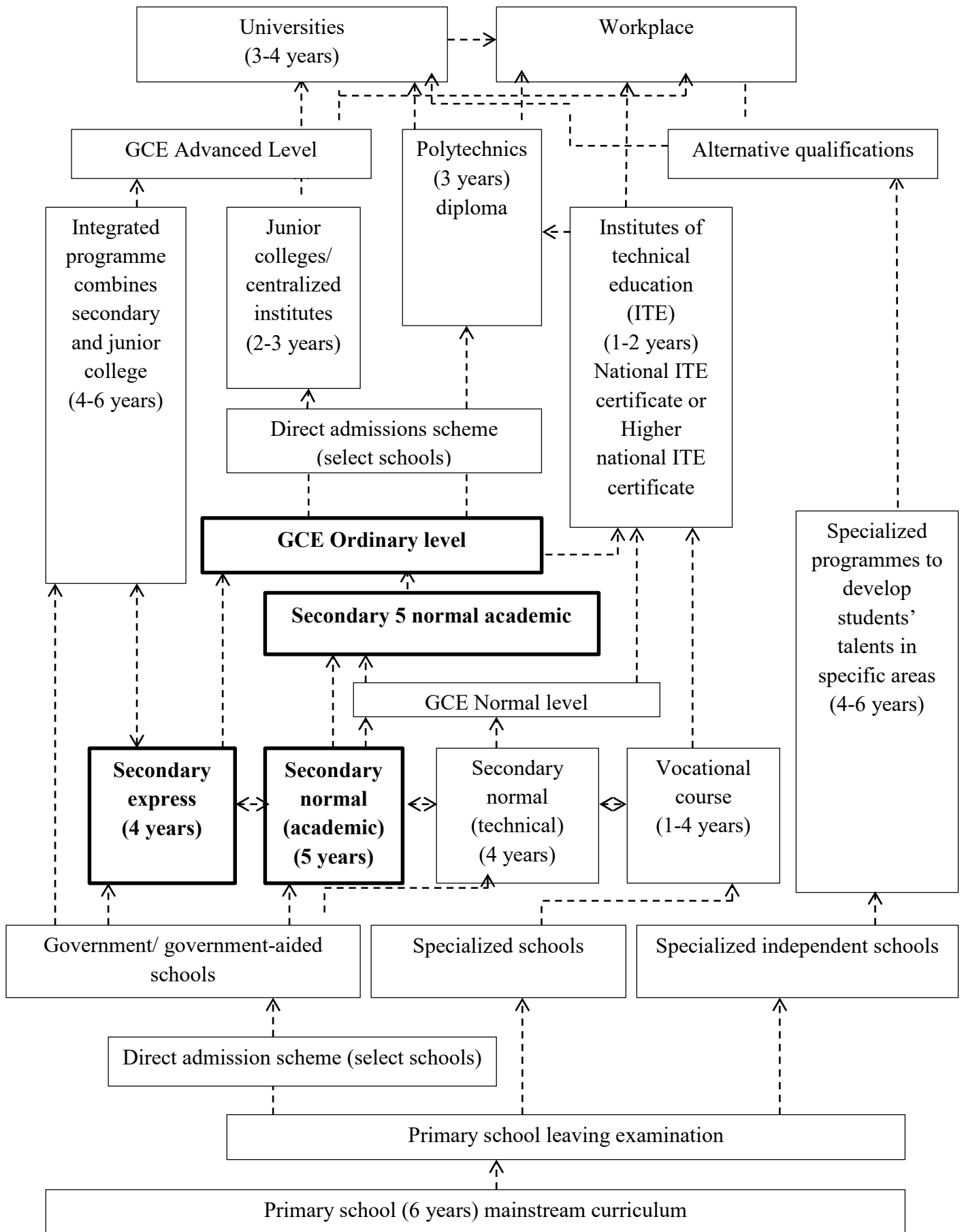


Figure 3a: Outline of Singapore's education and examination system (Tan, 2012: 47, emphases added)

At the secondary level, four main streams, designed to match students' learning abilities and interests, are offered (see Figure 3a). These four streams are the Integrated Programme, Express Course, Normal (Academic) Course and Normal (Technical) Course. The first of these, the six-year Integrated Programme, provides a seamless secondary school and junior college education in which students can proceed to junior college without taking the GCE O-Level Examination (with the exception of the higher mother tongue paper; a proportion of students also sit for the mother tongue paper). Implemented in 2004, there are currently 18 out of 154 secondary schools that offer this prestigious programme. The programme takes in high-performing students (approximately the top 10% of each cohort) most of whom study both English and mother tongue as first language. The Integrated Programme culminates in the GCE A-Level Certificate or other diplomas such as the International Baccalaureate. The second stream of the education system is the Express Course. This is a four-year course leading to the GCE O-Level Examination. In this course, mother tongue is taken as a second language. Approximately 50% of the cohort is streamed into the Express Course (MOE, 2014a). The third stream is the Normal (Academic) Course. This is a four-year course leading to the GCE N-Level Examination. Students who perform well at the N-Level will be eligible to sit the O-Level Examination in the following year (Grade 11). Selected students may also sit certain O-Level subjects at Secondary Four (Grade 10). In the Normal (Academic) Course, students learn a range of subjects similar to those in the Express Course. Approximately 25% of the total cohort of Singapore's students is streamed into this Normal (Academic) Course (MOE, 2014a). The fourth stream is the Normal (Technical) Course. Students following this course study a maximum of seven subjects that have a more technical or practical emphasis, preparing them for post-secondary education at the Institute of Technical Education. Unlike students in the first three courses, Normal (Technical) Course students are only required to sit the basic mother tongue examination. Students in this course make up approximately 15% of the cohort (MOE, 2014a).

English is taught as a first language in all streams and is also the medium of instruction for most subjects. Under Singapore's bilingual policy introduced in 1966, mother tongue is a mandatory subject. Students learn either Chinese, Malay or Tamil depending on their father's ethnicity. The streaming of students into one of the four

courses is based on performance at each national examination milestone. This system aims to achieve an accurate match between merit and qualification routes, as well as appropriate resource allocations. The national examinations are powerful gatekeepers of the system. Success in these examinations is rewarded by attractive scholarships and places at local and overseas tertiary institutions. Ultimately, students who excel academically gain better access to prestigious positions in the labour market. Singapore's differentiated education system, a result of the Goh Keng Swee reforms, is not without its critics. Singapore's survival and economic development needs have given the national education system a very pragmatic bent. Some critics feel that the obsession with high-stakes national examinations not only leads to an increased level of anxiety for students but also stifles their creativity and passion for learning. Others regard streaming as elitist and point out the stigmatizing effect of labelling students.

Responding to these concerns, MOE has in recent years expended efforts 'to soften the harshness and rigidity of the system's tracking mechanism' (Lim, 2013: 5). Alternative modes of assessment are encouraged and there has been a gradual move away from the overly strong emphasis on major summative examinations especially at the primary level. Avenues of lateral transfer from lower- to higher-prestige academic tracks have also been introduced to facilitate upward mobility. It comes as no surprise that Mr Heng Swee Keat, former Minister for Education (2011-2015), has repeatedly called for multiple pathways for success (Heng, 2015, 2011). National examinations and qualifications, though critical, 'are not the be-all and end-all' (Ong, 2016) of formal education. More importantly, Singapore's education system has to nurture students with character and integrity and equip them with 21st century competencies such as critical and inventive thinking, global awareness, cross cultural skills and a zest for life-long learning (Heng, 2012). In other words, one of the fundamental objectives of education is to 'prepare our students for life, rather than to teach for tests and examinations' (Tharman, 2005). In light of these initiatives, current issues in Chinese language assessment have to be contextualized and examined within Singapore's education system. Before we delve deeper, a brief account of UCLES and SEAB, the two main actors in Singapore's national examinations, is in order.

3.2.1 University of Cambridge Local Examinations Syndicate

UCLES, known also by its brand name Cambridge Assessment, is Europe's largest assessment agency, responsible for designing and delivering assessments to over eight million learners in more than 170 countries (Cambridge Assessment, n.d.). UCLES was established by the University of Cambridge in 1858. Its aim was to raise standards in education by administering local examinations for students who were not members of the university and inspecting schools. UCLES soon began operating examinations in territories overseas and this aspect of its work expanded quickly. Today, UCLES owns and manages the university's three examination boards, namely Cambridge International Examinations, Oxford, Cambridge and the Royal Society of Arts and Cambridge English Language Assessment. Cambridge International Examinations is the international arm of UCLES that supplies education programmes and qualifications worldwide, including the Singapore-Cambridge GCE O-Level (which is equivalent to the General Certificate of Secondary Education in England, Wales and Northern Ireland) and A-Levels.

The relationship between UCLES and Singapore is long-standing, dating back to the late nineteenth century. Modern Singapore was founded in 1819 by Sir Thomas Stamford Raffles as a British colony. A treaty was signed between Sir Thomas Stamford Raffles and the Sultan of Johor which allowed the British East India Company to establish a trading port in Singapore. Trade flourished and Singapore soon became the most strategically important colony in Britain's eastern empire. However, up to 1867, the British government had paid scant attention to education in Singapore, and 'the schools were left unchecked, without any form of government supervision' (Tan et al., 2008: 7). In 1867, Singapore, as part of the Straits Settlements, came under the direct rule of the Colonial Office in London. Reforms were undertaken to 'thoroughly re-organize all existing educational establishments [...] and to place [schools] on a more satisfactory and improved basis' (Wong & Gwee, 1980: 12). As an integral part of these reforms, the Department of Education in Singapore began conducting annual examinations in government and grant-in-aid schools from 1872. A significant milestone was reached two decades later when UCLES set up an examination centre in Singapore.

In 1891, Sir Cecil Clementi Smith, Governor and Commander-in-Chief of the Straits Settlements, wrote to UCLES to explore the possibility of setting up an examination centre in Singapore. This request was granted and the Reverend George Forrest Browne, Secretary of UCLES, replied saying that UCLES ‘shall be glad to form a centre at Singapore. It will be quite satisfactory to us that the management of the examination should be in the hands of the officers of the Education Department of the Colony’ (cited in Tan et al., 2008: 13). In the following year, the first Cambridge School Certificate Examinations were offered in Singapore. Thus began a relationship between UCLES and Singapore that continues to this day.

From 1892 to 1970, the Cambridge School Certificate Examination, taken mainly by students in English stream schools, was offered alongside other examinations, namely, the Government Secondary IV School Certificate (Chinese), School Certificate (Malay) and School Certificate (Tamil) examinations. In 1971, all four parallel examinations were replaced by the Singapore-Cambridge GCE O-Level Examination. UCLES was responsible for papers examined in English; and MOE, for papers examined in Chinese, Malay and Tamil and the subject English as a second language (MOE, 1971). The GCE O-Level examination, jointly certified by UCLES and MOE, was to become the standard national examination at the end of secondary education (see Figure 3a). This development was in tandem with the implementation of the PSLE in 1960 and the GCE A-Level Examination in 1975 as centrally coordinated and common examinations for school leavers. Together the PSLE and GCE examinations continue to form ‘the pillars of the national examination system’ in Singapore (Tan et al., 2008: 84).

Shortly after the *Report on the Ministry of Education 1978* was published in 1979, a single national stream was announced in 1983. All students would be required to study English as a first language and all schools were to use English as the main medium of instruction. This announcement signalled the end of the parallel systems of education which arose during the British colonial period. The GCE O-Level continues to be adopted as Singapore’s national examination, with Cambridge International Examinations of UCLES, MOE and SEAB being the present joint examining authorities (SEAB, 2015a). Approximately 40,000 students in Singapore sit GCE O-Level papers from more than 50 different subjects annually (SEAB,

2015b; MOE, 2014a). These papers can be categorized into four groups: Cambridge subjects (e.g. English, Physics), Cambridge O-Level school initiated elective subjects (e.g. Drama, Business Studies),³ applied subjects (e.g. Biotechnology, Media Studies)⁴ and local subjects examined in the mother tongue (e.g. Chinese, Literature in Malay). Local subjects, including the GCE 1162 Chinese language paper (or 1160 from the year 2016), are formulated and marked in Singapore using guidelines from UCLES.

UCLES has been a trusted education partner in Singapore for more than 120 years, providing Singaporean students with national examinations which are widely recognized by local and overseas institutions, universities and employers. Dr Toh Chin Chye, a prominent member of Singapore's first generation of political leaders and the then Deputy Prime Minister (1959-1968), clearly articulated the value of this partnership at the meeting of the Singapore Advisory Committee of the UCLES in 1964. He observed that 'proconsuls have come and gone and politics have taken a new colouring', but the working relationship with UCLES 'like a certain brand of Scotch whisky, [...] is still going strong' (Toh, 1964). UCLES likewise celebrated this 'enduring and fruitful partnership (of) developing successful students in the past, present and future' (cited in Tan et al., 2008: 14).

3.2.2 Singapore Examinations and Assessment Board

SEAB was established on 1 April 2004 as a statutory board under the MOE. Formerly the Examinations Division of MOE, SEAB was formed to develop and provide quality assessment services, with its core business being national examinations (SEAB, 2013; MOE, 2004b). The setting up of SEAB was in tandem with MOE's efforts to exercise greater autonomy over the examination system (MOE, 2004b). At the start of the millennium, a new form of collaboration between UCLES and Singapore began to take shape. Although this does not affect the local subjects

³ In 2005, MOE introduced greater flexibility in the curriculum by giving secondary schools the option to offer O-Level school initiated electives to build up their niche areas. These electives are taken in addition to, or as replacement for, current curriculum offerings by the MOE.

⁴ Applied subjects have been developed since 2008 by polytechnics in partnership with secondary schools to better cater to the interests and aspirations of students who are keen to progress along an applied and practice-oriented path of education. These subjects are examinable by the polytechnics.

directly, it is worth noting that since 2006, MOE together with SEAB have assumed greater control over Cambridge Subjects examined at GCE O-Level.⁵ In order to customize the curriculum and examinations to meet the evolving educational needs of Singapore, MOE and SEAB have taken greater responsibility for developing the syllabuses and their examination formats, setting standards and awarding grades (Tharman, 2004). MOE and SEAB, however, continue to tap into the expertise of UCLES in relation to designing syllabuses. The setting of question papers and the marking of examination scripts for Cambridge Subjects are also outsourced to UCLES.

Since its establishment, SEAB has strived to be ‘a trusted authority in examinations and assessment, recognized locally and internationally’ (SEAB, 2014c: 2). Its mission is to ‘assess educational performance so as to certify individuals, uphold national standards and advance quality in assessment worldwide’ (SEAB, 2014c: 2). At present, SEAB delivers more than 200 subjects to about 180,000 candidates each year and manages examination operations that handle more than 1.8 million scripts, asserting itself as a key actor in Singapore’s education structure. In 2014, at SEAB’s tenth anniversary celebration, Ms Ho Peng, Chairman of SEAB and Director-General of Education, MOE, reiterated the Agency’s commitment to upholding integrity and confidence in national examinations (SEAB, 2014c). To assure the quality of examinations invariably requires research and innovation. As the education system in Singapore undergoes transformation, the objectives, methods and content of its examinations need to be reviewed and updated to match ongoing changes in curriculum and pedagogy. Quality assurance of the national examinations is complex, cutting across many government agencies, including SEAB, MOE, the National Institute of Education (NIE), schools and the various language centres and teacher training academies. Incipient concerns over accountability and transparency in recent years further accentuate this complexity. Moving forward, SEAB could play a vital role in coordinating the efforts of these various actors, particularly in relation to the Chinese language examination at national level.

⁵ In 2002, MOE took greater control of the GCE A-Level examinations.

3.3 Chinese language assessment issues in Singapore

An exploration of issues in Chinese language assessment in the Singaporean context cannot afford to ignore the changing relationship of Chinese with the state's other official languages. Since independence, the Singaporean government has adopted a clearly interventionist stance when it comes to the management of societal multilingualism (Gopinathan, 2003). English, Chinese, Malay and Tamil were designated as official languages (All-Party Committee of the Singapore Legislative Assembly, 1956) and given equal status. In a bid to balance the sociocultural sensitivities of the different ethnic groups, English was chosen as the common language to facilitate communicative integration and to forge a new national identity. Bilingualism was set as a target for the younger generations and Chinese (also known as Mandarin, Standard Chinese or the Putonghua equivalent in mainland China) was actively promoted among the majority Chinese community to replace the various Chinese dialects, such as Hokkien, Teochew, Hakka and Cantonese. These language policies manifest themselves in education and work their influence through the agency of national examinations. By bringing into focus language policies and their implications for Chinese language assessment, this section lays the groundwork necessary for an evaluation of the GCE 1162 reading examination.

Language and language policies are imbued with values, beliefs and power relations. Language power asymmetries are rarely natural occurrences in a multi-ethnic, multi-lingual society (Tan, 2003). In Singapore, the language ecology is clearly the result of decades of 'successful centralised control of the nation's overall communicative structure' (Tan, 2003: 46). In other words, the distinct domains of English and mother tongue languages are not coincidental; rather, they have been consciously shaped and moulded by language policies enacted by the government since independence. Language policies are explicit plans, usually but not necessarily written in formal documents, about language use; and they form an integral part of the government's language management process (Spolsky, 2004). Gopinathan (2003) contends that language policies are often characterized as deliberate attempts at social change in language behaviour by a decision-making administrative structure. Further, Gopinathan claims that central to the Singaporean government's language management plans is the careful balance between racial sensitivity and the economic

and cultural value of the various languages. The Singaporean government's language policies have had considerable effect across all levels and sectors of society in the last 50 years. In the realm of Chinese language education, these policies have dictated to a very large extent how the language is taught, used and eventually, assessed.

3.3.1 History of Chinese language education in Singapore

Historically, Chinese language teaching in Singapore first started in the old-style private Chinese schools known as *sishus* (私塾). When Sir Thomas Stamford Raffles arrived in 1819, Singapore had only about a thousand inhabitants living in small fishing communities, of whom a few dozen were Chinese (Chew & Lee, 1991). Under British colonial rule, Singapore rapidly emerged as an important trading post, and with trade came a huge influx of Chinese immigrants from Southern China. As the Chinese population grew, *sishus* were set up to cater to the education needs of immigrant children. Textbooks were written in classical Chinese (*wenyanwen* 文言文) and Chinese dialects were used as the medium of instruction (Ang, 2003).

Following the Xinhai Revolution in 1911, many modern Chinese schools were set up in Singapore and other parts of Southeast Asia under the influence of Chinese revolutionaries such as Dr Sun Yat-Sen (Tan, 2013). Chinese was taught as a first language in these schools using textbooks and materials from mainland China; and Chinese culture, Chinese nationalism and patriotism were inculcated and fostered. Students also learned subjects such as history, geography, mathematics and science in Chinese. The structure of these schools followed the education system in mainland China: six years of primary schooling, three years of junior middle schooling and three years of senior middle schooling (the 6-3-3 system). Common examinations for students in Chinese stream schools were initiated by the Hokkien Association in 1931. Senior Middle III examinations were conducted for students upon the completion of senior middle schooling till 1961 when the government implemented the Government Secondary IV School Certificate (Chinese) examination.⁶ Students

⁶ As mentioned earlier, the Government Secondary IV School Certificate (Chinese) examination was replaced by the Singapore-Cambridge GCE O-Level examination in 1971.

who performed well in the examinations had the opportunity to further their studies when Nanyang University, Singapore's only Chinese language post-secondary institution, was established in 1956.⁷

Although the British government felt that its own interests would be best served by the English elite, little was actually done to further these ideals other than to provide free or subsidized education for students in English stream schools (Gopinathan, 1974). The fundamental reason for this provision being that, as Bokhorst-Heng (1998a: 136) aptly sums up, 'access to English needed to be managed in close tandem with the administrative needs of the colony [...] anything more than that would certainly result in social instability'. As the British government believed that mass English education might not be beneficial for colonial order and sovereignty, they adopted a generally neutral attitude towards Chinese stream schools, as with other vernacular schools (Bokhorst-Heng, 1998b).⁸ The Chinese stream schools existed alongside English stream schools and were largely managed and funded by the Chinese community itself. Even though students in Chinese stream schools were exposed to English, and Chinese as a second language was introduced as an optional subject in English stream secondary schools in 1938, most students remained monolingual. At the societal level, the Chinese community became segmented into the Chinese and Chinese dialect-speaking majority and an English-speaking elite minority. Perhaps even more worrying was the mounting inter racial tensions due to the lack of a common language and identity.

3.3.2 Bilingual policy rationale

In May 1959, Singapore achieved full internal self-government. The first fully-elected government was formed, with the People's Action Party (PAP) winning most of the seats to form a majority administration. The leader of the PAP, Lee Kuan Yew, was declared the first Prime Minister of Singapore (1959-1990). In 1965, Singapore became a sovereign state after separating from the rest of Malaysia. Its political

⁷ Nanyang University was established in Singapore in 1956. During its existence, it was Singapore's only Chinese language post-secondary institution. In 1980, Nanyang University merged with the University of Singapore to form the National University of Singapore.

⁸ The British colonial government, however, took more interest in Malay stream schools as Malays were recognized as indigenous (Chia, 2015).

leaders were immediately ‘faced with the unenviable task of ensuring the political and economic survival of the small city state’ (Goh & Gopinathan, 2008: 12). At independence, Singapore was a plural society with neither a common language nor a unifying social, cultural or religious system. This lack of cohesion accentuated the racial and ethnic fault lines which made the nation increasingly fragile and vulnerable. Riots and bloodshed in newly independent nations such as India and Sri Lanka and Singapore’s own turbulent beginnings had taught Singaporean leaders that devastating consequences could occur if inter-racial tensions were not addressed. At the same time, Singapore was plagued by economic problems. The economy was suffering from high population growth and significant unemployment. The entrepôt trade system that Singapore inherited from its colonial past was insufficient to sustain, let alone grow, the economy. In the absence of natural resources and arable land, Singapore had to quickly adopt an export-oriented industrialization strategy.

To mitigate these twin challenges, the Singaporean government readily harnessed the usefulness of the English language. First, English in principle is a ‘neutral’ language, giving no ethnic group an advantage. Although the Chinese formed the vast majority of the population, ‘making Chinese the official language was out of the question as the 25% who were non-Chinese would revolt’ (Lee, 2011). Similarly, it would have been unlikely that the predominantly Chinese society would have accepted Malay or Tamil as a substitution for Chinese and Chinese dialects. At the outset, designating English as the official working language and main medium of instruction in schools would have been seen to favour those who already had an education in English (Lee, 2008), but the government contested that in the long term the choice of English would create ‘an open level field [...] and equal opportunities’ for ‘all Singaporeans, whatever their race’ (Goh, 1999). Second, the last half of the twentieth century saw English fast becoming the lingua franca of the world. British political imperialism had spread English around the globe during the nineteenth century and after the Second World War, the widespread use of English was further reinforced by the economic supremacy of the new American superpower (Crystal, 2012). The Singaporean leaders viewed English as the up and coming language of international commerce and industry and the key for guaranteeing access to Western science and technology. The use of English has been fervently defended for its utilitarian value

since the early years of Singapore's independence. The late Prime Minister Lee Kuan Yew argued strongly for the necessity of English (Josey, 2012: 589):

The deliberate stifling of a language which gives access to superior technology can be damaging beyond repair. Sometimes this is done, not so much to elevate the status of the indigenous language, as to take away a supposed advantage a minority in the society is deemed to have, because that minority has already gained a greater competence in the foreign language. This can be most damaging. It is tantamount to blinding the next generation to the knowledge of the advanced countries. Worse, it leads to an exodus of the bright and the promising who do not intend themselves or their children to be blinded from new knowledge.

Ironically, while English was seen as a necessity for Singapore's survival, it was also perceived as a significant threat to the nation. Since Singapore gained independence in 1965, various political leaders of Singapore have expressed deep concern over the excesses of westernization that the English language indirectly propagated. Such westernization, if left unchecked, they argued, could lead to an erosion of moral and personal values, which in turn would weaken the fabric of society. This fear of 'deculturalization' was clearly articulated by the late President Wee Kim Wee (1985-1993) in the following address (Wee, 1989):

Singapore is wide open to external influences. Millions of foreign visitors pass through each year. Books, magazines, tapes, and television programmes pour into Singapore every day. Most are from the developed countries of the West. The overwhelming bulk is in English. Because of universal English education, a new generation of Singaporeans absorbs their contents immediately, without translation or filtering. This openness has made us a cosmopolitan people, and put us in close touch with new ideas and technologies from abroad. But it has also exposed us to alien life-styles and values. Under this pressure, in less than a generation, attitudes and outlooks of Singaporeans, especially younger

Singaporeans, have shifted. Traditional Asian ideas of morality, duty and society which have sustained and guided us in the past are giving way to a more Westernized, individualistic, and self-centred outlook on life [...] The speed and extent of the changes to Singapore[’s] society is worrying. We cannot tell what dangers lie ahead, as we rapidly grow more Westernized.

To counter Westernization, the post-independence government reasoned that the identity of Singaporeans must be anchored in their ethnic and cultural origins. The cultural role of the ethnic languages, or mother tongues was given prominence; and the learning of mother tongue, either as a first or second language, was made compulsory for all Singaporean students with the introduction of the bilingual policy in 1966. It is worth noting at this juncture that the perceived dichotomy between West/East, utilitarian/cultural, modern/traditional, decadent/virtuous, has been criticized by some politicians and scholars as problematic if not specious and ‘dangerously simple-minded’ (Tan, 2003; Ho & Alsagoff, 1998; Woon, 1992). As a Singaporean poet so succinctly puts it in a quatrain poem: ‘The East is red, /the West is blue, /Elvis is dead, /Confucius too’,⁹ Singaporeans are a complicated mix of collective and individualized values—they embrace neither Confucian ideals nor Western mores whole-heartedly. I will revisit this theme when I relate issues in Chinese language assessment to the bilingual policy as a whole in the following subsection which addresses the limitations and implications of the bilingual policy.

The three chosen mother tongues (Chinese, Malay and Tamil), together with English, fitted neatly with the nation’s four major ethnic blocs of Chinese, Malay, Indian and ‘Others’. By granting the ‘corresponding language’ of each ethnic bloc equal official status and legitimacy, the government was seen, in a broad sense, to grant cultural recognition to the multi-ethnic population (Tan, 2003). In the years following the implementation of the bilingual policy, the enrolment of children in Chinese, Malay and Tamil stream schools fell sharply. Between 1968 and 1978, the number of students enrolled in Chinese stream schools declined rapidly from 18,927 to 5,289

⁹ The poem, first featured in Damien Sin’s collection of poems (Sin, 1998: 18), was quoted in *Time* magazine (McCarthy & Ellis, 1999) in an article on politics, controversy and the arts and culture scene in Singapore.

students. The same decade witnessed an increase in enrolment in English stream schools, from 34,090 to 41,995 students. The national stream was thus introduced in 1983 as a result of the overwhelming preference of parents for an English-medium education (Yip, Eng & Yap, 1997). English was taught as a first language in all national schools and mother tongue relegated to a second language for the majority of students. Singapore's bilingual policy, with English as the dominant language and the mother tongues as transmitters of traditional values and culture, remains the bedrock of the state's education system and ideology.

The epochal scale of change in education from 1959 to the early 1980s can be seen as a consolidation of political power in a Foucauldian sense. A discourse of national survival was repeatedly drawn upon to reinforce the narrative of a young nation in crisis and conflict. It is no coincidence that education in the two decades following self-government has also been dubbed the 'survival-driven phase' (OECD, 2011: 161). In the words of the late Prime Minister Lee Kuan Yew (1965), 'for (Singapore), survival has always been hazardous [...] We are on our own [...] in the centre of an extremely tumultuous arena of conflict'. The people of Singapore were therefore called upon by political leaders to 'exercise self-restraint and self-sacrifice' (Lim, 1965) for the sake of the nation's survival. This official discourse provided justification for the policies that the ruling party implemented, which in turn legitimized its authority.

As illustrated above, the fragility of Singapore's economic structure and social fabric was perpetuated through government speeches, documents, the media and education to form a powerful political oratory. What ensued after self-government was not only concern about the survival of a fledgling nation but also the sustainability of a young and vulnerable political party. While the much less apparent and influential narrative of a political party coming into power is equally worth studying, it is beyond the scope of this study. Nonetheless it is necessary to point out that Chinese stream schools were perceived by the ruling party as hotbeds of communist-aligned political activities, including demonstrations and rioting, which served to undermine its leadership (Trocki, 2005). The PAP holds the view that communist organizations, 'knowing how dear Chinese education, language and culture were to the [Singaporean] Chinese [...] exploited these issues to the hilt' and rallied students in

Chinese stream schools against the authorities (Singh, 2015: 210). The threat became more menacing when the leftist faction of the PAP splintered to form the Barisan Sosialis (Socialist Front) in 1961, which the late Prime Minister Lee Kuan Yew labelled a ‘communist-front organization’ (Josey, 2012: 66).

The next few years in the 1960s saw a battle for political hegemony between the PAP and the Barisan Sosialis. In the process of assuming supremacy, the PAP leadership took punitive measures against those identified as communists or pro-communists, by detaining key Barisan Sosialis figures, student leaders and trade unionists under the Internal Security Act (Kwok, 2001). Although the PAP emerged victorious, it also attracted criticism, especially from academics abroad, who argued that the eventual disappearance of Chinese stream schools and the relegation of Chinese to a second language was a strategic move to keep the Chinese-educated in check (Trocki, 2005; Tremewan, 1996). These critical academics postulated that the PAP, through government policies, had strengthened the social and economic forces that favoured the dominance of the English language; and that the bilingual policy, while quelling potential revolts of the Chinese-educated, had left them and the Chinese language marginalized. Such marginalization, along with the generally lower socioeconomic status of the Chinese-educated, became even more pronounced with the demise of Chinese stream schools in the early 1980s. In 1991, *Lianhe Zaobao* (《联合早报》), the local Chinese daily newspaper, published a series of articles that documented the ‘resignation and agony’ felt by Chinese intellectuals in Singapore.¹⁰ These sentiments continued to be echoed by opposition parties in an attempt to appeal to the Chinese masses, even in the recent 2011 general election (Koh, 2011; Gopinathan, 2003). While the recurring theme of political and language marginalization needs to be addressed with regard to Chinese language assessment in the Singaporean context, - it is beyond the scope of this study.

In conclusion, unlike the British colonial government which adopted a laissez-faire attitude towards education in Singapore, the newly-elected PAP government views education as a powerful tool of state control and regulation. Using Foucault’s

¹⁰ These articles were penned by the former Director of the Institute of Education, Singapore, Dr Lau Wai Har, who documented the low morale and frustrations of Chinese intellectuals and argued strongly against the labelling of Chinese intellectuals as ‘chauvinists’. For a more thorough discussion, see Gopinathan (2003).

concept of discipline and punish makes it possible to see the bilingual policy, streaming system, standardized curriculum and national examinations as mechanisms for amassing and wielding power (Foucault, 2012). It is not uncommon in modern societies for the government to instil discipline in the individual through the intersection of hierarchical observation, social definitions of normality, material institutions and rituals of examination. In fact, humble modalities introduced after Singapore's independence, such as the everyday flag-raising and pledge-taking ceremony in all schools, the display of portraits of the president and their spouse in school halls, to more major moves, such as the compulsory study of Civics and the establishment of the Institute of Education in 1973 to provide centralized training for teachers,¹¹ could all be seen in this light. By controlling education, the government is shaping official discourse. Foucault (2012: 183) drove the message home in the chapter titled, *The Means of Correct Training*:

The individual [...] is also a reality fabricated by this specific technology of power that I have called 'discipline'. We must cease once and for all to describe the effects of power in negative terms: it 'excludes', it 'represses', it 'censors', it 'abstracts', it 'masks', it 'conceals'. In fact, power produces; it produces reality; it produces domains of objects and rituals of truth.

Drawing on this Foucauldian version of 'reality', the limitations and implications of the bilingual policy in Singapore are addressed next.

3.3.3 Limitations and implications of the bilingual policy

As outlined above, discussion of Chinese language and assessment in Singapore is problematic because of the country's complex ethnic-linguistic composition. To further complicate matters, the linguistic legacy in Singapore, derived from its historical development, has been tempered by a policy of bilingual education implemented since 1966. Bilingualism in Singapore has taken on a meaning peculiar

¹¹ In 1991, the Institute of Education merged with the College of Physical Education, which had been set up in 1984 to train specialist teachers in Physical Education, to form the NIE. NIE is the sole teacher education institute for teachers in Singapore.

to the needs of the country. It is defined as proficiency in English plus one of the officially recognized mother tongues, namely, Chinese, Malay or Tamil, which is automatically assigned according to ethnicity. The policy clearly compartmentalizes the role of English and mother tongues in Singapore's society—English functions as the 'elaborated code' being the language of education, government and commerce; while mother tongues function as 'restricted codes', used in informal intra-ethnic community interactions, acting mainly as 'cultural ballast' against undesirable Western influences (Bernstein, 1971). The government's position on the relationship between English and mother tongues was clearly laid out in a statement by Dr Tony Tan Keng Yam, former Minister of Education (1985-1991) (Tan, 1986):

Our policy of bilingualism that each child should learn English and his mother tongue, I regard as a fundamental feature of our education system [...] Children must learn English so that they will have a window to the knowledge, technology and expertise of the modern world. They must know their mother tongues to enable them to know what makes us what we are.

Essentially, what exists in Singapore is an 'English-knowing bilingualism' (Kachru, 1992), a term which acknowledges the primacy of the English language in defining what it means to be a bilingual person (Pakir, 1997). The goal, as specified by Tan, is to 'educate an entire population so that everyone is *literate* in English, and at the same time, has a *reasonable knowledge* of his mother tongue' (The Straits Times Editorial, 1990, emphases added). While the bilingual policy has helped in ameliorating the problem of illiteracy in the post-independence years and increased the proportion of Singaporeans with a minimum standard of proficiency in both English and their mother tongue, it is not without flaws. It has been pointed out that decisions about the bilingual policy are primarily made by political leaders and then communicated to subordinate levels which are then charged with the technical, managerial, and administrative tasks of putting policy into practice (Ng, 2011; Kuo & Jernudd, 1994). Given the highly centralized and regulated nature of the bilingual policy, there are bound to be gaps between the intended, enacted and experienced policy. The top-down approach in decision making and implementation may also

mean a delayed response to changes at ground level. Ball (1994: 10) takes these concerns into account when commenting about policy at large:

Policy is both text and action, words and deeds, it is what is enacted as well as what is intended. Policies are always incomplete insofar as they relate to or map on to the ‘wild profusion’ of local practice.

In what follows, the scope of discussion on the inherent limitations of the bilingual policy is limited to four aspects that are of direct relevance to Chinese language assessment.

First, the bilingual policy entails a reconceptualization of the internally heterogeneous Chinese population into one community with one mother tongue paired with one set of culture and values. Chinese of various descendancy and background were cast in a fixed over-simplified ethnic category and a homogenous notion of *Chineseness* was socially engineered to foster intra-ethnic cohesion (Guo, 2011; Chua, 2003). At the time of independence, the Chinese community remained divided into the English-educated and the Chinese-educated. Various Chinese dialects, or vernacular Chinese, were spoken as the predominant home language by the majority of Chinese although Chinese was well established as the language of Chinese education (Chua, 1964). In the span of a decade following the *Speak Mandarin Campaign* launched in 1979, the government successfully curtailed the use of dialects—the proportion of Chinese families who spoke mainly dialects at home declined steeply from approximately 62% in 1980 to below 10% in 1989, a downward trend which continued.¹² Since 2001, fewer than 2% of Chinese students in each primary one cohort have come from dialect-speaking homes. Chinese,

¹² The *Speak Mandarin Campaign* (讲华语运动) is a government initiative to promote the use of Chinese (i.e. Mandarin). The campaign was launched in 1979 by the late Prime Minister Lee Kuan Yew with the objective of persuading all speakers of Chinese dialects to switch to Chinese. The campaign continues to be an annual event, although its focus now encourages English-speaking Singaporean Chinese to use the Chinese language more frequently. Campaign slogans in the past decade have included, *Immerse Yourself Today. Mandarin. It Gets Better with Use* (华文华语, 多用就可以) (2015), *Be Heard in Chinese* (华文? 谁怕谁!) (2009) and *Mandarin COOL!* (华语 COOL!) (2007).

however, did not become the unifying thread within the Chinese communities as envisioned.

Despite the shift away from Chinese dialects, a significant and growing proportion of Chinese speak English in the private informal sphere of family and friends (Ng, 2014). This changing language landscape suggests that the Chinese community is now segregated into English and Chinese speakers, as argued by several academics (Goh, 2010). In reality, available data from Singapore's national census reports, international comparison studies and small-scale studies reveal a structure more complex than a rigid dichotomy (Beardsmore, 2003). The relationship between language and culture has been an area of concern for researchers but has not been given due attention in the discussion of bilingual policies by policy makers (Tan, 2004). Beardsmore (2003) hypothesizes that there is seldom a neat correspondence between a language and culture. The bilingual policy in Singapore and its resulting language shift, he observes, has produced at least five different types of Singaporean Chinese, namely the monocultural monolinguals, monocultural bilinguals, bicultural bilinguals, bicultural monolinguals and acculturated mono-/bilinguals (Beardsmore, 2003: 87).

Such linguistic and cultural diversity has a direct impact on the design of the O-Level Chinese language examination, which caters for the majority of the Chinese students. Although test-takers are in the same age group and have received similar hours of classroom training, they have varying linguistic profiles and cultural perceptions and hence, different expectations of and attitudes towards the examination. The study does not set out to investigate the statistical significance of each of the five types of test takers, or for that matter, Singaporean Chinese. It is anticipated, however, that such information, which is not comprehensive at the moment, would be useful in the management of Chinese language examinations in Singapore and should not be overlooked by the MOE and SEAB.

Second, the concept of mother tongue in the Singaporean context is hugely problematic as it undermines the legitimacy of the bilingual policy. The United Nations Educational, Scientific and Cultural Organization (UNESCO) (1953: 46) defines mother tongue as 'the language which a person acquires in early years and

which normally becomes its natural instrument of thought and communication'. Mother tongue is generally accepted as a language which the child first listens to and speaks and is often used at home (UNESCO, 2001). This language is perceived to frame thinking as a child is essentially learning how to think through the initial acquisition of their mother tongue (Simpson & Wigglesworth, 2008). Mother tongue is the language that is inextricably linked to cognitive development, as best captured in the words of Halliday (1978: 205), 'A child who is learning his mother tongue is learning how to mean. He is building up a potential, a potential for symbolic action which in a large measure is going to determine the kind of life he leads'. In Singapore, however, mother tongue is automatically ascribed based on ethnicity, irrespective of home language. Hence, a 'mother tongue' and a 'mother's tongue' are not always the same, and 'mother tongue' should not be confused with 'first language', 'native language' or 'dominant language'.

Bearing in mind that mother tongues are state-assigned in Singapore, it comes as no surprise that there is a group of Chinese students who neither feel comfortable using the Chinese language nor identify themselves readily with the Chinese culture. For this group of students whose dominant language is often Standard English (or its vernacular form, Singlish), Chinese is more of a 'step-tongue' (继母语) rather than mother tongue (Chew, 2007). Even though the GCE O-Level Chinese language curriculum and examination are pitched at second language level, many of these students struggle with the learning of the Chinese language. While most acquire a minimum level of oral proficiency owing to the presence of a sizeable community of Chinese speakers in Singapore, few find the opportunity and motivation to read (and write) in the language. This situation necessarily raises questions about the objectives and authenticity of the GCE O-Level Chinese reading examinations. Ensuring that this group of predominantly English-speaking Chinese students passes the national Chinese language examinations has also become a constant concern among parents, educators and policy-makers (Loke, 1994).

Third, another term that warrants discussion is 'bilingualism'. Singapore's bilingual policy has come to be accepted by the general public without any critical engagement with the key issue of bilingualism itself. What is bilingualism? What are the dimensions of bilinguality? What level of proficiency in both languages must a

student achieve to be legitimately called bilingual? The answers to these questions remain unclear and key terms undefined in the Singaporean context. Increasingly, academics have voiced their concerns about the efficacy of Singapore's bilingual policy in producing effectively bilingual students. In the rhetoric of the policy, English is taught at first language level and Chinese at second language level for approximately 90% of ethnic Chinese students. Furthermore, English is used as the primary medium of instruction and assessment in schools and Chinese is mainly taught and learned as a single language subject. With an average of only 3.75 hours of instruction time per week at secondary school level, many students grow up to function predominantly in English. As pointed out by Beardsmore (2003: 90):

Few specialists consider as bilingual education any programme that does not use two languages both for subject-matter and content-matter learning. They consider the learning of a language as a subject akin to any other school material as unlikely to produce sophisticated bilingual proficiency, let alone bicultural sensitivity.

James (2003) elaborating on the idea of language and power similarly observes that bilingualism in Singapore is in actual fact, highly selective. Students who offer both English and Chinese as first language at secondary school come from the top 10% of the PSLE candidates, although provisions have been made in recent years to expand this group of students (MOE, 2015b).¹³ Even fewer sit Chinese language or literature papers at the end of their post-secondary education, with only a handful of academically inclined students being hand-picked for the Chinese Language Elective Programme and Bicultural Studies Programme (Chinese) in junior colleges. Effective bilingualism is therefore seen to be reserved for a group of elite students.

For the majority of students, Chinese is in essence a single subject which discontinues after secondary school. Loke (1994) presents a foreboding scenario where Singapore becomes a functionally monolingual English-speaking country, not

¹³ Students in the top 11% - 30% of the cohort who meet the language criteria (i.e. an A* grade in Chinese or at least a Merit in Higher Chinese at primary school) may also be offered Higher Chinese (Chinese as a first language) at secondary level. Schools may also allow students who do not meet the above criteria to opt for Higher Chinese if they are assessed as having exceptional ability in Chinese and are able to study Chinese at a higher level without affecting their performance in other subjects.

very different from other predominantly English-speaking countries such as Britain, America, Australia and New Zealand. Whether this prognostication is accurate is, of course, debatable; however we must acknowledge the hegemony of English in Singapore's education system. After half a century of conscious language planning, most Singaporean Chinese have at least a rudimentary level of spoken Chinese. Yet, the number of Chinese who demonstrate high levels of competency in all four language skills has shrunk noticeably. Kirkpatrick (2010) notes that producing students who are truly effective in the Chinese language is an uphill task as there is simply not enough curriculum time to read and write Chinese under the current 'English + 1' bilingual policy. As early as the 1980s, some parliamentarians lamented that 'standards in Chinese had declined to such a degree that students could neither write a simple essay nor read Chinese newspapers with comprehension' (Gopinathan, 2003: 28). The problem persists: there is increasing evidence, as cited by the Chinese Language Curriculum and Pedagogy Review Committee (CLCPRC), to show that the Chinese language proficiency level of the average Chinese student in Singapore is still in decline and Chinese language teachers in Singapore are facing greater challenges in motivating students to learn the language (CLCPRC, 2004).

The ruling political party asserts that it is unrealistic to expect standards in Chinese as a second language to be comparable to those achieved when students used Chinese as a medium of instruction. Such a trade-off, it argues, is inevitable. The learning of two non-cognate languages simultaneously is deemed highly demanding and the government will not allow the curriculum to be overloaded by requiring higher standards in Chinese from all Chinese students. When asked in the late 1990s about the impact of an emerging China, the late Prime Minister Lee Kuan Yew (The Straits Times Editorial, 1997) famously said:

China may be the greatest power in the world. The Chinese language may be one of the world's leading languages. We stay where we are, bilingual. Working language English, everybody level playing field. Or be prepared for big trouble. You know, the tide is not receding.

It seems unlikely therefore that the government's stand on bilingual education and Chinese language will change drastically in the near future. Given its limited room to

manoeuvre, the Chinese language reading curriculum and assessment in Singapore might be unable to respond as quickly to the demands necessitated by new global trends. A growing literature emphasizes that reading is best taught and learned when it is put to work in the service of other purposes, activities and learning efforts (English Language Institute of Singapore, 2014; Pearson, 2009). Reading instruction in the English language can be achieved through the teaching of other subjects (or disciplines) in Singapore. The same cannot be said of the Chinese language. As explored in Chapter 2, the Chinese language being a single subject in most Singaporean secondary schools, would find it almost impossible to cultivate and assess the disciplinary literacy much sought after by students today.

Fourth, the underlying premise of Singapore's bilingual policy, namely that English and Chinese assume sharply different roles in society, has broken down rapidly in recent years, with English infiltrating into social spheres and Chinese being increasingly promoted as an economically valuable language. Building on research by Canadian linguist William Mackey (1987), Beardsmore (2003) maintains that linguistic compartmentalization, both at an individual and societal level, shifts across time depending on the demands and needs of both the society and its members. Bilingualism, which normally implies linguistic compartmentalization, is therefore by definition, unstable.

As evident in the discussions above, English is now very much the language of education in Singapore. As a result of the younger generations being educated mainly in English, English has penetrated beyond formal domains, progressively becoming a language of personal communication, informal interaction and cultural expression (Ng, 2014). The government's dichotomized view of English as having economic utility and Chinese as having cultural functions has generated different attitudes towards these languages (Zhao & Liu, 2010, 2008), especially because English language proficiency to a large extent determines career progression and socio-economic status (Silver, 2005). More Chinese parents are choosing English as the preferred language of communication with their children (MOE, 2011), speeding up the rate at which English infiltrates the social sphere. Even the initiator of the bilingual policy, the late Prime Minister Lee Kuan Yew, expressed concern that the pendulum had swung too much in the direction of English (Toh & Ong, 2011). Yet,

at the same time, new elements in favour of learning Chinese have also entered the policy frame. China's remarkable economic growth and emergence as a political and technological powerhouse have strengthened the currency of the Chinese language.

Since initiating market reforms in 1978, China has rapidly changed from a centrally planned system that was largely closed to international trade to a market based economy with a growing private sector (The World Bank, 2015). In 2010, China surpassed Japan to become the second largest economy in the world after the United States of America. Many countries are recognizing the study of Chinese language and culture as a strategy to ensure the global competitiveness of their citizens in the future. Former British Prime Minister, David Cameron (2010-2016), advocated the learning of Chinese in British schools during his official visit to China in 2013. Cameron urged students to look beyond the traditional focus on French and German and instead learn Chinese, the language that will 'seal tomorrow's business deals' (The Guardian Editorial, 2013). In the United States of America, Chinese as a foreign language is also growing in popularity. A recent survey by the Modern Language Association (2015) reported that over 61,000 students are studying the language in colleges and universities in the United States of America, a number that has more than tripled since the mid-1980s. Reports such as the *Expanding Chinese Language Capacity in the United States* (Stewart & Wang, 2005) called for a national commitment to new investments in teaching Chinese language and culture. The learning of Chinese has become a growing global phenomenon, with the estimated number of non-native learners in the world exceeding 100 million at present (Statista, 2017).

Being a small nation committed to a pragmatic ethic, Singapore's education and assessment system is invariably shaped by global trends and imperatives. National survival necessitated a fundamental restructuring of the education and assessment system very early in the life of the nation and today the expanding influence of the Chinese language undoubtedly calls for a review on how it is taught and assessed. With China set to be the world's largest economy in the near future, it can be asked whether Singaporean students are ready to take full advantage of their language skills to engage with China. As more people around the world learn Chinese as a second language, Singaporean Chinese will have to increase their competence in the

language to retain their competitive edge. That is to say, it may no longer suffice to have conversational fluency in informal and casual settings. To take full advantage of the rise of China and its attractive market of 1.3 billion consumers, Singaporean Chinese will have to be adept users of the language even in formal settings. Being able to read critically in the Chinese language with speed and efficiency is becoming a prized work skill in the age of the Internet and information explosion.

In view of China's growing global influence, Singapore's government has since the 1990s promoted the learning of Chinese as a doorway to trade and business dealings with China. Such linguistic instrumentalism has been repeatedly reinforced through various government initiatives such as the Speak Mandarin Campaign (讲华语运动) and Business China (通商中国). As encapsulated in the MOE's directives on mother tongue education, Chinese students are encouraged to study the Chinese language for as long as possible and to as high a level as they are capable of in order to ride the wave of growth in China (CLCPRC, 2004). Chinese language is no longer constrained to the singular purpose of maintaining traditional values and providing a sense of Chinese identity. It has stepped into the sphere of economy, a domain traditionally reserved for the English language in Singapore.

The blurring of the dividing line between the functions of English and Chinese has exposed conflicts and tensions underlying Singapore's bilingual policy. On the one hand, the government affirms the legitimacy of Chinese as a powerful resource for opportunities in China; on the other hand, Chinese remains a single subject in government schools, with far less prestige afforded to it in comparison with English. In particular, the latest Chinese language curriculum and assessment reforms have arguably further lowered the common standards for the subject. In the paper *Planning for Development or Decline? Education Policy for Chinese Language in Singapore*, Curdt-Christiansen (2014) wrote of the educational uncertainty and cultural confusion that stem from the conflicting ideologies behind the nation's language planning and use.

While the promotion of the Chinese language for its richness and commercial benefits was underway, Curdt-Christiansen (2014) noticed that the educational discourse took a different turn. In 2004, MOE introduced the Chinese Language 'B'

Syllabus at secondary school level. This simplified syllabus which gives higher weighting to practical communication skills is ‘designed for students who, despite additional support in school and beyond, have exceptional difficulties coping with the Chinese language’ (MOE, 2004a). Students who obtain a pass in Chinese Language ‘B’ will be deemed to have met the mother tongue requirement for admission to junior college. In other words, students are no longer required to include their grade for the GCE 1162 Chinese language paper in their overall score for academic progression. In the same year, the CLCPRC proposed that for the majority of students, the emphasis should be on effective oral communication, followed by reading and then, writing (先听说、再读、后写), as ‘in adult life, most Singaporeans will more often hear and speak Chinese than read Chinese, and more often read than write Chinese’ (CLCPRC, 2004: ii). To solve the problem of increasing language learning difficulties and to enthuse students in learning Chinese, the content of the secondary syllabus was reduced while the curriculum and assessment no longer emphasized full command of literacy skills. In 2009, the Chinese curriculum underwent further change to allow teachers the flexibility to use English to facilitate the teaching of Chinese. All these changes have done little to elevate the status of the Chinese language in Singapore. On the contrary, Curdt-Christiansen argues that they run counter to the government’s efforts to promote Chinese as a language in vogue. The government’s ‘mixed messages’ may make it difficult for students to ‘appreciate the value, be it economic, cultural or educational, of the Chinese language’ (Curdt-Christiansen, 2014: 23).

3.4 Conclusion

When this chapter was prepared in 2015, Singapore was gearing up for its 50th anniversary of independence. It seemed especially appropriate to reflect on ‘the hard truths’ articulated by the late Prime Minister Lee Kuan Yew (Rose, 2009):

You cannot maintain your relevance by just staying put. The world changes. There are shifts in the geopolitics and the economics of the world. We have to watch it and ride it. You surf with them. As the surf comes this way you ride the surf.

As a young nation, Singapore did not merely survive, it thrived. In the words of the late Prime Minister Lee Kuan Yew, Singapore was able to stay ahead of the game because of its ability and willingness to reinvent itself and to stay relevant. Such thinking remains integral in today's competitive world. In the field of educational testing, the late Prime Minister Lee Kuan Yew's 'hard truths' challenge policy makers, specialists and education practitioners to re-think and re-examine why and how the Chinese language is being assessed in light of the changing local and global linguistic landscapes. To a small and vulnerable state like Singapore, the major impetus for any revamp and restructuring is undoubtedly economic. At the same time, any policy refinements will have to ensure that the nation's founding principle of racial equality is not compromised. A balance must therefore be struck between promoting the Chinese language and accommodating the sensitivities of non-Chinese ethnic groups.

In this chapter, I have traced briefly the historical trajectory leading to the implementation of Singapore's bilingual policy. I have also highlighted some of the inherent limitations of the bilingual policy and their implications for Chinese language education and assessment. It is clear that there are many factors at play simultaneously in the design and operation of a national language examination. This raises the fundamental question of who or what factors are able to define the constructs of Chinese language proficiency, and more particularly the constructs of CL2 reading comprehension, in Singapore. Specifically, it needs to be asked whether institutions such as MOE, SEAB and UCLES exercise direct control over the determining of the constructs or whether the needs of the Singaporean Chinese community and demands of a global workplace provide the major influences. Questions should also be raised regarding the extent to which ex normative standards and new developments in the field of educational testing are determinants. Even if tensions among various factors can be reconciled and a clear, consistent set of test constructs agreed, the need remains for coordinated efforts to ensure that the examinations measure that which they propose to measure. Further complicating matters is the fact that some constructs such as reading motivation and critical reading skills are deemed hard to measure for a variety of reasons such as intangibility, the lack of a widely accepted definition and subjectivity of scoring.

Notwithstanding the accumulating volume of literature on reading and reading assessment, our knowledge in these areas is still limited.

The purposes of the examination raise other issues that require attention. It appears that the GCE 1162 reading examination is used for several different purposes and these purposes have been reshaped, and will continue to be altered, as the power dynamics between Chinese and other languages in Singapore change. From qualification and placement, to institution and system monitoring, to social and programme evaluation, scores from this high-stakes examination have a multitude of intended and actual uses. It is useful to bear in mind ‘that a system which is fit for one purpose will not necessarily be fit for all purposes’ (Newton, 2007: 149). If validity is ‘the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment’ (Messick, 1989b: 13), having multiple purposes for one examination makes validity estimates much more challenging.

Thus, it is all the more important for all actors involved in the examination process to remain resolutely receptive to recommendations and current influences. It is also desirable that more opportunities for regular dialogues among policy makers, ground level personnel and the public are made available, so that new ideas and solutions can be generated. As academics Sharpe and Gopinathan (1997: 370) so aptly put it nearly two decades ago, ‘It is precisely the effectiveness of the established system in terms of conventional measures that opens up the possibility in Singapore of a redefinition of effectiveness’. More recently, Deputy Prime Minister and Finance Minister Tharman Shanmugaratnam (Khamid, 2015) has also called for ‘more debate and peer review within civil society itself, with participants evaluating each other’s analyses and proposals, and pointing to the trade-offs thoroughly and dispassionately’ so as to help Singapore mature and advance as a society. To sum up, to stay relevant and effective, we must not be afraid to abandon tried and tested methods that have produced enviable results in the past. With this in mind, the next chapter presents the methodology adopted for this study.

Chapter 4 Methodology

4.1 Introduction

I have analysed in detail the research questions and key concepts of validity, the reading construct and the Singaporean context in the first three chapters. Comprehensive coverage of these three key concepts and in-depth discussion of the relevant literature have been presented. In this chapter, the focus turns to how the research was planned and conducted, which in essence, concerns the methodology of the research. A methodology chapter connects the research questions and key concepts to the findings in order to produce an architecturally sound plan of the study. The three key components of methodology, namely the philosophical paradigm, research design and research methods are explored in turn in the following sections. In addition, ethical concerns and quality assurance measures are elaborated.

4.2 Philosophical paradigm

Any cogent research is built on the premise of thoughtful planning. An important initial step in planning research is to consider different philosophical paradigms and to adopt one that is consonant with the research design and methods. Philosophical paradigms or worldviews can be understood as how we perceive the world and, therefore, go about conducting research. All research includes assumptions, whether implicit or explicit, about the ontology, epistemology and axiology which a philosophical paradigm encompasses (Greene & Caracelli, 1997).

The main philosophical paradigms that are traditionally depicted as being in direct opposition are those of positivism and constructivism. In a brief outline, positivism claims a singular reality and researchers execute objective and unbiased studies to discover the one and only truth. Positivism is often associated with quantitative methods which are based on determinism, reductionism, empirical observation and measurement, and the acceptance or rejection of theories that are continually refined (Slife & Williams, 1995). Positivism is contrasted with constructivism which advocates multiple realities that need to be interpreted, and for this reason

constructivists tend to favour qualitative methods. Research guided by constructivism is often naturalistic, inductive and value-laden, where researchers begin with their data and work bottom-up to offer patterns, theories and generalizations (Burton & Bartlett, 2009). There are, of course, other influential philosophical paradigms such as critical theory, postmodernism, subjectivism, advocacy and participatory design and many more nuanced positions within these broad frameworks. I will, however, be focusing on pragmatism, the philosophical paradigm underpinning the present study. Pragmatism recognizes that not all research falls comfortably within a positivist or constructivist paradigm; it instead, embraces pluralism and is oriented towards ‘what works’ to best address the research questions (Feilzer, 2010).

Pragmatism, with its roots in the philosophic writings of Dewey (1925), James (1909) and Peirce (1878) highlights practicality and contextual responsiveness to the demands, opportunities and constraints of the research in hand. Pragmatism as a philosophical tradition originated in the United States of America at around the 1870s, dispensing with Hegel’s (1985) notion that philosophy aims at knowing what is imperishable, eternal, and absolute. Since the 1970s, pragmatism has undergone a revival through the work of notable modern pragmatists such as Rorty (1982) and Brandom (2011). The contentious ontological issue of what is reality is circumvented by pragmatism which accepts that philosophically, that there can be singular and multiple realities open to empirical inquiry. Pragmatism posits that objective truth is unattainable through the faculty of reason, hence the strength of propositions and hypotheses have to be judged by the results they produce when put into practice or in other words, their practical consequences (Diggins, 1994). At the epistemological and axiological levels, researchers are guided by practicality, collecting data to solve practical problems in the ‘real world’, taking both biased and unbiased perspectives (Creswell & Plano Clark, 2007). Freed from the dichotomy between positivism and constructivism, pragmatism can integrate aspects from different philosophical paradigms. Naturally, pragmatism calls for a convergence of qualitative and quantitative methods which explains why pragmatism appears to be the dominant philosophical paradigm employed in mixed methods studies (Creswell, Plano Clark, Gutmann & Hanson, 2003). Academics such as Morgan (2007) and Tashakkori and

Teddlie (2003) even go one step further to assert that pragmatism represents the single most appropriate philosophical paradigm for mixed methods studies.

The following sections link pragmatism to mixed methods research design and subsequently to the choice of research methods. Pragmatism's contributions to this study are two-fold. First, pragmatism enables the study to draw from the full complement of philosophical paradigms available. For example, whilst the interviewees and documents studied provided multiple realities as in the case of constructivism, the research was not purely inductive as it was structured by Weir's (2005) socio-cognitive validity framework and Kane's (2006) argument-based approach to validation. Second, pragmatism promotes the use of research methodologies and methods to answer research questions that 'aim at utility for us' (Rorty, 1999: xxvi). The study essentially aims to be a useful and grounded evaluation of the GCE 1162 reading examination, in order to inform future test designs and validation studies of similar nature through the triangulation of research methodologies, methods and data which are discussed below.

4.3 Research design

Research design refers to the plan of action that links the philosophical paradigm to specific methods (Crotty, 1998). As De Vaus (2001) argued, a research design constitutes the blueprint for the collection and analysis of data. To answer the research questions in this study as unambiguously as possible, an embedded mixed methods research design, which privileges a qualitative rather than quantitative approach, was chosen (Creswell & Plano Clark, 2007). Data collected from the different methods were then used in a convergent parallel approach to arrive at an overall interpretation.

It is useful to begin with a brief history and definition of mixed methods design. Researchers for many years have collected both qualitative and quantitative data within the same studies, however, mixing these datasets as a distinct research methodology is relatively new (Creswell & Plano Clark, 2007). The formative period for mixed methods design began in the 1950s and went through a challenging period from the 1970s to 1990s, which saw the paradigm wars over the qualitative versus

quantitative debate (Johnson & Onwuegbuzie, 2004). Paradigm purists were adamant that mixed methods research was incommensurable as it propounded that paradigms could be combined (Smith, 1983). The traditional quantitative methodology rested on a positivist paradigm while the qualitative methodology pivoted on a constructivist paradigm and paradigm purists argued that the differences between these paradigms were irreconcilable. It was not until the early 2000s that academics such as Johnson and Onwuegbuzie (2004), Tashakkori and Teddlie (2003) and Creswell (2003) succeeded in positioning mixed methods research as a natural complement to traditional quantitative and qualitative research. This study adopts Creswell and Plano Clark's (2007: 5) comprehensive definition of mixed methods research as follows:

Mixed methods research is a research design with philosophical assumptions as well as methods of inquiry. As a methodology, it involves philosophical assumptions that guide the direction of the collection and analysis of data and the mixture of qualitative and quantitative approaches in many phases in the research process. As a method, it focuses on collecting, analyzing and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone.

Using this definition of mixed methods research, I now proceed to explain the specific methods used in this study. Qualitative data were collected and interpreted simultaneously through semi-structured interviews and document analysis. Further, a part of the data amassed through document analysis was analysed quantitatively by a panel of subject matter experts (SMEs). The quantitative component embedded in the qualitative data provided a supportive, secondary role in this study. All findings were subsequently structured according to the main parameters that defined the scope of the study. The embedded mixed methods design is captured visually in Figure 4a. The capitalized abbreviation 'QUAL' is used in the figure to denote the dominance of the qualitative component in the study and the lowercase 'quan' to imply less emphasis on the quantitative component.

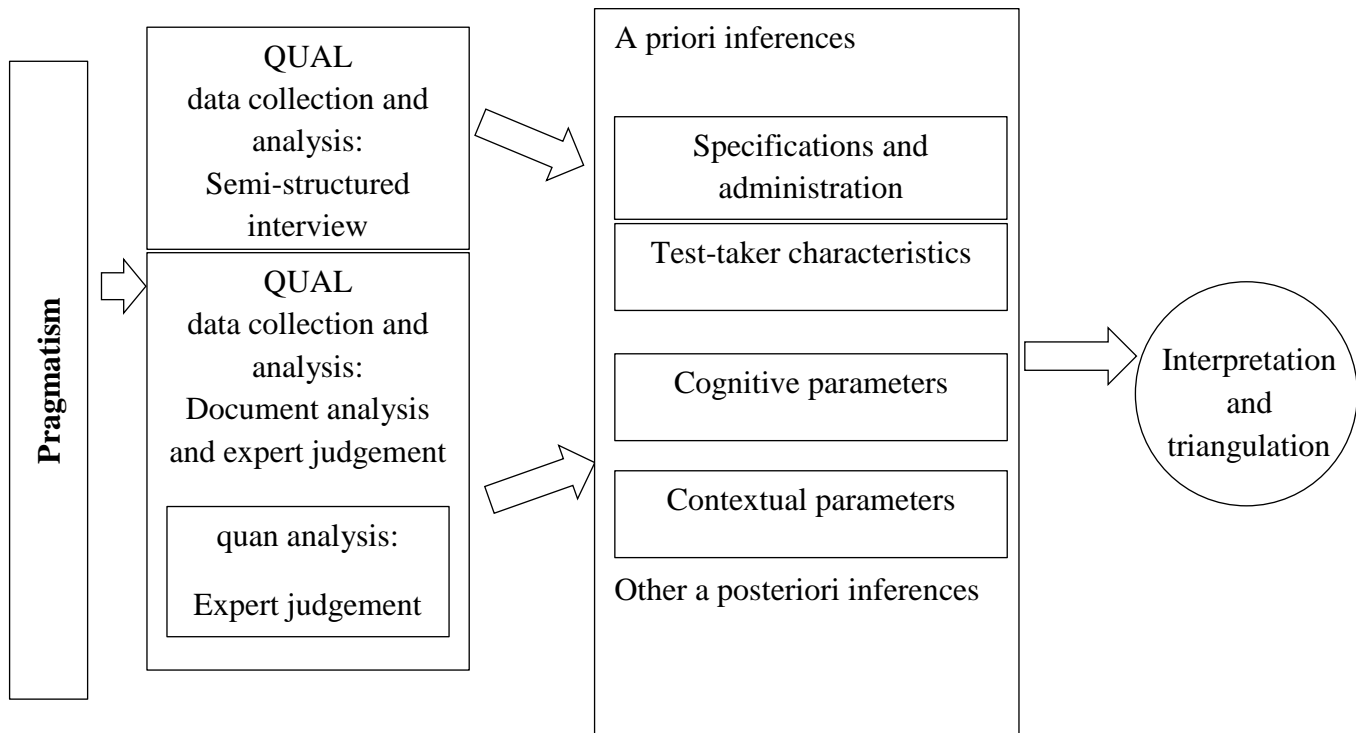


Figure 4a: Overview of the mixed methods research design used in this study

Methodologically, mixed methods research was chosen for this study underpinned by the rationale that the strengths of qualitative research can be complemented by the strengths of quantitative research while simultaneously compensating for the weaknesses of each methodology. The dominant qualitative method promotes listening to both the researcher and the researched and brings with it the strengths of sensitivity to meaning and in-depth understanding of smaller samples. In addition, qualitative methods, because of their exploratory nature, ‘tend to be oriented toward discovery of new phenomena and ways of understanding’ (Hesse-Biber, 2010: 64). Using qualitative methods to garner validity evidence is also in agreement with the recognition of validity as a unitary yet multi-dimensional concept. As Haertel (2013) observes, the measurement field has developed various quantitative methods to understand and explain score meaning where differential item functioning, scaling, norming, studies of score precision, reliability and generalizability are but some of the many examples. Considerably less effort and attention, however, are devoted to the qualitative aspects of validity evidence. Haertel (2013) contends that when it comes to accounting for the ways testing is supposed to function in the real world, such quantitative or technical evidence hardly suffices. Expanding the scope of

validity evidence to include qualitative evidence, I would argue, is therefore essential if the intended positive effects of testing are to be realized and the unintended effects avoided.

A quantitative methodological approach was added to the core qualitative approach when inspecting the GCE 1162 reading examination papers to foster a more robust understanding of the large sample of papers collected. Quantitative methods were used with the goal of profiling the cognitive and contextual features of the papers, tracing trends and formalizing comparisons which would be challenging with the use of qualitative methods alone. Triangulating within a mixed methods framework through the use of qualitative data collection methods, namely semi-structured interviews, document analysis and expert judgement, complemented by the descriptive statistical presentation of data amassed through expert judgement, aimed to increase the generalizability of the findings. The term triangulation is borrowed from land surveying, where ‘knowing a single landmark only locates you somewhere along a line in a direction from the landmark, whereas with two landmarks (and your position between the third point of the triangle) you can take bearings in two directions and locate yourself at their intersection’ (Fielding & Fielding, 1986: 23). In the same light, understanding the similarities and inconsistencies across data yielded from different methodologies can be illuminative. The resulting study is one that demonstrates a more complete and insightful account of the a priori inferences of the GCE 1162 reading examination.

4.4 Research method: Semi-structured interview

This section and the following section elaborate on the research methods used in the study, namely semi-structured interviews and document analysis. Each method is explored at three stages—at the pilot study, main study and data analysis stage.

Interviewing as a research method typically involves the researcher asking questions and, hopefully, receiving answers from people who are being interviewed (Robson, 2016). The way in which an answer is made, for example hesitation, the tone of voice and facial expression, can provide supplementary information that cannot be obtained through a written response. Although interviewee paralinguistics was not

analysed in this study, stress, pauses and laughter were still annotated in the transcripts for potential use in later studies. Interviews were incorporated into this study largely to find out what stakeholders know, perceive and feel about the GCE 1162 reading examination and more generally, Chinese as a second language (CL2) reading in the Singaporean context. The interviews involved semi-structured, open ended questions centred around the various a priori inferences. For semi-structured interviews, essential questions and a default order for the questions are predetermined ahead of the interview. However, additional unplanned questions may be asked during an interview to follow up on what an interviewee says. The pre-set order and wording of questions may also be substantially modified to allow for more natural and focused two-way communication. Interviewees are viewed as meaning makers, not passive conduits for retrieving information from an existing vessel of answers (Holstein & Gubrium, 1995).

The advantage of a semi-structured interview is that the researcher can adapt the research method to the interviewee's level of comprehension and articulacy, taking into account the fact that in responding to a question, people often also provide answers to questions that were going to be asked later (Fielding & Thomas, 2008). Semi-structured interviews allow systematic and consistent investigation that yield comparable qualitative data and yet afford enough flexibility to probe for elaboration when interesting points are raised by an interviewee. This type of interview also provides opportunity for identifying new ways of understanding the topic in hand (Lewis-Beck, Bryman & Liao, 2004).

The main limitation of interviewing as a research method is that the quality of the data collected is heavily dependent on the individual skills, opinions and idiosyncrasies of the interviewer. Moreover, an ethical challenge would be the openness and intimacy of the interview process which might lead to interviewees disclosing information that they later wish to withdraw, rendering the data void. Further, translation of interview transcripts potentially introduces bias and raises the question of how to ensure agreement on the translation. As Barrett (1992: 203) observes, researchers 'have accepted to varying degrees the view that meaning is constructed in rather than expressed by language'. Speaking for the interviewees in another language, therefore, encompasses a responsibility for the researcher to

represent them neutrally. I was the primary translator as I am fluent in both Chinese and English and well-acquainted with Singapore's assessment and education system. As an additional step to safeguard objectivity, selected data to be quoted in the study were also back-translated by a certified translator.

Another disadvantage is that interviews are often very time consuming. For instance, interviews were audio recorded, transcribed and coded using NVivo10, although with high-quality source materials and good typing speed, each hour of interview recording took at least four hours to transcribe. In addition to these processes, during the interview the main points were noted and shared with the interviewees at the end of each interview to avoid misinterpretation. From designing the interview schedule, obtaining ethical clearance and liaising with the interviewees, to transcribing, coding, analysing and translating the interview data from Chinese into English, the entire process, including both the pilot and main studies, took nearly 14 months to complete. The subsections below detail the data collection procedures.

4.4.1 Pilot study

Two interviewees, Alpha and Beta, were selected for the pilot study. Alpha is an experienced secondary school Chinese language teacher. She was also involved in teacher training for several years. Beta taught Chinese language pedagogy and assessment in a higher education institution and is an active community leader promoting reading, writing and appreciation of the literary arts. Letters of invitation in English were emailed to both interviewees to seek their consent for me to conduct the interviews (see Appendix E for the final version of the invitation letter). For my pilot interviews, an interview guide that served as a checklist of topics to be covered was used (see Figure 4b). The interviews, each lasting about one and a half hours, were conducted in Chinese. These interviews comprised questions relating to the four main topics listed in my interview guide.

1. Defining the CL2 reading construct and test-takers in the Singaporean context.
2. The purposes and administrative procedures of the GCE1162 reading examination.
3. The cognitive and contextual parameters of the GCE1162 reading examination.
4. Improvements that could be made to enhance the assessment quality of this paper.

Figure 4b: Interview guide for the pilot study

The quality of an interview depends very much on the interviewer. As cleverly articulated by Kvale and Brinkmann (2009: 166), ‘the interviewer is the key research instrument of an interview inquiry’. An able interviewer must learn to draw out rich and specific answers, to encourage the reserved or inarticulate and to be neutral toward the topic while displaying genuine interest in what the interviewees are saying (Fielding & Thomas, 2008). Pilot interviews provide opportunities for an interviewer to develop these critical interviewing skills. In retrospect, the objectives of my pilot interview should have been more clearly conveyed to the interviewees in the invitation letter. The expected length of the interview should also have been stated. Reassurance that only my supervisors and I would have access to the password-protected digital audio recordings and transcripts of interviews kept on my laptop and hard discs could have been given before interviewees expressed concern about issues of security and anonymity. More care should also have been taken when selecting venues for interviews—the background noise and disruptions could have been avoided in the case of Beta if the interview had been conducted at a place more conducive than a café in a busy shopping mall. In addition, there were instances where leading questions and probes could have easily led to bias. In addition, technical terms such as ‘construct’, ‘validity’ and ‘measurement error’ were sometimes used without checking for understanding.

Despite these oversights, answers from Alpha and Beta were essentially rich, specific and relevant (Kvale and Brinkman, 2009). Attempts were made throughout the interviews to follow up and clarify the meaning of the interviewees’ responses. The interviewees were generally at ease and responses were spontaneous and relevant. The interviews were subsequently transcribed, coded using NVivo10 and translated. Themes that emerged from these two interviews were then integrated into a revised interview schedule along with other key questions and probes. The interview

schedule was then modified and refined with suggestions and feedback from my supervisors at University College London and Harvard Graduate School of Education who themselves are skilled and experienced in interviewing skills. Oversights during my pilot interviews were also promptly rectified before the main interviews were conducted.

4.4.2 Main study

The main study which took place two months after the pilot study involved interviews with 20 stakeholders. Interviewees were selected using maximum variation sampling, a type of purposive sampling technique (Patton, 2015). Unlike probability sampling, the goal of maximum variation sampling is not to select individuals from a population randomly and entirely by chance to create a sample with the intention of making statistical inferences from that sample to the population of interest. Instead, it seeks to capture a wide range of perspectives with regard to the phenomenon being studied. By using this sampling technique, I had hoped to examine the GCE 1162 reading examination from various angles through the lenses of stakeholders involved in areas as varied as coordination, test design, item construction, marking and reviewing. The interviewees included an elite policy group with privileged access to the detailed test specifications and procedures. Interviews were also carried out with secondary school CL2 teachers and students, whose perspectives are seldom considered in validation processes. In addition, opinions were sought from experts in the field of CL2 reading and assessment.

Figure 4c provides more details about the stakeholders interviewed, although due to reasons of confidentiality, their profile and involvement in the GCE 1162 reading examination were made more general in some cases. All names were replaced by letters in the Greek alphabet and actual examples when quoted in the study were anonymised. All interviewees in the main studies received a letter of invitation informing them of the purpose and length of the interview (see Appendix E). Written consent was also sought from the interviewees and parents or guardians of the student interviewees (see Appendix F). Interviews were conducted in Chinese or English, and sometimes in a mixture of both, depending on the language with which the interviewee felt more comfortable.

Name	Gender Male (M)/Female (F)	Age	Profile	Involvement in the GCE 1162 reading examination
Pilot study				
Alpha	F	Early 30s	Secondary school CL2 teacher. Taught in government and independent schools. Involved in teacher training.	Invigilation, marking and preparing test-takers.
Beta	M	Early 40s	Academic. Taught Chinese language pedagogy and assessment at undergraduate level. Community leader promoting reading, writing and appreciation of the literary arts.	Review and research.
Main study				
Gamma	F	Early 50s	Academic and key personnel in teacher training institution. . Involved in the design and execution of IGCSE examinations and teacher training.	Research and reviewing of examination systems.
Delta	M	Early 40s	Curriculum specialist. Taught CL2 in government secondary schools.	Training, invigilation, marking, preparing test-takers and coordinating at school and national level.
Epsilon	M	Early 60s	Master teacher. Taught CL2 in government and independent schools. Involved in teacher training.	Reviewing of examination papers, research, test design, item construction, invigilation, marking, preparing test-takers and coordinating at school and national level.

Zeta	F	Early 30s	Secondary school vice-principal. Taught CL2 in government schools.	Invigilation, marking, preparing test-takers and coordinating at school level.
Eta	F	Mid 30s	Head of Chinese language department in secondary school. Taught CL2 in government schools. Ex-curriculum specialist and teacher trainer.	Invigilation, marking, preparing test-takers and coordinating at school and national level.
Theta	M	Early 60s	Master teacher. Taught CL2 in government and independent schools. Involved in teacher training.	Reviewing of examination papers, research, test design, item construction, invigilation, marking, preparing test-takers and coordinating at school and national level.
Iota	M	Mid 30s	Curriculum specialist. Taught CL2 in government secondary schools.	Training, invigilation, marking, preparing test-takers and coordinating at school and national level.
Kappa	F	Early 40s	Academic. Researching into CL2 pedagogy and assessment. Ex-secondary school CL2 teacher.	Invigilation, marking, preparing test-takers and research.
Lambda	F	Late 30s	Head of Chinese language department in secondary school. Taught CL2 in government schools. Ex-curriculum specialist.	Invigilation, marking, preparing test-takers and coordinating at school and national level.

Mu	M	Early 40s	Academic and member of key personnel in teacher training institution. Taught CL2 in government and independent schools. Involved in teacher training.	Reviewing of examination papers, research, test design, item construction, invigilation, marking, preparing test-takers and coordinating at school and national level.
Nu	F	Mid 40s	Head of Chinese language department in secondary school. Taught CL2 in government and independent schools. Involved in teacher training.	Invigilation, marking, preparing test-takers and coordinating at school and national level.
Xi	M	Mid 60s	Academic and key member of personnel in the MOE. Involved in the design and execution of examinations and teacher training. Taught CL2 in government and independent schools.	Planning, training, reviewing of examination papers, research, test design, item construction, coordinating at school, institutional and national level.
Omicron	F	Early 30s	Secondary school CL2 teacher. Taught in government and independent schools.	Invigilation, marking and preparing test-takers.
Pi	F	Early 40s	Secondary school CL2 teacher. Taught in government and independent schools.	Invigilation, marking and preparing test-takers.
Rho	F	Late 40s	Academic and key member of personnel in teacher training institution. Taught CL2 in government and independent schools. Involved in teacher training.	Reviewing of examination papers, research, test design, item construction, invigilation, marking, preparing test-takers and coordinating at school and ministry level.

Sigma	M	Late 30s	Head of Chinese language department in secondary school. Taught in government and independent schools.	Invigilation, marking, preparing test-takers and coordinating at school level.
Tau	F	16	Express Course student at an independent all-girls secondary school. High CL2 reading proficiency, obtained A1 in the GCE 1162 examination.	Test-taker of the GCE 1162 examination in 2014.
Upsilon	F	16	Express Course student at a co-ed mainstream secondary school. Upper intermediate CL2 reading proficiency, obtained B3 in the GCE 1162 examination.	Test-taker of the GCE 1162 examination in 2015.
Chi	F	16	Express Course student at a co-ed mainstream secondary school. Lower intermediate CL2 reading proficiency, obtained B4 in the GCE 1162 examination.	Test-taker of the GCE 1162 examination in 2015.
Omega	M	17	Normal (Academic) Course student at a co-ed mainstream secondary school. Low CL2 reading proficiency, obtained D7 in the GCE 1162 examination.	Test-taker of the GCE 1162 examination in 2015.

Figure 4c: Biographical data of the 22 interviewees in the pilot and main study

The final interview schedule included opening, body and closing sections (see Appendix G). The body of the interview schedule included below was divided into five segments (Segment A to E). Each segment comprised questions, closely following the four sub research questions set out in Chapter 1, designed to understand the degree to which the intended measurement objectives of the GCE 1162 reading examination had been achieved.

Segment A inquired into the interviewee's work or schooling experiences and involvement in national examinations. The prepared questions were:

Segment A: Work/schooling experience

Questions for interviewees other than students

Q1. 请问您现在在哪里工作？主要负责的项目是什么？

Where are you working? Could you tell me more about your job?

Q2. 可否请您简单叙述一下您过去的工作经验？

Could you describe briefly your previous work experience?

Q3. 请问您曾参与国家级考试的设计或执行工作么？请说明工作内容。

Have you been involved in the design and execution of any national examination? Could you elaborate on the nature of the work involved?

Questions for student interviewees

Q1. 请问你现在在哪里念书？可否简单叙述一下你过去学习华文的经验？

Where are you studying? Could you tell me more about your experience of learning Chinese?

Segment B probed the interviewee's understanding of the CL2 reading construct and how it could be defined in the social-cultural context of Singapore:

Segment B: The reading construct

Q1. 华文作为第二语文的新加坡考生具有什么特点？有哪些特点是我们在设计 1162 试卷时所必须注意的？

Are there any unique characteristics of Singapore's CL2 readers that have to be taken into consideration when designing the GCE 1162 reading examination?

Q2. 您认为我们的学生在完成了中小学教育后须具备什么样的华文阅读能力和思维能力？

In your opinion, what is the level of CL2 reading proficiency a student is expected to attain after completing secondary education? What about a student's cognitive ability?

Segment C focused on the cognitive parameters of the GCE 1162 reading examination. Emphasis was placed on the cognitive demands of the examination and the reading skills it aims to elicit. The prepared questions and probes consisted of:

Segment C: Cognitive parameters of the GCE 1162 reading examination

Q1. 可否请您谈谈您对 1162 阅读试卷的总体印象？ Could you tell me about your general impression of the GCE 1162 reading examination paper?

Q2. 您认为 1162 阅读试卷的考试目标为何？

In your opinion, what are the assessment objectives of the GCE 1162 reading paper?

Q3. 您认为 1162 试卷的成绩有哪些用途？

In your opinion, what are the purposes of the GCE 1162 reading paper?

Q4. 您觉得这份试卷考核了什么样的阅读技能和思维能力？您可以按项目逐步分析（综合填空、阅读理解一选择题、阅读理解二简答题）。

What are the reading skills and cognitive processes involved in answering the GCE 1162 reading paper? You may wish to look at the different sections in turn, namely, multiple-choice cloze, reading comprehension multiple-choice and reading comprehension constructed response.

Q5. 除了阅读技能和思维能力，这份试卷还考核了什么要素？

Besides reading and cognitive skills, does the GCE 1162 reading paper elicit other aspects of learning?

Q6. 考试实际测量的构念与《中学华文课程标准》和《考试纲要》中列出的目标是否契合？为什么？

Does the GCE 1162 reading examination paper measure what it proposes to measure? Why?

Q7. 您在较早前指出新加坡学生在完成中小学教育后所应该具备的华文阅读能力和思维能力。您认为 1162 阅读试卷能否有效地测量这些要素？为什么？

How effective is the examination paper in eliciting the reading and cognitive skills which, as you mentioned earlier, are critical to a student after completing secondary education?

Q8. 您认为 1162 试卷能否有效地区分出读者的优劣？为什么？

Is the GCE 1162 reading examination able to differentiate between the competent and less competent reader? Why?

Reading skills and levels of cognitive processing are mediated by the contextual parameters of the passage and items in hand (Khalifa & Weir, 2009). The purpose of Segment D was therefore to solicit the interviewee's views on the features of the passage and items that influence validity. The prepared questions and probes were as follows:

Segment D: Contextual parameters of GCE 1162 reading examination

Q1. 1162 阅读试卷的题型有三种，即综合填空、选择题和简答题。您认为这些题型是否真实有效？

The GCE 1162 reading examination has three item formats, namely, multiple-choice cloze; reading comprehension multiple-choice and reading comprehension constructed response. What do you think of these item formats? Are they effective and authentic?

Q2. 您可否举出其他考查学生阅读能力和兴趣的题型或方式？

Can you think of other types of items or ways of assessing reading ability and interest?

Q3. 您所列举的这些题型和考核方式可否包括在中学生阅读能力的终结性评估中？倘若可以，应该如何融入？

Can the existing GCE 1162 reading examination include a wider range of item formats and assessment methods? How can they be incorporated?

Q4. 有的老师认为，应该恢复过去阅读试卷中填写汉字和造句的题型。对此您有何看法？

Some teachers recommend the revival of item formats used in older versions of the GCE 1162 reading examination, such as filling in the Chinese character and sentence construction. What is your opinion on this?

Q5. 您认为考评局根据什么标准选择考试篇章？

What do you think the selection criteria for the GCE 1162 reading passages are?

Q6. 您觉得篇章的数目、长度和类型一般是否适合？为什么？

Are the number, length and genre of the passages generally appropriate? Why?

Q7. 您觉得题目的顺序和权重一般是否合适? 为什么?

Are the order and weightage of the items generally appropriate? Why?

Q8. 您认为一个半小时的作答时间是否足够? 为什么?

The duration of the GCE 1162 reading paper is one and a half hours. Do you think the time given is sufficient? Why?

Building on the questions in Segment A to D, interviewees were asked in Segment E to suggest ways of improving the quality of the examination paper and evaluation system.

Segment E: Evaluation

Q1. 您认为可以通过什么方式提升 1162 的质量?

What improvements can be made to enhance the quality of the GCE 1162 reading examination paper?

Q2. 您认为维持或提高一套试题的质量是否需要持续性的监督与审查? 在这个过程中, 考评局、教育部、教育学院、学校和家长又能扮演什么样的角色?

Are ongoing evaluation and validation needed to ensure the quality of an examination? What roles can the different actors, e.g. the Singapore Examinations and Assessment Board (SEAB), Ministry of Education (MOE), National Institute of Education, schools and parents play in these evaluation and validation processes?

To ensure quality inferences from the interviews during the analysis stage, I practised reflexivity to minimize my possible influence on the findings and interpretations. I was aware that many of the interviewees were apprehensive at first because of my role as an MOE officer, and doubly so as they are public servants and students. Exactly who the interviewees perceive themselves to be talking to, and why, will naturally affect what they say; hence, being interviewed by someone from the same professional circle might prove inhibiting (Dowling & Brown, 2010). To avoid

interviewees offering politically correct answers, I tried to downplay my role as an MOE officer. I went to considerable lengths to explain the nature of my study and my role as a student researcher. Invitation letters were printed on University College London instead of MOE letterhead paper (see Appendices E and F) and interviews were carried out in informal settings such as cafés and benches outside libraries to make interviewees feel at ease.

At the beginning of the main body of the interview questions, I provided adult interviewees with initial stimuli with which they could easily engage, such as getting them to share an anecdote about assessment and testing. With student interviewees, questions were paraphrased to avoid jargon and examples from the GCE 1162 reading examination papers were provided to stimulate their thinking. If interviewees agreed with an official statement or observation that I made, I often asked them to elaborate or provide a supporting example when appropriate. No matter how radical an interviewee's perspective, I tried to remain neutral and resisted presenting their responses using politically correct terms. Clarification was always provided when interviewees experienced difficulties or appeared to have misinterpreted a question. Further, samples of the GCE 1162 reading examination paper, Syllabus 2011 and the official *GCE 1162 Examination Information Booklet* were always kept at hand during the interview in case interviewees needed to refer to them. It is worth noting that one of the interviewees, Xi, followed up post interview via email with elaboration on his comments. The email was added to his interview transcript, bearing in mind that there had been time for after-the-event rationalization and reflection.

4.4.3 Data analysis

The resultant dataset is a rich and informative collection of stakeholders' views and opinions about the GCE 1162 reading examination, with 37 hours of recording and more than 20,000 lines of transcription (see Appendix H for excerpts). Analysis of the interview data comprised four stages where I sought to transform, evaluate, refine and synthesize the data. Although the four stages are described in a linear sequential manner below, it was, in practice, an iterative process with constant revisiting of data within and across the four stages described below.

The first stage is the stage of transformation. This stage began with re-reading notes taken during the interviews and listening back through the recordings. The data were then transcribed and entered into the qualitative software package Nvivo10. Interview data collected during the pilot study were first open coded. After working through the entire transcripts for Alpha and Beta, notes and comments were grouped and compared with my research questions to form a provisional list of codes, which is an ‘organising system of entering the text and identifying units of interest for further analysis and interpretation’ (Miller & Crabtree, 1999: 135). The initial list of 49 provisional codes was constantly updated during stages two and three.

The second stage is the stage of evaluation. Transcripts were read and tagged using NVivo10, guided by the provisional list of codes (or ‘nodes’ in NVivo10). In stage one, the focus was primarily to derive codes. In stage two, however, the focus had changed to applying existing codes to the transcripts to ascertain that the provisional list of codes was a good reflection of the recurring regularities in the interview data. As more transcripts were evaluated, wider and deeper insights were gained and the provisional list of codes was amended accordingly. Pilot and main study transcripts were revisited to include new codes and to delete old ones. The evaluation process thus comprised several repetitive cycles.

The third stage is the stage of refinement. As the codes were progressively reviewed and refined, the coding process at stage three became increasingly deductive. The final 27 codes were subsumed under five categories, namely, specifications and administration, test-taker characteristics, cognitive parameters, contextual parameters and a posteriori inferences (see Figure 4d).

Code		Category
▪ Purposes	(S-PUR)	Specifications and administration (SPAD)
▪ Constructs	(S-CON)	
▪ Authenticity and packaging	(S-AUTPAC)	
▪ Transparency	(S-TRANS)	
▪ Aligning with new curriculum	(S-ALIGN)	
▪ Administrative procedures	(S-ADMIN)	
▪ Suggestions	(S-SUGG)	
▪ Proficiency	(T-PROF)	Test-taker characteristics (TAKER)
▪ Motivation	(T-MOTIV)	
▪ Difficulties	(T-DIFF)	
▪ Experience	(T-EXP)	
▪ Suggestions	(T-SUGG)	
▪ Overall impression	(C-OVER)	Cognitive parameters (COGNI)
▪ Reading dimensions	(C-DIME)	
▪ Reading level (local/global)	(C-LEVEL)	
▪ Reading type (careful/expeditious)	(C-TYPE)	
▪ Cognitive levels	(C-COG)	
▪ Differentiating readers	(C-READER)	
▪ Suggestions	(C-SUGG)	
▪ Multiple-choice gap-filling test	(X-GAP)	Contextual parameters (CONTEXT)
▪ Multiple-choice items	(X-MCQ)	
▪ Constructed response items	(X-CON)	
▪ Passages	(X-PASS)	
▪ Suggestions	(X-SUGG)	
▪ Scoring	(P-SCORE)	A posteriori inferences (POST)
▪ Washback effects	(P-WASH)	
▪ Suggestions	(S-SUGG)	

Figure 4d: Final list of categories and codes for tagging qualitative data

The categories were constructed to meet the following criteria. First, categories should be responsive to the purpose of the study. Second, categories should be exhaustive, that is, all data considered relevant could be placed under one of the categories. Third, categories should be conceptually congruent, in other words, all categories should be at the same level of abstraction and fit together to answer the research questions. The number of categories was kept small in view of Creswell's (2007) recommendation to have only five or six final categories. A large number of categories is not only difficult to manage but also likely to reflect an analysis reliant more on description than critical thinking.

The fourth stage is the stage of synthesis. Sections of transcripts were re-coded and re-categorized to check for dependability. As Nvivo10 was used to manage the data, it was relatively easy to filter and retrieve segments of transcripts by codes or categories. The data were subsequently triangulated with those derived from other research methods to minimize the inadequacies associated with semi-structured interviews. Segments extracted and placed under each code and category were studied and translated into English, if necessary, to be quoted in the study. Care had been taken in the selection of quotes to ensure they were representative of the findings obtained from the interviews. Unless otherwise stated, selected quotes are not contradicted by, or contradictory to, the general position held by all interviewees. Nevertheless, it is impossible for the quotes to illustrate the views of all interviewees and they are, of course, part of my own construction of the validity argument in the study.

4.5 Research methods: Document analysis and expert judgement

A document may be defined briefly as a record of an event or progress (Cohen, Manion & Morrison, 2011). It is a form of crafted communication—a visual, graphic or electronic representation of language and objects (Freebody, 2003). Documents may be produced by individuals or groups and can take many different forms, such as school textbooks, corporate reports, report cards, historical archives, government websites, photographs and drawings. Documents are commonly classified as either public or personal documents. The former being official records of a society's

activities, for instance, association manuals and newspaper articles, and the latter being any first-person narrative that describes an individual's experiences and beliefs, for instance, diaries and travel blogs. Prior (2003) contends that documents form a field of research in their own right and document analysis 'consists of selecting, as opposed to generating, documents [...] and analysing their contents' (Guest, Namey & Mitchell, 2013: 252).

There are many reasons why documents are a good source of data. First, many documents are free, easily accessible and contain information that would take a researcher a considerable length of time and energy to gather otherwise. For example, *The Mother Tongue Languages Review Committee report: Nurturing active learners and proficient users* (MOE, 2011), readily accessible at public libraries, presents findings based not only on an extensive survey of 22,000 teachers, students and parents but also on study trips to several countries to observe the latest developments in teaching, learning and testing of languages. Such information would be impossible to acquire through the efforts of an independent researcher. Second, documents are stable, 'non-reactive' sources of data; in other words, the researcher's presence does not affect what is being studied, unlike methods such as interviewing and observation. Third, documents can contain information that interviewees are unaware of or have forgotten. With the Internet and search engines, the uncovering and retrieval of documents relevant to a research study becomes much simpler. Last, documents are versatile, for example, they can furnish descriptive and statistical information, provide background details, offer historical understanding and track change and development.

As with other research methods, there are concerns to keep in mind when using document analysis. An initial consideration is that most documents are not produced for the research in hand. The documents may therefore contain information that is incomplete or partially useful. For instance, although there are many newspaper articles on the national examinations in Singapore, the vast majority centres upon examination stress among young students, which is only tangentially related to the scope of my study. Another major problem with documents is determining their authenticity, accuracy and objectivity. With the public documents gleaned from official sources in this study, authenticity may be less of an issue; however, it was

still necessary to question their accuracy and objectivity. A further challenge of using document analysis in this study was the wealth of documentary materials that, unfortunately, I could not access. Detailed test specifications, mark schemes, markers' reports, guidelines for item setters and test-takers' answer scripts are examples of restricted and confidential documents that could have provided invaluable insights in a validation study of the GCE 1162 reading examination.

4.5.1 Pilot study

Finding relevant, extant and accessible documents was the first step in my pilot study. Data collection was a systematic process guided by my topic of inquiry and research questions. An evaluation of the GCE 1162 reading examination would logically lead a researcher to track down documents through the MOE, SEAB and University of Cambridge Local Examinations Syndicate's websites. Besides the MOE press release webpage, the government press release webpage of Singapore provides a repertoire of past speeches relating to Chinese language education in general and Chinese language assessment and testing in particular. I also had privileged access to the MOE's intranet and daily news briefs which highlight local as well as international reports on Singapore's education and provide links to relevant news clippings. Documents were also located at the National Institute of Education Library and the Resources for Education and Development Library at the Academy of Singapore Teachers. I kept an open mind when it came to discovering useful documents, for instance, I asked my interviewees if they had any documents to suggest. This process led to the serendipitous discovery of the National Library Board's BookSG digital collection of more than 2,500 copies of Chinese textbooks spanning the period from the 1920s till now.

Once the public documents had been located through the tracking strategies stated above, their objectivity and accuracy had to be established. Who is/are the author(s)? What is the author trying to accomplish? Who is the intended audience for the document? What are the author's sources of information? What is the author's possible bias? To what extent is the author likely to want to report the truth? Is the document complete or have parts of it been censored? It should be noted that public documents, especially policy texts, often 'seek to persuade their readership of the

truthfulness and credibility of the arguments which they are deploying [...] by suggesting that there is only one way of representing the world and this way resonates with common sense views of representation' (Scott, 2000: 26). Public documents, however, generally represent one outlook or ideology and it cannot be assumed, therefore, that such documents reflect consensus or concerns on assessment and educational practices. Ideally, document analysis is employed in conjunction with other research methods, such as interviewing, or substantiated by judgements from a panel of SMEs.

The past GCE 1162 reading examination papers are a set of documents that this study investigated in detail through a panel of SMEs. SEAB distributes past GCE O-Level papers through authorized publishers, and compilations known colloquially as the *Ten Year Series* are readily available in print form. Original official papers rather than commercial practice tests were used in the study as they better reflect the overall nature and characteristics of the GCE 1162 reading examination. There were 78 sets of examination papers available when sampling started, given that the paper was first implemented following the 1978 education reforms in Singapore. The examination is offered twice a year, in May/June and October/November, thus two sets of paper are available for each given year.

In the pilot study, a secondary school Chinese language teacher and I reviewed a specimen GCE 1162 reading examination paper released by SEAB in 2014¹ (see Appendix A). Both of us had taught Chinese as a second language in Singapore for more than five years, and as such, we were qualified to provide a preliminary scan of the passages and items in the specimen paper. An Excel evaluation spreadsheet (see Appendix I) was duly designed based on our two 90 minute discussions (see Figure 4e). The specimen paper was then systematically re-evaluated using the Excel evaluation spreadsheet. Of the 148 components evaluated,² the secondary school Chinese language teacher and I concurred 124 times. Hence, SME inter-rater reliability for the pilot study was $(124 \div 148) \times 100\% = 83.78\%$.

¹ The specimen GCE 1162 reading examination paper released by SEAB in 2014 relates to the new reading examination format (GCE 1160) which came into use in 2016.

² The 148 components evaluated for the specimen paper is based on the eight parameters listed in Figure 4e. There are seven passages and 30 items in total for the specimen paper, therefore $(7 \text{ passages} \times 4 \text{ parameters}) + (30 \text{ items} \times 4 \text{ parameters}) = 148 \text{ components}$.

Passage		
1.	Summary	
2.	Literary merit	0 (not applicable or low)
		1 (moderate)
		2 (high)
3.	Discourse mode	Narrative
		Expository
		Argumentative
		Functional
		Others
4.	Propositional content	Values and attitudes
		Traditions and festivals
		Local news and culture
		Global awareness
		Aesthetic appreciation
		Advertisements and lifestyle
		Others
Item		
5.	Cognitive demand	Lower-order thinking
		Higher-order thinking
6.	Specific cognitive demand	Remember
		Understand
		Apply
		Analyse
		Evaluate
		Create
7.	Reading approach (reading level)	Local
		Global
8.	Reading approach (reading type)	Expeditious
		Careful

Figure 4e: The eight parameters used for evaluating the 22 sets of GCE 1162 reading examination papers

4.5.2 Main study

The main study proceeded to narrow the selection of documents to be examined. Documents were chosen from those collected based on the following four criteria. First, documents about the construction and evaluation of the GCE 1162 reading examination were chosen which included those that state the purposes of this examination, the reading and cognitive skills it assesses, how the examination aligns with the curriculum and the processes in place to ensure its quality and relevance. Second, documents that record the development and reforms of the secondary Chinese language reading curriculum and assessment were selected. Third, documents about recent MOE initiatives and findings, in particular, those pertaining to trends in language use, assessment and testing, literacy and 21st century skills were chosen. Fourth, documents were selected that assist in determining the interpretations derived from GCE 1162 reading examination scores, such as publicly available data on the technical and statistical aspects of the examination, the uses of test scores and the consequential effects of such high-stake assessment.

The final selection included the GCE O-Level Chinese examination information booklets, secondary Chinese language syllabuses, recent Chinese language and mother tongue languages reviews, SEAB annual reports, SEAB presentations, SEAB's instructions to test-takers, press releases and speeches. These public documents, produced mainly by the MOE and SEAB are a significant source of validity evidence. Further, 22 GCE 1162 reading examination papers, comprising 142 passages and 660 items, administered in the last decade (2006-2016) were scrutinized by a panel of SMEs. The number of passages and items were calculated (see Figure 4f). A slight change in the number of passages per examination from the year 2012 onwards was noted. I did not randomly choose passages and items from the 78 available papers since 1978. Only passages from the latest examination papers, between 2006 and 2016, were selected as this study is primarily an inquiry into the measurement objectives of the GCE 1162 reading examination in the *present* context. Since this study is not intended to be a detailed documentation of the progress and development of the GCE 1162 reading examination since its inception, the selection of more recent papers is justified.

Year	Number of examination papers	Number of passages	Number of items per examination		
			Multiple-choice gap-filling test	Multiple-choice for passages	Constructed response for passages
2006-2011	12	6 passages x 12 papers = 72	10	10	10
2012-2015	8	7 passages x 8 papers = 56	10	10	10
2016	2*	7 passages x 2 papers = 14	10	10	10

Total number of examination papers: 12 + 8 + 2 = 22	Total number of passages: 72 + 56 + 14 = 142	Total number of items: 30 x 22 = 660
--	---	---

Figure 4f: The number of official GCE 1162 reading examination papers, passages and items available from 2006-2016

* Includes the SEAB specimen paper used in the pilot study

The GCE 1162 reading examination is administered in standard simplified Chinese and held twice yearly (May/June and October/November) in secondary schools and various examination centres. The reading examination forms Paper 2 of the entire GCE 1162 examination, with Paper 1 being writing and Paper 3 being listening and oral communication. The time allocated for the reading examination is 1 hour and 30 minutes. The paper consists of three sections (Section 1, multiple-choice gap-filling; Section 2, passages with multiple-choice items; and Section 3, passages with constructed-response items) with 30 items in total, accounting for 70 marks (carrying a weightage of 35% of the entire examination). Amendments were made to the reading examination in 2006, 2012 and 2016 as reflected in Figure 4g.

	Subject code	Marks/ Weightage	Sections		
Previous examination format (before May 2006)	GCE 1162	80/40%	<u>Section 1</u> Fill in the blank with the correct Chinese character. <i>Hanyu pinyin</i> is provided. 5 items, 10 marks. <u>Section 2</u> Grammar (function words). 5 items, 5 marks. <u>Section 3</u> Sentence construction. 5 items, 15 marks. <u>Section 4</u> Multiple-choice gap-filling; short passage with selected words removed. 10 items, 10 marks.	<u>Section 5</u> 1 passage with multiple-choice items. 5 items, 10 marks.	<u>Section 6</u> 1 passage with constructed-response items. 6 items, 30 marks.

Old examination format (May 2006- November 2011)	GCE 1162	70/35%	<u>Section 1</u> Multiple-choice gap-filling; short passage with selected words removed. 10 items, 10 marks.	<u>Section 2</u> 3 to 4 short passages with multiple-choice items. 10 items, 20 marks.	<u>Section 3</u> 2 to 3 short passages with constructed-response items. 10 items, 40 marks.
New examination format (May 2012- November 2015)				<u>Section 2</u> 3 to 4 functional passages (e.g. advertisements, flyers and newspaper articles) or short passages with multiple- choice items. 10 items, 20 marks.	
Latest examination format (May 2016 onwards)	GCE 1160				

Figure 4g: GCE 1162 (and GCE 1160) reading examination formats (emphases added)

Three SMEs were subsequently invited to form a panel of four with me to appraise the passages and items. The four of us had between us nearly fifty years of experience teaching Chinese as a second language to secondary students in Singapore. There was a half day briefing and training session to increase consistency, reduce review time and to discuss ambiguities and special situations. A letter of invitation stating the objectives of the investigation was given out before the training (see Appendix E). SMEs were then grouped into pairs and each pair was given 10 to 11 sets of examination papers to peruse. They first assessed the passages and items independently and responses and feedback were keyed into the accompanying Excel evaluation spreadsheet (see Appendix I).

There were eight parameters to be evaluated (see Figure 4e). First, SMEs were asked to identify the discourse mode, namely narrative, expository, argumentative and functional, in accordance with the Secondary Chinese Language Syllabus 2011 (Syllabus 2011). Next, SMEs summarized the propositional content of the passages and classified them under one of the six themes listed which were derived from the thematic concerns highlighted in Syllabus 2011. Subsequently, they gauged the literary merit of the passages, with 0 being not applicable or of low literary calibre, 1 being moderate calibre and 2 being high calibre. The cognitive demands of the items were then analysed. SMEs had to determine whether an item was a lower-order or higher-order thinking item based on the 2001 revised Bloom's taxonomy (Anderson & Krathwohl, 2001) and to ascribe a specific cognitive level to the items, namely, remember, understand, apply, analyse, evaluate or create. Zhu's (2015) helpful descriptors and examples specific to the cognitive levels in reading comprehension (Figure 2e) were also given to SMEs to facilitate evaluation. SMEs then had to indicate the level of reading (local or global) and the type of reading (expeditious or careful) that an item corresponded to. These two parameters are based on Weir's extensive research on reading assessment (Khalifa & Weir, 2009).

The review process took approximately two months and was completed by SMEs mainly at home. The SMEs conferred with their partner and attempted to reach an agreement when there were differences in judgement. In cases of conflicting responses, the passage or item was flagged and raised to the other pair of SMEs for resolution. To check for inter-SME reliability, all four SMEs evaluated an identical set of three passages and corresponding 13

items. Of the 64 components to be analysed,³ the SMEs concurred on 58 components. Hence, it may be said that SME inter-rater reliability for the main study was $(58 \div 64) \times 100\% = 90.63\%$. Furthermore, each SME re-evaluated the set of three passages and 13 items a month after the first assessment. An average of 60 components were given identical evaluation in the repeated assessment. Hence, it may be said that the intra-SME reliability was $(60 \div 64) \times 100\% = 93.75\%$. Based on the responses from the SMEs, I was able to garner data on the cognitive and contextual design of the GCE 1162 reading examination. The relevance of the passages and items in relation to the intended domain and congruence to the test specifications could also be better understood from the data.

4.5.3 Data analysis

Once the document sample had been decided, the selected documents were inventoried and organized using NVivo10 for easy retrieval and manipulation. The documents were subsequently coded using pre-existing categories generated from the interview data (see Figure 4d). The main content and messages under each category were then integrated and compared with and linked to data collected through other methods.

With regard to the 22 sets of GCE 1162 reading examination papers, a quantitative mode of analysis and presentation was selected to supplement the fundamentally qualitative investigation. First, the mean literary value of the 142 passages was calculated. Next, the distribution of items in terms of cognitive demands, specific cognitive demands and reading approaches were tabulated and charted to provide an overview of the items as a whole. Frequencies of the occurrence for each type of item as well as percentages were given for comparison. Similarly, quantitative summaries of the frequencies and percentages of passages of different discourse modes and propositional content were provided. Further, the 22 sets of reading examination papers were grouped into three, namely, the old (May 2006-November 2011), new (May 2012-November 2015) and latest (May 2016 onwards) (see Figure 4g), corresponding to the three changes in examination format in 2006, 2012 and 2016. For parameters three to eight (see Figure 4e), the Chi-square test of independence and P value were then used to determine whether differences in the distribution of items and passages

³ The 64 components to be evaluated for the specimen paper were based on the eight parameters listed in Figure 4e. There were three passages and 13 items in total, hence $(3 \text{ passages} \times 4 \text{ parameters}) + (13 \text{ items} \times 4 \text{ parameters}) = 64 \text{ components}$.

across the three groups were statistically significant. Quantification enabled the coverage of a larger amount of data than was possible using only the elaborated description of qualitative analysis. The next step in planning the study was to ensure ethical compliance. Ethical concerns are considered in the following section.

4.6 Ethical concerns

This study abided by the *British Educational Research Association Ethical Guidelines for Educational Research 2011*. All necessary measures were taken to ensure that my study was conducted in an ethically defensible manner, with utmost regard for the person, knowledge, democratic values, the quality of educational research and academic freedom (British Educational Research Association, 2011). Sensitive areas in my study emanated primarily from the participation of public servants and secondary school students. Care had to be taken from the outset to respect the rights, autonomy and dignity of these individuals. Insensitive storage and handling of data and later dissemination of the findings may inconvenience or harm them.

Approval to interview students, teachers and specialists had to be sought from the MOE before I could embark on data collection. The lengthy four-month approval process required the submission of my detailed research proposal, methodology chapter and timeline for clearance. When approval was granted, a formal letter of invitation was emailed to potential interviewees. They were informed of the purpose of the research, the role they would engage in (either as an interviewee or SME), how the data would be used and to whom the research outcomes would be reported. There was never any intention on my part to engage in secret or covert research. Participation in this study was entirely voluntary and interviewees had the right to withdraw from the research at any time, for any or no reason, without negative consequences.

Interviewees were then requested to sign a consent form and to provide essential demographic information. Although student interviewees are ‘capable of forming their own views [and therefore] should be granted the right to express their views freely in all matters affecting them, commensurate with their age and maturity’ (British Educational Research Association, 2011: 6), in accordance with MOE requirements, however, consent was also sought from the student interviewees’ parents or guardians as they were below the age of 21.

To minimize the impact of my research on the workloads of the interviewees and to put them at their ease, all interviews were conducted at a time and place of their convenience. All interviewees were given a small token of appreciation. They were, however, not told before the study was completed, hence, their decision to participate would not be influenced by the reward. To protect interviewees from external scrutiny, they were each associated with an arbitrary name, specifically a letter from the Greek alphabet, and the same name was used in my transcripts and study. The interviewees' individual identity will remain strictly confidential. All records and feedback are securely stored on my laptop and hard discs and the data will only be used for academic purposes.

4.7 Conclusion

This chapter has explicated and justified the use of pragmatism, the mixed methods research design and the methods of semi-structured interview, document analysis and expert judgement with the embedded quantitative analysis in the present study. Mindful of the pitfalls revealed in the literature, the study relied on a combination of qualitative and quantitative methodologies and methods. Data collected from multiple sources were triangulated to test for the consistency of findings and to discover divergent perspectives. To enhance the defensibility of the findings, steps were taken to strengthen the quality of the research.

As seen earlier, the mixed methods design was underpinned by a qualitative methodology and allied research methods complemented by quantitative methods. The quality assurance of a fundamentally qualitative study is concerned with the soundness or trustworthiness of findings emerging from the study, warranted by its methodology (Mertens, 2010). Put differently, a qualitative study must be evidenced by documentation that shows how the research was carried out, what methods were used and how the data collected were analysed and interpreted. Guba and Lincoln (1989) propose four criteria for ensuring the rigour of qualitative investigation, specifically, credibility, transferability, dependability and confirmability.⁴ The following section considers briefly each of these four qualitative

⁴ The four criteria of credibility, transferability, dependability and confirmability are used here in preference to indicators often considered in quantitative studies, namely, internal validity, generalizability, reliability and objectivity.

research trustworthiness criteria and discusses how provisions were made during the research to meet these criteria.

The first criterion to be addressed is credibility which may be defined as the congruence between research findings and reality (Lincoln & Guba, 1985). In this study, credibility was validated if the research findings represented plausible extrapolation from interviews and document analysis. The following strategies were implemented to ensure credibility. The first strategy implemented was prolonged engagement in the field of research. I had taught CL2 to secondary students for five and a half years and had participated in the administration of the GCE 1162 reading examination throughout this period. As such, I had adequate familiarity with the examination as well as the working culture of Singapore's secondary schools, the MOE and SEAB. In addition, I had known many of the adult interviewees and all the student interviewees on a professional basis before the study was conducted and there were, thus, established relationships of rapport and trust. A further strategy ensured that that data and interpretations were continuously tested as they were derived from interviewees and documents. To verify whether the interviewees' opinions had been accurately captured, main points were recounted at the end of each interview. Evaluation of the passages and items in the GCE 1162 reading examination papers were counterchecked by other SMEs. A third strategy was to practise peer debriefing. During the research process, I actively sought support from my supervisors and peers to 'test [my] growing insights and to expose [myself] to searching questions' (Guba, 1981: 85). I had presented my research at regular laboratory meetings at the Harvard Graduate School of Education, sharing sessions and international conferences and the feedback received was channelled into improving the quality of my inquiry. The fourth strategy used in this study was triangulation. Data were collected from different methods, namely interviewing, document analysis and expert judgement to compensate for their individual limitations and to exploit their respective strengths. Data were then synthesized to obtain corroborating evidence. Together, these four strategies minimized researcher bias and bolstered the credibility of the study.

The second criterion is transferability. Transferability refers to the extent to which the analyses and conclusions of qualitative research can be transferred to other contexts with different interviewees and documents. To facilitate transferability, I provided thick descriptions of the theoretical basis, framework, context and methodology underpinning the study. Whilst the burden of transferability lies with the reader (Mertens, 2010), it would be

impossible for them to gain a proper understanding of the phenomenon under investigation, let alone apply or generalize the findings, without the presence of thick descriptions (Shenton, 2004). Purposeful sampling was also practised as opposed to random sampling, whereby the selection of interviewees and documents was based on specific purposes associated with answering the research questions. The resultant study provides a baseline understanding with which subsequent validity investigations for other CL2 reading examinations can be compared.

The third criterion is dependability, which is concerned with the stability of findings over time. To enhance the dependability of the research, I took the following three steps. First, different methods were used in tandem for the triangulation of data. Second, parts of interview transcripts and selected documents were re-coded and re-categorized after a gestation period of two weeks. Results were compared with previous coding and categorizing outcomes and differences were addressed. When the GCE 1162 reading examination papers were analysed quantitatively, checks for inter- and intra-SME reliability were also conducted. Third, raw data such as documents and interview recordings, transcripts and notes were kept and thoroughly audited at the end of the research. Excerpts of these raw data can be found in Appendix H for easy examination by an external party.

The fourth criterion is confirmability. Confirmability in qualitative research is essentially the researcher's commitment to an inquiry free of bias and prejudice (Guba & Lincoln, 1989). Although absolute objectivity is unattainable as 'all types of research involve selective and thus value-laden interventions of different types during their conduct' (Scott, 1997: 155), I had taken measures to promote confirmability, including practising reflexivity, triangulating data and providing thick descriptions. Further, the limitations of the research together with a reflective commentary on executing a validation study in the Singaporean context are documented in Chapter 9.

In sum, I utilized the Guba and Lincoln (1989) qualitative research trustworthiness criteria discussed to safeguard the quality of my research. The ensuing chapters present the findings which are used to address the research and sub research questions set out in Chapter 1 which pertain to the degree to which the intended measurement objectives of the GCE 1162 reading examination have been achieved.

Chapter 5 Specifications and administration

5.1 Introduction

The primary research question guiding this study interrogates the degree to which the intended measurement objectives of reading in the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) have been achieved. To facilitate the presentation of findings, an adaptation of Weir's (2005) socio-cognitive validity framework is used to structure information (see Chapter 1). The four a priori components of specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters are identified as the main inferences. Each inference with its corresponding validation question, claim and assumptions form an interpretive argument (IA) in accordance with Kane's (2009, 2006) argument-based approach to validation. The interpretive argument is linked to a validity argument (VA) where evidence which supports or counters each claim is investigated. The credibility of the VA hinges on the plausibility of the assumptions that underlie each claim. Hence, the more compelling the supporting evidence or the weaker the counterevidence, the stronger the VA.

Chapters 5 to 8 explore the four main inferences and their accompanying claims in turn, outline the IA, and substantiate and challenge the VA with data collected using the methodology presented in Chapter 4. At the end of each chapter, the assumptions are revisited and careful conclusions are drawn as to whether the assumptions should be accepted, rejected or not investigated (Shaw & Crisp, 2012). The closing chapter, Chapter 9 supplements these findings by offering a brief discussion of a posteriori validation components before providing an overall evaluation of the GCE 1162 reading examination.

5.2 Interpretive argument

The first inference to be examined relates to the specifications and administration of the GCE 1162 reading examination. Expressed as a validation question, the inference probes whether the intended purposes, constructs and administrative procedures of the GCE 1162 reading examination are clearly and sufficiently articulated. The

associated claim is, therefore, that the intended purposes, constructs and administrative procedures of the examination are clearly and sufficiently articulated. For the claim to be justified, the following 12 assumptions have to hold true:

1. The purposes of the examination are indicated.
2. It is possible to identify the primary purpose(s) when there is a multiplicity of purposes.
3. Purposes attributed to the examination are achievable and non-conflicting.
4. Purposes for which the results are unfit are indicated.
5. The constructs of the examination are indicated.
6. Detailed explanations of what the constructs entail are given.
7. The constructs reflect a general consensus of the views of experts in relevant fields with specific consideration of Singapore's context.
8. The constructs align with the recommendations and learning outcomes of the broader curriculum.
9. Security procedures are in place to ensure confidentiality and fairness.
10. Feedback channels are available.
11. Administrative procedures are documented and accessible for public scrutiny.
12. Intra and cross organizational collegiality and research are promoted.

The 12 assumptions can be further grouped under the headings of test purpose, construct and administrative structure. Each assumption prompts consideration of the pertinent validity evidence, which can be both supporting evidence and rebuttals. Rebuttals are threats to validity and they form the basis for recommending changes to the specifications and administration of the GCE 1162 reading examination. The following section sets out the VA by inspecting both the supporting evidence and rebuttals in relation to the assumptions.

5.3 Validity argument

5.3.1 Purpose

Large-scale national examinations often encompass multiple purposes. Over time, these high-stakes assessments have evolved as forms of social contract mutually

agreed upon by stakeholders. As such, national examinations may be used to determine whether students have achieved the stipulated learning outcomes and attained the standard required for the award to which they lead. They may also be administered for selection and placement purposes, where students are systematically ranked or grouped and important decisions made such as university admissions. In addition, information generated by these examinations may be utilized to send signals about the successes and failures of schools and the education system. If validity is ‘the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests’, as defined by the American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA & NCME, 2014: 11), it is paramount that policy makers and test designers first understand the proposed uses or *purposes* of these examinations. As articulated by Stobart (2009: 166), ‘Any validity argument begins and ends with purposes’. Validation studies need to inquire into the purpose or purposes of the assessment in question and consider potential threats to validity, such as lack of clarity, competing purposes and unachievable purposes (Stobart, 2009).

The official *GCE 1162 Examination Information Booklet* issued by the Singapore Examinations and Assessment Board (SEAB) in 2014 states that the examination is ‘constructed based on curriculum objectives and content listed in the *Secondary Chinese Language Syllabus 2011* (Syllabus 2011) [...] with the purpose of assessing the Chinese language competencies of students’ (SEAB, 2014a: 2). The language competencies assessed include listening, speaking (spoken interaction), vocabulary knowledge and language application, reading comprehension, writing (email and different text types) and integrative language skills. The booklet then highlights the learning objectives in Syllabus 2011 stated below (SEAB, 2014a: 2, emphases added), namely that students will be able to:

1. Listen to and understand narrative, expository, argumentative and functional materials.
2. Present opinions and convey emotions on more complex issues and engage effectively in conversations.

3. *Read narrative, expository, argumentative and functional texts of appropriate difficulty and appreciate literary texts.*
4. Write narrative, expository, argumentative and functional texts of appropriate standard and produce simple literary texts.

Taken as it stands, without further elaboration, the *GCE 1162 Examination Information Booklet* is unsatisfactory—it leaves too many unanswered questions and loose ends. First, the GCE 1162 Chinese as a second language examination, one paper of which is the GCE 1162 reading examination, is a mandatory examination for the vast majority of Chinese secondary students. The grade awarded to examinees of the GCE 1162 usually counts towards the aggregate score used for admission to post-secondary institutions. Students are scored based on their performance relative to that of the cohort and are awarded a grade from A1 to F9.¹ For admission to junior colleges, students must satisfy the minimum requirement of an L1R5 aggregate score (first language and five relevant subjects) of 20 or less,² including a pass in English language, at least grade D7 in mathematics or additional mathematics and no less than minimum proficiency in Mother Tongue. For Chinese students sitting the O-Level, grade D7 in the GCE 1162 is required (or E8 in Higher Chinese), although since 2004 students who face exceptional difficulties in learning the Chinese language can be exempted from sitting the GCE 1162 following the introduction of the Chinese language ‘B’ syllabus. Admission to polytechnics with GCE O-Level results is based on ELR2B2 (English language, two relevant subjects and two other best subjects). The GCE1162 is listed as one of the relevant subjects and is often included in the computation of net points for polytechnic admission.

Given the importance of the association between the points system and national examinations it is surprising that tertiary education selection and placement processes are not clearly communicated in the *GCE 1162 Examination Information Booklet*. Selection and placement, often perceived as two of the key purposes of

¹ The grades for GCE O-Level examination subjects are, in descending order, A1, A2, B3, B4, C5, C6, D7, E8 and F9. Grade C6 or above is considered an O-Level pass.

² The L1R5 aggregate score is the combined score of six subjects tested at O-Level. Each grade has a respective point value, for example, grade A1 attracts one point and C6 six points. Hence, if a test-taker scores A1 in all six subjects, their L1R5 aggregate score will be $1 \times 6 = 6$; whereas the aggregate score of a test-taker who obtains C6 in all six subjects is $6 \times 6 = 36$.

high-stakes national examinations, are intricately linked to the exercise of power. Foucault in *Discipline and Punish: The Birth of the Prison* (2012) contends that the examination is an invisible mechanism used to distribute and disseminate power. For Foucault, examination is what separates the ‘juridical’ state from the modern state. High-stakes assessment is in itself one such form of examination. By means of selection and placement, high-stakes assessment ‘establishes over individuals a visibility through which one differentiates them and judges them’ making it possible ‘to qualify, to classify and to punish’ (Foucault, 2012: 174). High-stakes assessment introduces ‘individuality into the field of documentation’ (Foucault, 2012: 178) and constructs ‘each individual as a “case”’ (Foucault, 2012: 181). Consequently, high-stakes assessment is an institutional process that invariably transforms test-takers into objects, scientifically arranging and documenting them ‘through a modality of power where difference becomes the most relevant factor’ (Scott, 2011: 163).

Further, a synthesis of the data collected from semi-structured interview and document analysis suggests that a multiplicity of purposes is attributed to the GCE 1162 in general and its reading examination in particular by the different stakeholders rather than the single aim expressed in the *GCE 1162 Examination Information Booklet*. These purposes may be grouped into five categories, namely, political symbolism, student selection and placement, informing comparisons among educational approaches, educational management and improving instructional guidance and student learning (see Figure 5a).

Purpose	
1.	Political symbolism
1.1	To serve as a symbolic action to convey the message that the bilingual language policy remains the cornerstone of Singapore's education system.
1.2	To shape public perceptions about Chinese language and culture.
1.3	To appease the Chinese educated community.
1.4	To ensure that students remain motivated to learn the Chinese language in school.
2.	Student selection and placement
2.1	To provide a summative assessment of students' Chinese language learning achievement in secondary school.
2.2	To assess a student's ability to read and understand materials such as advertisements, blogs and newspaper articles in daily life, relative to their peers.
2.3	To predict students' Chinese language performance at higher education institutions.
3.	Informing comparisons among educational approaches
3.1	To assess the quality of Chinese language education in Singapore.
3.2	To ensure that the raft of changes to the Chinese language curriculum and pedagogy in recent years have been well implemented.
3.3	To evaluate reading initiatives and interventions at school level.
4.	Educational management
4.1	To hold schools and teachers accountable for their students' performance
4.2	To rank and evaluate teachers for promotion and performance bonuses.
5.	Improving instructional guidance and student learning
5.1	To promote certain learning objectives and outcomes in the classroom.
5.2	To provide feedback, though minimal, to the students about their Chinese language proficiency.

Figure 5a: Purposes attributed to the GCE 1162 reading examination by different stakeholders

These ambitious purposes have been subsequently added to the GCE 1162 and its reading examination, both intentionally and unintentionally, by the various stakeholders and users of the examination, either in response to reforms in education and assessment policies or as a result of changes in political climate. The sheer range of purposes that can be associated with the GCE 1162 reading examination needs to be emphasized, since an examination which is fit for one purpose will not necessarily be fit for all purposes, even for purposes that are ostensibly similar (Newton, 2007). Even if it is assumed that these are non-conflicting purposes, it is highly improbable that a single examination can validly sustain all of them and their related interpretations.

Emerging from the interview data is the prevalent notion that Chinese language education and assessment are deeply politicized in the Singapore context. Interviewee Alpha contemplates this politicality when considering the status of Chinese in Singapore:

I would imagine the main purpose of GCE 1162 is to reaffirm our collective identity as Singaporean Chinese. It is a symbolic action [...] to convey to our students that as a Chinese, you must be able to converse in Chinese, you must be able to read and write in Chinese [...] that Chinese language is still important because it is examinable [...] It's examinable therefore it's important—that's how pragmatic Singaporeans are.

Eta, another interviewee, reiterates this view when arguing that:

Chinese language education and assessment in Singapore is bound up with politics [...] Removing GCE 1162 as an examinable subject would be unimaginable [as] the government needs to be highly sensitive to the sentiments of the Chinese educated and Chinese speaking community.

Many of the interviewees express their doubts that the GCE 1162 reading examination could fulfil its purpose of improving instructional guidance and student learning. They are aware that Singapore's Ministry of Education (MOE) has placed a

strong emphasis on the use of assessment for learning in recent years but remain sceptical that the very limited test information released about the national Chinese language examinations could be useful in this respect. When told that the *MOE Secondary Chinese Assessment Guide for Educators* produced by the Curriculum Planning and Development Division, of the Ministry of Education, Singapore (CPDD) states that ‘any form of assessment, be it the national examinations or a classroom test, has the potential to bring about meaningful learning’. (CPDD, 2014: 5), interviewee Kappa remarks:

It may be written down [in the *MOE Secondary Chinese Assessment Guide for Educators*] but in practice, this is almost unachievable. Providing feedback to help learners learn more effectively, that’s the purpose of formative assessment, of classroom assessment, not the purpose of summative assessment.

In the same vein, interviewee Zeta reasons that:

Students receive only a grade. There is no breakdown of that grade, no qualitative feedback, ok, we have a separate grade for their oral component [...] but essentially the grade does not reveal much about each student’s reading ability or his or her strengths and weaknesses. Even if the student were to re-sit GCE 1162, there is little teachers could do with the information received to help him or her improve [...] Sometimes I feel that even if there is more test information available, teachers being so overwhelmed at work are likely to lack the time and other affordances to take action, or they might simply not be equipped with the skills to do anything with it.

Interviewees speak of the crucial need for mark schemes and more detailed test specifications to be available if the GCE 1162 reading examination were to inform instructional guidance. Omicron laments:

The mark scheme for the [GCE 1162] reading comprehension paper has always been held [as] confidential. The mark schemes that we get are

produced by publishers of test prep and practice books. Even the Specimen Paper 2 released by SEAB does not come with answers [...] And yes, detailed test specifications are confidential too. This sometimes leaves teachers rather exasperated [...] From the students' perspective, this is not beneficial for learning either.

Another concern amongst some interviewees is that the GCE 1162 and its reading examination may be unfit for the purpose of teacher evaluation. It seems to be common practice for Chinese language teachers teaching graduating classes to report the number of GCE 1162 distinctions and passes that their students have produced during work reviews. Some schools may also use 'value-added' models to identify teachers for career advancement, performance bonuses or mentoring and retraining. Pi candidly puts it:

GCE 1162 might be more of a high-stakes assessment for teachers rather than students (laughs) [...] Teachers spend so much time drilling their secondary 4 and 5 students with past year papers [...] we know there is so much more to [the] reading [curriculum] than completing past year papers, but many teachers think this is the way for students to do well at the exams [...] and you know, their results affect how teachers are being ranked and graded.

All interviewees, however, agree that the GCE 1162 and its reading examination provides a necessary lens to understand and compare how students are performing even though a clearer articulation of its purposes and constructs is much needed. Without a standardized national Chinese language examination at the end of secondary education, it would be difficult to gauge the language proficiency of Singaporean students and to ensure that all students are provided with adequate educational opportunities. Valuable information about the effects and implications of education and curriculum reforms would also be lost. Further, a need is felt by a number of interviewees for the continuous monitoring of the examination's various purposes as their relevance and significance are likely to change over time.

As evidenced by the data collected, there are multiple and accumulating purposes of the GCE 1162 and its reading examination. It is at least questionable whether the test design remains optimal for each and every purpose. Newton (2007) warns that policy makers commonly conflate qualitatively different purposes into broad misleading categories such as formative and summative and, in doing so, overlook the intricacies of assessment planning and design. Test scores are often put to diverse uses and the key issue here is that stakeholders and users of the GCE 1162 and its reading examination need to come to a consensus not only about their *primary* purpose but also their other purposes. In other words, an explicit prioritization of purposes should be defined and the characteristics of the examination should be determined by a mutually agreed priority of purposes. Newton (2007: 168) also foregrounds the obligation that policy makers and test designers have to ‘identify, for all stakeholders, those purposes for which results are unfit (not simply those for which results are fit)’ so as to ‘ensure that results are not used for inappropriate purposes’.

The difficulties of doing so, however, can be formidable. Even within MOE itself there are numerous branches and units, not to mention the various autonomous institutions that MOE works with, such as SEAB and the National Institute of Education (NIE), to deliver the examination. This situation necessitates further questioning of who is ultimately responsible for the national Chinese language examinations in Singapore and who should be responsible for integrating and coordinating efforts in evaluating the purposes, constructs and administrative structure underlying the GCE 1162 reading examination. In the words of Rho, an experienced teacher trainer and academic:

It is natural that everyone wants to avoid the onus of evaluating an examination system [...] There has to be someone, or rather a group of people, who has the necessary knowledge and expertise to oversee the maintenance of quality [in Singapore’s Chinese language papers]. For a start, they need to identify what main purposes does the GCE 1162 examination serve [...] and orient design features to the most important ones.

5.3.2 Construct

There is an equally impelling need to establish exactly what is being assessed by the GCE 1162 reading examination. In Chapter 2, I outlined the development and emerging trends in second language reading research and suggested ways that the reading construct could be represented in the national examination and curriculum. In Chapter 7, I will investigate and assess the degree to which the GCE 1162 reading examination represents and adequately measures all facets of the secondary reading curriculum. In this subsection, I will focus on whether the reading construct is well articulated and if there is general agreement among stakeholders about how it may be interpreted.

The broadest declared reading construct measured by the GCE 1162 reading examination is reading comprehension (SEAB, 2014a). This is further refined as a ‘student’s ability to read narrative, expository, argumentative and functional texts of appropriate difficulty and appreciate literary texts [...] in congruence with the secondary Chinese language syllabus’ (SEAB, 2014a: 2). In Syllabus 2011, other dimensions of reading, such as cognitive skills, cultural awareness and reading interest are noted, though they are not explicitly stated in the *GCE 1162 Examination Information Booklet*. Interviewees point out that none of the official documents offers a clear and detailed explanation of what reading comprehension entails. This is problematic, not only because there is a lack of clarity in what is being assessed but also because it renders the construct potentially contestable. Rho questions:

I don’t think the reading construct has been clearly defined. What does ‘reading narrative, expository, argumentative and functional texts of *appropriate* difficulty mean’? What does ‘*appreciating* literary texts’ mean? (original emphases) [...] When it is not clearly defined, it is open to interpretation.

Delta, an interviewee who works for MOE’s CPDD shares the following thoughts:

There is ambiguity about what the reading construct is and how it is being defined [...] Without clarity, it is difficult to demonstrate [how]

what is covered in the mandated reading examination [GCE 1162] aligns with what occurs in the classroom, both in terms of curriculum and instruction [...] This is one of the reasons why when we went to schools to provide training or consultation for the teachers, they were often anxious to know what the reading paper is measuring [...] They asked us, ‘Why does our reading examination remain more or less the same when our curriculum and pedagogy have changed?’

As illustrated by Rho and Delta, when the intended construct is ill-defined, distortion could easily occur in the process of enactment. It also becomes doubly difficult to detect construct underrepresentation and irrelevance. Although any attempt to capture the complexity of reading comprehension would necessarily be inadequate, the reading construct still needs to be articulated to the best possible extent. It could be made accessible to stakeholders and users of the GCE 1162 reading examination as well as to interested members of the public. As argued in Chapter 2, in an ideal situation, the reading construct will reflect a general consensus of the views of experts in the field of second language reading, with specific regard to the Singapore context. The reading construct will also show evidence of being modern, in that it embodies current research and knowledge about the second language reading process and the assessment of reading proficiency.

Understanding of the reading process and product has certainly evolved, as demonstrated in Chapter 2. The definition of literacy is fast changing. A competent reader is no longer just one who is able to read street signs, labels and newspaper articles. It no longer suffices to decode, that is to read individual words, and to construct meaning effectively; competent readers need to develop knowledge, a repertoire of skills and strategies, and awarenesses that enable them to interact analytically and critically with continuous and non-continuous texts of different text types and genres, in both print and electronic media. Such knowledge, skills, strategies and awarenesses are subsumed under the 21st century competencies learning outcomes that the Singapore education system is now geared toward. MOE envisages that these competencies are critical if Singapore’s students are to be able to face the challenges and seize the opportunities brought about by globalization, changing demographics and technological advancements—some of the key driving

forces in the 21st century. It is therefore not uncoincidental that the foreword of Syllabus 2011 highlights that the new Chinese curriculum ‘is steered by the framework for 21st century competencies and student outcomes, adapting to 21st century trends and needs, and focusing efforts to develop our students’ cognitive abilities, communication, information and technology skills’ (CPDD, 2011: 5). It remains unclear, however, when examining the construct defined in the *GCE 1162 Examination Information Booklet*, the extent to which the GCE 1162 reading examination exemplifies the recommendations and learning outcomes of the broader curriculum. This ill-defined construct thus becomes a threat to validity.

Interviewees extol the benefits of a well-defined reading construct. Several interviewees holding leadership positions make reference to the Common European Framework of Reference for Languages (CEFR) project implemented by MOE, in collaboration with SEAB, in the early 2010s. Recognizing that a lucid description of language proficiency would not only enhance the transparency of the curriculum but also form the underlying construct for the GCE 1162 reading examination, MOE pooled resources to develop its own framework for Chinese as a second language. Theta recollects:

I remember many of our colleagues at the MOE, NIE and Singapore Centre for Chinese Language were involved in this project [...] It was in the early 2010s. I think the Minister for Education even mentioned it at MOE’s Work Plan Seminar [...] Building on the CEFR, MOE wanted to produce its own Chinese language descriptors, to make explicit what it means to be able to read, write, listen and speak in Chinese [...] This would in turn provide a basis for designing the GCE 1162 examination [...] It was a laudable effort [...] The project was largely completed, however, the framework was never released. There were concerns that there might be negative impacts, such as causing undue anxiety among parents.

This conversation about constructs led to concerns about MOE replacing the more tangible and specific explanations provided earlier with vague and overbroad statements. Interviewees with greater years in service recall the *Secondary Chinese*

Language Syllabus 2002 (Syllabus 2002) (CPDD, 2002) and suggest that the reading construct was less ambiguous in this syllabus. Syllabus 2002 was seen to have unpacked the reading construct into 13 components for the upper secondary level, specifying that students should be able to:

1. Recognize the 3,000 Chinese characters in the MOE stipulated character list in addition to mastering the form, pronunciation and meaning of these characters.
2. Recognize different punctuation marks and understand their uses.
3. Recognize the 135 Chinese idioms and proverbs in the MOE stipulated idiom and proverb list.
4. Identify implied meaning from texts.
5. Deduce the meaning of unfamiliar words from the context.
6. Deduce the sequence of events.
7. Deduce the traits of characters.
8. Differentiate between plot and subplot
9. Employ appropriate reading strategies such as skimming, scanning, careful reading and surveying along with expeditious reading and skipping
10. Possess adequate reading competency, which is demonstrated by the ability to comprehend materials of appropriate standard, including texts that are richer in content and more varied in expression, and local and international newspaper articles, including speeches by government officials and community leaders. Candidates should also demonstrate the ability to read independently relatively simple works of popular literature.
11. Read aloud fluently and with expression suitable materials such as short passages, verses and articles.
12. Apply reference tools for self-study.
13. Expand vocabulary through extensive reading outside the classroom.

In addition, Syllabus 2002 detailed the cognitive skills involved in language use, identifying 16 types of desirable skills, many of which were interwoven with the processes of reading. As observed, the definition of the reading construct became

briefers in Syllabus 2011. Some interviewees, such as Eta, reason that clearly defining the constructs ‘requires much research which is both tedious and costly’; moreover, according to Xi, ‘it is sometimes hampered by a lack of technical expertise’. These and other interviewees also speculate that more transparency could potentially lead to unwanted contention and even heated disagreement. If constructs are expressed too precisely, as Eta maintains, ‘trust in these agencies and actors will be lost if the public perceives that the construct is not being measured accurately, hence “too much” transparency may not be beneficial’. This is a classic example of information asymmetry, where government agencies ‘are averse to sharing information [...] not just because of the sensitivity of secrets, but because information is power, and asymmetry between seeker and owner of information shapes their relative power relationship’ (Ho, 2016: 120). Informed by Ho’s (2016) assertion, Singapore’s paternalistic governance culture may need to change to a participatory democratic model in the future, where the public has access to freely available and largely unrestricted information imperative for robust discussions. There are also interviewees who see the reduced construct description in the light of the seminal speech on Chinese language learning delivered by the late Prime Minister Lee in 2004. Responding to the recommendations made by the Chinese Language Curriculum and Pedagogy Committee, Lee proposed ‘taking out the drudgery of rote memorising of words and passages for examinations [...] to get the textbooks revised; wordlists revamped and reduced; examinations recast to lessen rote learning and focus on testing ability to listen, speak and read’ (Lee, 2004). The massive reforms in Chinese language curriculum and pedagogy that quickly ensued, even with the best intentions, might have led to the muddled constructs.

Issues pertaining to the lack of clarity around the reading construct are, however, not insurmountable. Interviewees who have provided training to pre-service and in-service teachers speak of ways to mitigate the problem of an ambiguous construct, such as encouraging teachers to refer to frameworks of established international assessment such as the Programme for International Student Assessment (PISA) and defining the GCE 1162 reading examination reading construct through the process of reverse engineering (Fulcher & Davidson, 2007). Reverse engineering is ‘an analytical process of test creation that begins with an actual test question and infers the guiding language that drives it, such that equivalent items can be generated’

(Fulcher & Davidson, 2007: 57). Through the reverse engineering process of critically analysing the items in the GCE 1162 reading examination, teachers can form a clearer understanding of what the examination is assessing. Mu, who has been involved in teacher training for close to a decade, discerns:

There is a lot done at ‘street-level’ by SCCL [Singapore Centre for Chinese Language], NIE, various clusters and schools [...] but there’s a call for more coordinated efforts [...] something that is more *official* (original emphasis) [...] If not teachers will always be wondering what the reading examination is trying to measure.

It seems that not only are the teachers bewildered but also students sitting for the GCE 1162 reading examination. As Rho reflects:

Students need to know, when they sit for the examination, what the examination is trying to assess and how the curriculum helps to prepare them for this [...] In second language assessment, students must know exactly what and how to prepare. There must be clearly defined constructs [...] Students are very upset. It’s not that they dislike the Chinese language [...] the anxiety of learning Chinese [in Singapore] may very well stem from the fact that these second language learners don’t have a clue as to what and how to prepare for the [GCE 1162] examination.

Rho’s perceptions are consistent with the information collected from the student interviewees. Perhaps most telling is student Omega’s response:

Interviewer: What do you think the GCE 1162 reading examination is trying to test?

Omega: (Pauses) My ability to read? (Pauses) How good my Chinese language is? [...] How many characters I recognize? I don’t really know [what the GCE 1162

reading examination is trying to assess] *leh*.³

Interviewer: How do you prepare for the reading examination?

Omega: Practice past year papers [...] No, I don't read my textbooks. In fact a lot of my classmates, including myself didn't even buy the Secondary Five Chinese textbooks (laughs) [...] Yes of course if you read a lot outside of the classroom, you will definitely do better for the reading examination [...] but I don't read in Chinese unless I absolutely have to, I don't even have a Chinese book at home.

While student Omega admits to being distracted and unfocused during Chinese language lessons at lower secondary level, he says he 'pulled up [his] socks at higher secondary level because ultimately everybody wants to do well at the national examinations'. As with the other three student interviewees, student Omega 'is not satisfied with [his] Chinese language results and wishes [he] had performed better in the examination'. Being awarded a good grade for the GCE 1162 is a significant aim for all student interviewees who are acutely aware that obtaining good results will assist them in achieving their goals and aspirations, such as gaining entrance into competitive post-secondary institutions and courses. They are also confident that increased effort would lead to improved results. Their responses become less definite, however, when asked about what the reading paper intends to assess. Their understanding of the connection between the reading paper and reading curriculum is similarly vague.

An ill-defined construct could contribute significantly to why these student interviewees are unsure of where to direct their efforts, or as Rho contends, 'erroneously believing that completing practice papers is the best, if not the only way to score well in the GCE 1162 reading examination'. Interviewees note that this misconception is likely to be reinforced in some schools by educational malpractice such as endless drilling. Providing adequate information on the examinations in the public domain may, therefore, help reduce anxiety among students. The perceptions

³ *Leh* is an expression in vernacular Singaporean English used to express doubt.

of these student interviewees are worthy of consideration and further research on a larger scale.

5.3.3 Administrative structure

I now turn to the third focus of the specifications and administration inference—the administrative structure of the GCE 1162 reading examination. Administrative structure in this context refers to the mechanisms behind ensuring confidentiality and fairness, selecting and training item setters, markers and advisers, advancing transparency in examination procedures, and promoting ongoing intra and cross organizational collegiality and research. The administrative structure relates directly to issues of reliability which is an integral part of test validity. If an examination is poorly administered, scored and reported, what can be confidently inferred about a student's performance, even when the items have been perfectly conceived and designed?

Fairness and confidentiality are highlighted by several interviewees and documents as major strengths of the GCE 1162 reading examination. Interviewees opine that keeping examination content secure is a top priority of SEAB as breaches compromise the integrity of the certification process. Interviewees who have been involved as advisers or item setters for the GCE 1162 reading examination reveal that processes are in place to ensure that advisers and item setters are screened and those who may have a conflict of interest, such as having children and immediate family members sitting for the examination in a given year, will not be selected. Advisers and item setters are also required to sign a non-disclosure agreement with SEAB, making them legally obliged to protect the confidential information they receive. In addition, rigorous measures are taken to guard the examination paper and its content closely until the examination commences. Schools and examination centres are also monitored to make certain that examinations are administered under the same conditions. In the event that cheating is detected or suspicious activity reported, invigilators and markers have to complete and file an irregularity report. The penalty for cheating and examination misconduct is severe and candidates may be barred from all GCE O Level examinations in the same year. With regard to candidates with disabilities, special modifications are made to administrative

arrangements such as providing extended time and using large print examination papers. Adherence to these provisions has meant that interviewees describe schools and teachers as being fully supportive in upholding the highest standards of security and fairness.

It is interesting to note, however, that not all security measures are deemed necessary and some interviewees perceive the need for a trade-off between security and communication with colleagues and students. Pi provides a cautionary note:

The pendulum appears to have swung too far [...] there is just too much security now. There were some hiccoughs a few years ago. I heard there was a marker who misplaced a script [...] left it in the newspaper he brought to the marking room [...] and there was another marker who dirtied the scripts when he accidentally knocked over his cup of coffee. Since then, SEAB has stepped up security measures. Now all markers have to leave their personal belongings in the lockers provided before entering the marking rooms [...] Not only that, teachers will be allocated to different rooms for a briefing based on the item they are marking [...] This way, no one will know the mark scheme for all the items [...] Are all these measures taken necessary? [...] After all, there needs to be a level of transparency so that we [as teachers] can be in a better position to help our students with the GCE 1162 examination.

This view is consistent with the need expressed by interviewees for more stakeholder engagement and transparency in SEAB's decisions and processes. When interviewees were asked about their understanding of the selection criteria for item setters, markers and advisers for the GCE 1162 reading examination, they acknowledge that they are unclear about the selection criteria and unsure if it exists. From personal experience, Theta observes that 'markers are usually nominated by schools; and item setters and advisers are recommended by officers at MOE and SEAB'. Theta adds that 'items setters and advisers are usually academics or teachers with extensive experience in Chinese language teaching'. In terms of the guidance and coaching provided, interviewees report mainly on-the-job training. Other than attending a joint briefing and calibration exercise prior to scoring the scripts, markers

work alongside presiding examiners and assessment specialists who are available to answer queries and impart useful skills. Item setters and advisers receive advisory guidelines from SEAB but as Theta, an accomplished teacher who has been involved in setting and reviewing the GCE 1162 reading examination papers alludes, they rely mainly on past papers and heuristics, in other words, rule of thumb and intuitive judgement:

When we write items, we often begin by looking at past papers. We would try and replicate the examination based on the item types and passages we see in past papers [...] We would also consult the syllabus and textbooks [...] There are also guidelines from SEAB [...] and in recent years, SEAB has conducted more workshops and training sessions for teachers to improve their assessment literacy [...] However, past papers are our most important sources of reference [...] We also depend on rules of thumb derived from experience [...] It's the same case when we review items.

Responses from interviewees reveal systemic issues that could prove to be the weak links in the validity chain. First, the processes for selecting and training item setters, markers and advisers are not publicly documented and they ironically remain unknown even to those involved in these roles. This runs contrary to best practice as highlighted in the *Standards for Educational and Psychological Testing* (AERA et al., 2014: 25):

Standard 1.9: When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether interviewees reached their decisions independently, and should report the level of agreement reached. If interviewees interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Second, when the selection processes are not openly articulated, the larger implications of the absence of a publicly available code of practice and standard operating procedures are drawn out. Interviewees interpret the dealings and operations of SEAB as ‘highly confidential’ and ‘shrouded in mystery’. This has a direct link with SEAB’s relatively opaque policy measures. When an examination board does not set forth clearly its approach to assessment and the standards against which its self-evaluation will be conducted, it indubitably leads to confusion and misunderstanding among its stakeholders and users. Public confidence in the examination board may also be easily undermined by media stories and unfounded speculation due to low assessment literacy.

The theme of transparency which recurs throughout the collected interview data becomes especially prominent when opinions are solicited with regard to the disclosure of procedures for determining the cut-scores of the GCE 1162 examination. Cut-scores are selected points on the score scale of an examination. These points are used to determine whether a particular score when attained is sufficient for the stated purposes of an examination. In the GCE 1162 examination, a test-taker’s performance is classified into one of the nine grades ranging from A1 to F9 on the basis of cut-scores. There appears to be no information on the process of how cut-scores are set for the GCE 1162 examination in official MOE and SEAB reports made available to the public. Some interviewees, like Iota, believe that ‘it is not unusual for the “technicalities” of high-stakes summative assessment to remain hidden from public view’ and that SEAB ‘as a national-level organization subject to constant scrutiny from the government must have a team of statistical experts tasked with defining the cut-scores for each examination objectively, locating them within the range recommended by the panel of subject matter experts’. Other interviewees like Beta, however, warn about threats to validity—without the disclosure of procedures in determining cut-scores, external political agenda could influence the way cut-scores are set and results are presented. Beta questions:

Many of us teaching at schools sense this [threats to validity] [...] there is a constant decline in Chinese language standards, even in schools that pride themselves on Chinese heritage, students there hardly read in Chinese too. Yet, the pass and distinction rates [for the GCE 1162

examination] seem to be stable [...] and I would say that the difficulty of the examination is comparable for the past ten, fifteen years? [...] We don't know how much external political influences are at play [...] maybe to keep schools happy and students motivated to learn the Chinese language? [...] and we all know that if you torture the numbers long enough, they will confess to anything (laughs).

Beta proceeds to share an unpublished paper presented in Singapore at the 40th *International Association for Educational Assessment Conference* (Teo, Soh, Wong & Chua, 2014). In the paper, a group of SEAB assessment specialists presents an exploratory study on the use of the Bookmark standard setting method for Singapore's mother tongue national examinations. The Bookmark method, developed by Lewis, Mitzel, Green and Patz (1999), is a promising standard setting procedure that is now widely used in large-scale assessment (Karantonis & Sireci, 2006). Teo et al. (2014) elaborate on the technicalities of the Bookmarking process, stating how a panel made up of assessment specialists, curriculum specialists and master teachers reviews the items, ordered from the easiest to the most challenging, in three sessions. In the first session, each panellist independently places a bookmark between the items judged to represent a cut-point. A borderline test-taker who possesses the minimum ability required at the specified cut-point will have less than 0.67 probability of correctly answering the items after the bookmark. Panellists discuss their bookmark placements and reset their bookmarks in the second and third sessions. The procedures described by Teo et al. (2014) remain, unfortunately, relatively vague—the specific mother tongue examinations for which the Bookmark method is used by SEAB are not known, examples are not shown and the makeup and size of the panel unrevealed. In addition, little is said by Teo et al. (2014) about what is done once the qualitative procedure is completed; the paper only indicates that 'the cut-score computations using Item Response Theory would then be carried out based on the bookmark placements' (Teo et al., 2014: 4).

Given that the results of the GCE 1162 examination have far-reaching repercussions for many test-takers and stakeholders, all examination procedures such as the determining of the cut-scores have to be documented and defensible. Rationales for any adjustments made by policy makers must also be responsibly reported. It is this

assurance of transparency and accountability that creates trust and confidence in the examination system, not unlike any other government institutions in a democratic state where access to information is an existential imperative for citizens, enabling them to trust that institutions work in their collective interests and that mechanisms are in place to review and redress any shortcomings.

To sum up, interviewees judge as inadequate the amount of disclosed and freely accessible information on the GCE 1162 and its reading examination as well as other national Chinese language examinations. Interviewees cite three main reasons why a culture of research and data sharing is vital. First, a research-led and evidence-based approach supports and encourages Chinese language assessment that is high quality, ethical and valuable. The uniqueness of Singapore's test-takers also calls for more localized inquiry into their characteristics and needs. Second, national politics appear to be the key driving force behind Chinese language assessment, with research 'taking a back seat'. Following a route defined by political ideology can restrict choice, thereby impeding reforms backed by research. Third, feedback from stakeholders and users, and insights gained from SEAB's self-assessments need to be properly channelled to target areas for development, innovation and continual improvement. Receptiveness to feedback is in line with SEAB's vision of being 'a trusted authority in examinations and assessment, recognized locally and internationally' (SEAB, 2017a).

The local data amassed by this study, however, reveal that research and data sharing are not always prioritized by policy makers and practitioners, as the words of Xi, a high-ranking officer who previously worked for the MOE Examinations Division indicate:

I think SEAB positions itself as an examination agency rather than a research and academic organization [...] The specialists at SEAB are mainly teachers with rich subject and pedagogical knowledge [...] but testing and assessment is essentially a distinct area of specialization, requiring expertise different from curriculum and instruction [...] A well-developed network for knowledge transfer and resource sharing, both internally and externally, is also lacking. Thus, when specialists

leave, they very often take along with them their research ideas, rendering their ideas inaccessible to the organization.

Xi's sentiments are echoed by Kappa, an academic and teacher who has participated in research projects commissioned by MOE:

Not all examination boards have the research capability to conduct and sustain extensive research. SEAB may also lack the research and academic talent [to do so]. This is why partnerships with universities and research centres are indispensable [...] We see a lot of successful collaborations between the Hong Kong Examinations and Assessment Authority and universities such as Hong Kong University, Hong Kong Institute of Education and Hong Kong Polytechnic University [...] This is something we could learn from the Hong Kong Examinations and Assessment Authority.

These pertinent problems are aggravated by the fact that most of the limited research on the GCE 1162 reading examination and most of the data generated by the examination, including item analyses, detailed examination reports and statistics are kept confidential. To date, SEAB has published only three academic books⁴ in addition to their annual reports and newsletters, *SEAB-Link*. This is barely sufficient for the purposes of promoting intra and cross organizational collegiality and research. Omicron's reference to an article in *The Straits Times* by Mahbubani (2015), Dean of the Lee Kuan Yew School of Public Policy at the National University of Singapore, seems apt here. In the article entitled *Trust the People, Share Government Data*, Mahbubani advocates the need for an educated and well-informed citizenry. Using the banyan tree⁵ analogy cited by the former foreign minister George Yeo, Mahbubani (2015) contends that the banyan tree needs pruning to allow more sunlight through. In other words, 'relatively speaking, the [Singapore] civil service

⁴ The three academic publications by SEAB are as follows: *Examinations in Singapore: Change and continuity (1891-2007)*, *Assessment in Singapore: Perspectives for classroom practice* and *Assessment in Singapore volume 2: Strategies and methods for classroom practice*. For more information, refer to SEAB (2017b).

⁵ The banyan tree is one of the most venerated trees in Asia. Featured extensively in Asian religions and myths, the banyan tree symbolizes wisdom and knowledge.

has been reluctant to share information', thus, hierarchy has to be cut down in order for robust debates on public policies to flourish. This delayering of hierarchy is in accordance with democratic governance which assumes a collective responsibility for decision making in which those who are bound by the outcomes are empowered to participate in the actual process in meaningful ways (Michael, 2006). The big question from interviewees that ensues concerns whether policy makers can trust test users and the public in general to make wise judgements with the information given to them. To which, Ho (2016) and Mahbubani (2015) rightly argue that if policy makers cannot trust the citizens of Singapore, what does that say about the strength and resilience of Singapore's society? The prevailing culture in Singapore with regard to sharing information, and specifically the information made available by SEAB and MOE, must therefore be changed.

5.4 Conclusion

The four a priori inferences, namely, specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters, are individually investigated, beginning in this chapter with specifications and administration. I have constructed an IA by stating the claim and generating assumptions for the three components of the claim, namely purpose, construct and administrative structure. The IA in turn prompts the construction of the VA. Based on the data gathered to form the VA, there are three key points as follows. First, an interpretation of the data suggests that there is considerable ambiguity surrounding the definition and description of the purposes and constructs of the GCE 1162 reading examination. In essence, ill-defined purposes and constructs are threats to validity as 'we risk designing assessments that are not actually needed, or that measure the wrong constructs, or that measure in the wrong way' (Newton, 2017c: 5). Second, there appears to be a common view among interviewees and the documents analysed that meticulous measures are taken by SEAB to uphold the fairness and integrity of the national examinations, although sometimes at the expense of public participation, research and innovation. Third, administrative practices for the GCE 1162 reading examination may need to be reconfigured to match the standards of transparency outlined in testing standards publications such as *The Standards for Educational and*

Psychological Testing (AERA et al., 2014) and *the Educational Testing Service Standards for Quality and Fairness* (Educational Testing Service, 2014).

Informed by data gathered in this study, the evaluation status of each assumption based on Shaw and Crisp's (2012) indicators are given below (Figure 5b):

Assumption	Provisional evaluation status based on semi-structured interview and document analysis data
1. The purposes of the examination are indicated.	Accepted with concerns
2. It is possible to identify the primary purpose(s) when there is a multiplicity of purposes.	Plausible rejection
3. Purposes attributed to the examination are achievable and non-conflicting.	Plausible rejection
4. Purposes for which the results are unfit are indicated.	Rejected
5. The constructs of the examination are indicated.	Accepted with concerns
6. Detailed explanations of what the constructs entail are given.	Plausible rejection
7. The constructs reflect a general consensus of the views of experts in relevant fields with specific consideration of Singapore's context.	Accepted with concerns
8. The constructs align with the recommendations and learning outcomes of the broader curriculum.	Accepted with concerns
9. Security procedures are in place to ensure confidentiality and fairness.	Accepted
10. Feedback channels are available.	Plausible rejection
11. Administrative procedures are documented and accessible for public scrutiny.	Rejected
12. Intra and cross organizational collegiality and research are promoted.	Plausible rejection

Figure 5b: Provisional evaluation status of the assumptions underpinning the specifications and administration inference relating to Singapore's GCE 1162 reading examination

The IA and VA of the specifications and administration inference connect to those of the other inferences to form a chain of reasoning where a weak link reduces the strength of the whole chain (Crooks, Kane & Cohen, 1996). The characteristics of test-takers sitting the GCE 1162 reading, which could potentially affect performance, will be considered next in Chapter 6.

Chapter 6 Test-taker characteristics

6.1 Introduction

It is important to keep in mind that ‘the test-taker, rather than the test task, is at the heart of the assessment event’ (Khalifa & Weir, 2009: 18). Reading examination performance invariably alters as a function of both the test-taker’s reading proficiency and of the cognitive and contextual parameters of the examination. Performance is also affected by test-taker characteristics that are not part of a test-taker’s reading proficiency. Test-taker characteristics include gender, age, socio-economic and cultural background, native language and length of formal instruction (Kunnan, 1998; Bachman, 1990). As early as the 1960s, Carroll (1961) had pointed out that the more diverse the range of test-taker characteristics of the population for which an examination is intended, the more demanding the task of test designers is in ensuring fairness and relevance. The general view shared by academics is that language ability, in this case reading proficiency, can be interpreted more meaningfully if relevant test-taker characteristics are taken into consideration during test design (Gu, 2014; Alderson & Banerjee, 2002).

Whilst there are many test-taker characteristics that can be studied, this chapter highlights the defining characteristics of test-takers sitting the reading component of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162), namely, adolescents in a highly modernized society with a rich tapestry of language and culture. As Messick (1989b) cautions, a universally applicable construct meaning across various populations of test-takers seems questionable. The GCE 1162 reading examination for Singaporean adolescent test-takers would, accordingly, need to be designed differently from, for example, a Chinese as a second language (CL2) reading examination for adolescents in Hong Kong or Tibet, or adult heritage learners in the United States of America. The next two sections construct an interpretive argument (IA) and validity argument (VA) for the test-taker characteristics inference. Constructing the IA and VA leads into discussions on the directions in which summative reading assessment could develop in the future, in as far as it is feasible to do so.

6.2 Interpretive argument

The second sub research question in this study asks, ‘Are the characteristics and needs of Singaporean test-takers taken into consideration?’ This question concerns the test-taker characteristics inference. The claim would be that test-takers characteristics and needs are taken into careful consideration by policy makers and test designers at the planning phase and duly reflected in the examination’s cognitive and contextual blueprint. Interviewee Beta, a lecturer who has extensive experience working with secondary and post-secondary students, describes GCE 1162 reading examination test-takers as follows:

We are assessing a group of adolescents who live and read in the 21st century [...] These post-millennials have very different reading interests and reading habits from us [when we were adolescents], those born in the 70s and 80s [...] Internet technologies have also affected conventional reading and thinking [...] Most of these test-takers are bilingual to a certain extent, however, the language terrain in Singapore is complex and they exhibit a wide range of Chinese language proficiency.

Building on interviewee Beta’s observations, the assumptions that lend credence to the claim are stated below:

1. The examination is supported by knowledge of adolescence and adolescent literacy.
2. The examination appeals to the reading interests of Singaporean adolescents.
3. The examination is relevant and authentic to Singaporean adolescents, paralleling their real life needs.
4. The examination takes into account new forms of reading literacy.

Supporting evidence and rebuttals that construct the VA are featured in the next section. The validity evidence collected through semi-structured interview and document analysis centres on adolescent literacy, motivation and new forms of literacy, as guided by the four assumptions.

6.3 Validity argument

There are a number of pertinent issues with regard to test-taker characteristics that attention must be directed towards. First is the issue of adolescence and adolescent literacy. From a biological perspective, adolescence is ‘the period in human growth and development that occurs after childhood and before adulthood, from ages 10 to 19. It represents one of the critical transitions in the life span and is characterized by a tremendous pace in growth and change’ (World Health Organization, 2011). Adolescence as a natural biological phenomenon is viewed to be universal and predictable in its characteristics and onset and to have far-reaching effects on all facets of human life, including cognitive, social and psychological aspects. It is also often seen as a grade-level designation, with entry into secondary school marking the start of adolescence.

Another way that adolescents are defined in the field of reading and literacy is through their literacy needs, which are positioned as separate and distinct from those of children and adults. This shift in literacy needs is brought about not only by an individual’s cognitive and psychological development per se but also by the increasing academic language demands of schooling and the individual’s expanding interests and abilities. The progression from a learning-to-read to a reading-to-learn stance underscores a set of more sophisticated skills, strategies and knowledge, which needs to be aptly reflected in the reading curriculum, instruction and importantly, assessment.

According to education theorist Chall (1996), individuals often advance through a series of stages in reading development (see Figure 6a). Chall’s model of reading development is intimately linked to other developmental models such as those of Vygotsky (1987), Piaget (1970) and Bruner (1966). These developmental models postulate that development involves changes in cognitive structures and that these changes are progressive and occur in an orderly pattern. They maintain that children pass through a series of qualitatively different stages; at any given time, the developmental level places limitations on learning possibilities. Chall’s six-stage reading development model, not unlike other developmental models, reflects a

constructivist perspective as it claims that knowledge is constructed actively by learners through interaction with the environment. Assimilation occurs when new information is fitted into an existing knowledge structure while accommodation involves adapting one's knowledge structure to what is perceived (Piaget, 1970). When assimilation and accommodation work together to create an equilibrium, the child may advance to a new developmental stage.

Stage	Name	Characteristics	Approximate Age
0	Pre-reading	Oral language develops, children learn how print functions; they acquire phonological awareness and knowledge of the alphabet.	6 months-6 years old
1	Initial reading and decoding	Children are able to read simple text containing high frequency words and phonically regular words; many words are now recognized automatically.	6-7 years old
2	Confirmation and fluency	Oral reading of grade-level text becomes relatively rapid, marked by natural phrasing and intonation.	7-8 years old
3	Reading for new learning	Reading is used to learn new ideas, to gain new knowledge and to experience new feelings, generally from one viewpoint.	9-14 years old
4	Multiple viewpoints	Reading widely from a broad range and genre of complex materials with a variety of viewpoints.	15-17 years old
5	Construction and reconstruction	Reading serves to integrate one's own knowledge with that of others, to synthesize knowledge and to create new knowledge and world views. Reading is rapid and efficient.	18+ years old

Figure 6a: Chall's six stages of reading development (adapted from Chall, 1996, emphases added)

According to Chall, normally by the end of secondary education, progressing students should be able to purposefully extract and interpret information from a variety of fiction and non-fiction texts to learn new ideas, to gain new knowledge and to experience new feelings (Stage 3). They should also begin to recognize that texts embrace multiple viewpoints and be able to discern differences in perspective (Stage 4). Expectedly, however, a significant percentage of students are unable to exhibit these key skills. In the current study, student interviewees Tau, Upsilon and Chi, who scored A1, B3 and B4 in the GCE 1162 reading examination respectively, assess themselves when interviewed about their CL2 reading proficiency as being at Stage 3 of Chall's six-stage reading development model. They affirm that they read a broader range and genre of materials with multiple viewpoints (Stage 4) in the English language, simply because, as Tau reasons, 'history, geography, social studies, science [...] most subjects are taught in the English language in Singapore'. Student Omega who did not manage to pass the GCE 1162 reading examination, considers reading in CL2 to be 'a rather tedious and slow process [...] and perpetually stuck in Stage 2 [of the Chall's six-stage reading development model]'.

Student Omega falls into the category of struggling readers, whom interviewees Epsilon and Theta, both senior educators, describe extensively during their interviews. Interviewee Epsilon observes that there are three main types of struggling test-takers. There is a group of test-takers who can read the examination passages with reasonable speed and accuracy. They, however, lack the vocabulary and higher order thinking skills needed to infer beyond the literal meaning of the passages and to answer the more challenging items. Next, there are test-takers, like student Omega, who can decode some of the characters (字) and words (词) but lack the required fluency to complete the GCE 1162 reading examination within the stipulated one and a half hours. As a result, most of their attention and time is spent on character and word identification at the expense of comprehension. Last, there are the weakest test-takers who have never successfully passed through the decoding stage. Reading is slow and halting, characterized by frequent stops at unfamiliar characters. Their reading level is several years below their grade placement and completing the items is extremely difficult for them.

The problem, then, is twofold. Interviewee Epsilon sums it up adequately:

There must be items pitched at higher order thinking skills, like synthesizing and evaluation [...] skills that are critical to adolescents in the 21st century [...] skills they should have acquired. Even though it is a second language paper, students have been learning the Chinese language for more than ten years [...] so examination standards must be upheld; yet there must be enough items for the less able students. They must still feel motivated enough to work hard and sit for the examination.

Put differently, a delicate balance must be maintained among test items of varying difficulty. This need should be clearly exhibited in the test specifications.

The second issue regarding test-taker characteristics, concerns motivation and authentic assessment. Interviews with the student interviewees suggest that their motivation to invest time and effort in preparing for the GCE 1162 reading examination is closely related to the value they attach to it. Personal incentives such as gaining entry into higher education institutions, scholarships and rewards from parents are all push factors. The value that a test-taker attaches to an examination 'is always a matter of perception, rather than designation, and this means that different types of students will be motivated to do well to different degrees' (Scott, 2011: 159). In accordance with this logic, groups of test-takers who require the GCE 1162 examination grade to progress to junior college and university and those who foresee themselves in professions that demand Chinese language proficiency will have a higher propensity to work hard for and concentrate during the examination.

Beyond the examination context, many Singaporean adolescents today are not motivated to read in the Chinese language, be it inside or outside the classroom (Aw, 2015). To be motivated can be understood as 'to be moved to do something; someone who is energised or activated toward an end is considered motivated' (Ryan & Deci, 2000: 54). In other words, motivation 'deals with the choices individuals make about which activity to do or not to do, their degree of persistence at the chosen activities and the amount of effort they put forth to do the activity' (Wigfield, 2000: 140). Interview data suggest that Singaporean adolescents are not energized or activated toward reading in the Chinese language, devoting very little time and effort

to it, as they do not perceive reading in the Chinese language as a vital aspect of their daily lives.

Further, the *2016 National Reading Habits Study: Findings on Teenagers* report, collated by Singapore's National Library Board (2017), reveals that the main barriers to reading are that Singaporean adolescents nowadays tend to be occupied with homework, co-curricular activities and most of all, screens—Internet, mobile phone applications, games, Facebook and Instagram, to name a few. Findings from my interviews with the student interviewees are consistent with the *2016 National Reading Habits Study*. Student Chi, for example, jests:

Unlike posting a photo on Facebook or completing the next level in Candy Crush [mobile game], reading does not provide instant gratification. Reading for pleasure is so passé.

Student Omega moans:

Reading frustrates our [adolescents'] smartphone sense of being connected to everyone and being everywhere at once [...] Reading is a solitary activity [...] it's even more difficult to concentrate and enjoy the process when I am not competent [in the Chinese language], more often than not I am checking the dictionary [...] and it doesn't help that a lot of the texts are so boring.

As Csikszentmihalyi (2008), a leading psychologist who created the concept of 'flow' states, humans need external incentives to take the first steps in an activity that requires a difficult restructuring of attention. Reading, like most enjoyable activities, is not natural and it requires an effort that initially many adolescents are reluctant to make. It is only when the literacy needs and reading interests of adolescents are understood that these 'external incentives' can be appropriated so as to render the design of reading examinations worth teaching to. In other words, it is time for policy makers concerned with the future of the Chinese language in Singapore to take stock of reading assessment and to ask: How can the GCE 1162 reading examination be repackaged so as to be more relevant and authentic for adolescents in

the 21st century? What are the skills and knowledge that adolescents need and how can reading examinations be created that generate validity evidence? Do the selected passages appeal to adolescents or at least reflect the kind of texts that students encounter in real life? Fully answering these questions will require a comprehensive nationwide undertaking which is beyond the scope of this study. The data amassed, however, uncover some intuitive insights into how these questions could be addressed.

At this juncture, comparison between the GCE 1162 and the Cambridge International General Certificate of Secondary Education (IGCSE) suite of CL2 examinations is relevant. Interviewee Gamma, who has been involved in developing, setting and marking the IGCSE suite of CL2 examinations for many years, draws on personal experience of starting to coordinate a team of item setters more than a decade ago:

To start with, I have a real trouble with a lot of my Chinese native speaker item writers, to get them out of the mind-set of don't write to me about Chinese New Year in Chinatown again, or quite traditional topics.

Gamma observes that the IGCSE suite of CL2 examinations, however, has focused more on authenticity in recent years. The new focus is in accordance with IGCSE adoption of the communicative testing approach, where examinations are designed to approximate the 'reality' of non-test language use (Weir, 2005; Alderson, 2000). The notion of authenticity emerged in the field of applied linguistics in the 1970s when the communicative curriculum and assessment were gaining influence and there was heightened interest in teaching and testing 'real-life' language. Studies on applied linguistics by Widdowson (1979) pointed to an overriding problem in second and foreign language testing—the frequent disparity between the way that language is evaluated and how language is used and assessed in real world communication. Widdowson (1979: 164) perceived that 'the learner's achievement was measured by examinations designed essentially to validate the syllabus rather than to reflect actual communicative needs'. Without a genuine relationship with real life language use, test-takers could fail to cope with language demands outside the classroom based on what they had been tested. Douglas (1997), Bachman and Palmer (1996) and Wood

(1993) echoed similar sentiments. Authenticity in assessment hence became a desirable characteristic embraced enthusiastically by policy makers, test designers and educators alike (Cumming & Maxell, 1999).

The next question is therefore, what is authenticity? Early on, in the 1970s, authenticity was simply associated with texts extracted from ‘real-life’ sources, as opposed to those constructed specifically for pedagogical and testing purposes (Kramsch, 1993). Authenticity has since been redefined to encompass a broader meaning. A more robust definition recognizes that authenticity is multi-dimensional, subjective and non-binary. Authenticity in reading examinations is not only about the text, but also the task, context, and cognitive processes, skills and knowledge elicited. The level of authenticity is hence defined by its degree of resemblance, or fidelity, to the criterion situation along these multiple dimensions (Bachman & Palmer, 1996). Authenticity also considers whether interactions between the examination and test-takers take place according to test designers’ plans. The definition hence covers both the situational and interactional facets of authenticity set forth by Bachman (1991).¹ Such an approach to authenticity provides adequate basis to justify test use in terms of content relevance and predictive utility (Bachman, 1990) and allows scores to be extrapolated to analogous situations in the real world.

Besides being multi-dimensional, authenticity is, to an extent, subjective and is dependent on perceptions (Gulikers, Bastiaens & Kirschner, 2004). This implies that test designers and test-takers may not necessarily share the same perception of what is authentic. If their respective views do indeed differ, the fact that examinations are often designed from a test developer’s vantage point raises validity issues. Care also needs to be taken to avoid a simplistic dichotomy between authentic and inauthentic assessment. Authenticity, as propounded by Breen (1985), is a relative rather than absolute quality. Complicating matters is the fact that reading examinations are by their very nature, artificial contexts for language use; there is therefore a debate about the degree of authenticity that can be realistically achieved.

¹ Bachman (1991) conceptualized authenticity as being composed of situational authenticity and interactional authenticity. Bachman and Palmer (1996) subsequently revised their understanding of these two aspects of authenticity which in turn came to underpin two independent test qualities—authenticity, which relates to the correspondence of test tasks to language use in real life situation, and interactiveness, which relates to the involvement of the test-taker’s traits and abilities in accomplishing a test task.

Although full authenticity may not be attainable in the examination situation, attempts have been made in the IGCSE suite of CL2 examinations to use passages and tasks which are likely to be more relevant and familiar to adolescents. Gamma proceeds to describe some of these features which have implications for how the authenticity of the GCE 1162 reading examination is evaluated. For example, the IGCSE pre-university CL2 reading examination is set around topics, such as education, media, work and leisure and the environment, that Gamma maintains are ‘more accessible to students [...] and when you have more culturally relevant topics, the students find that a lot more motivating’. Similarly, texts written by contemporary Chinese authors, such as *Yu Hua* (余华) and *Su Tong* (苏童) have been incorporated into reading examinations to introduce adolescents to a ‘contemporary China that has a very vibrant culture beyond tradition’.

The following quotation from Gamma draws special attention to the significance of having interesting passages in reading examinations to engage adolescent test-takers:

China tends to peddle a very traditional Chinese culture. What do they want to sell abroad? They want to sell Peking opera, fan dances, paper cutting, you know, very traditional things. When we look at the texts in [the] Singapore [GCE 1162 reading examination], which seem similar, I think [...] it’s how can we break out, for the learners, into something that’s more contemporary. Singapore is a highly modernized society [...] we have this dynamic and very international teenage life in many ways and they are on the Internet, they are playing Internet games, and they come back to Chinese where it is very traditional in its content.

For me as a teacher that’s the fundamental thing, you can work with different test and text types but to make my kids motivated, there has to be motivating content. And we are excited by Cambridge pre-u because we have more motivating content, it’s not so much of us trying to encourage learners, they are much more enthused by it because they can identify with it. I don’t really see that papers in Singapore will be any different. The content of anything they read to me is fundamental.

For Nu, who has taught both local Singapore students entering for the mainstream GCE 1162 examinations and international students sitting the IGCSE and International Baccalaureate (IB) suite of papers, the strength of IGCSE and IB reading examinations lies not only in passages that are more relatable to adolescents, but also in tasks that better equip adolescents with the reading abilities they will need for performing in a real-world context:

The reading tasks [for IGCSE and IB], I would say, are richer and more varied than GCE 1162. There are the information transfer items, where students are presented with a text and provided with incomplete visual stimuli, such as tables or charts, to be completed with information drawn from the text [...] This is in line with [the] PISA [reading framework] which recognizes understanding non-continuous texts as part of reading literacy [...] There are the reading tasks that require students to analyse two related passages, for instance, they are both on homework or environmental conservation but written from different angles. Students will have to draw information from both passages, compare and contrast and provide their opinions [...] Students get tested on their ability to summarize too. They read an interview script and choose from a 'heading bank' for identified paragraphs [...] These are all massively useful reading skills for adolescents to be able to take to the university or workplace later.

Interviewees generally agree that the GCE 1162 reading examination could benefit from a wider range of item types, forming a closer alignment with the real-world reading needs of adolescents. Interviewees acknowledge Alderson's (2000: 270) claims that 'any single technique for assessment will necessarily be limited in the picture it can provide [...] We should always be aware that the techniques we use will be imperfect, and therefore we should always seek to use multiple methods and techniques, and we should be modest in the claims we make.' Interviewees, however, highlight the practical dilemmas. As Lambda contemplates:

There is always a gap between the ideal and the feasible [...] Among the many challenges to consider for summative reading assessment is

how an array of reading skills vital for our adolescents can best be captured within the operational constraints of standardized testing. As much as we would like to have a wider range of item types [...] each of these new item types would have to go through a pilot study for reliability, item performance, fairness and so forth [...] that's highly consuming, in terms of time, effort and resources.

In closing this discussion on student motivation, it is worth mentioning that interviewees speak of the possibility of 'repackaging', of giving the GCE 1162 reading examination a modern revamp. Passages and item types that are more relevant and relatable, and therefore potentially more appealing, to adolescents could be introduced. Rho voices the following opinion:

It's high time we clear out the cobwebs [...] The English reading examinations have passages on mysteries, adventures and discoveries and we are still getting passages about senior citizens and the handicapped, about perseverance and diligence year after year (for the GCE 1162 reading examination) [...] that's not to say they are not essential but they are over-represented [...] and are tedious and burdensome for our 16 and 17 year-olds.

Rho adds that the motivation to read involves a set of beliefs, values and expectations, which in turn are influenced by an assortment of external and contextual factors such as peers, parents, classrooms, sociocultural expectations, curriculum and pedagogy. Whilst policy makers cannot expect a surge of interest in reading with a reformed GCE 1162 reading examination alone, using high-stakes examinations as a motivational mechanism will likely send a positive signal in the direction of change.

The third issue to be examined with regard to test-taker characteristics is adolescents and new literacies. Global economies, new technologies and exponential growth in information are rapidly transforming the world. The world is becoming 'flat' according to American journalist Friedman (2007). Drastic changes have occurred in the past two decades, enabling individuals to connect with the rest of the world much more easily than ever before. Many political and socio-economic barriers have been

removed, creating a more level global field. As a result, policy makers and test designers are now forced to re-evaluate the concepts of learning and adaptation. There appears to be a phenomenon, as described by Nu, that ‘the legitimacy of many large-scale assessments has been undermined since scoring well in these examinations does not guarantee a minimum level of competence in the set of critical abilities needed for future employment’.

Central to this shifting landscape, is the appearance and spread of the Internet. Never in the history of civilization has a new technology been adopted by so many, in so many different places, in such a short period of time and with such powerful consequences for both education and life (Coiro, Knobel, Lankshear & Leu, 2008; Friedman, 2007; International Reading Association, 2002). At present, there are more than 3.6 billion Internet users in the world, accounting for 49.2% of the world’s population (Internet World Stats, 2016). In Singapore, Internet usage is so prevalent that 81.3% of Singaporeans now use the Internet (Internet World Stats, 2016). Student Upsilon remarks:

I think the Internet is ubiquitous among Singapore adolescents. I can’t imagine life without Internet access! [...] I use the Internet for research when I have assignments and projects to complete [...] to keep up with the latest entertainment news and celebrity gossips. I read Chinese web fiction too [...] and group messages and emails [...] and I buy things on Taobao (淘宝)² [...] Life without my mobile phone and laptop would be pure torture (laughs).

The Internet has been so swiftly integrated into the private and social lives of the younger generation of Singaporeans that student interviewees allude to it as being a crucial determinant of an engaged and successful teenage life. Research has also shown that ‘the top reason for Singaporean teenagers to read is that Internet and digital devices have helped them to read more’ (National Library Board, 2017: 37). The meaning of literacy has evolved with the widespread use of the Internet (Coiro et al., 2008). To have been literate yesterday, in a world defined primarily by relatively

² *Taobao* (淘宝) is a Chinese online shopping website founded by China’s Alibaba Group.

static book and print technologies, does not guarantee that one is fully literate today in an online age of information and communication. Although the three Information and Communication Technologies Masterplans carried out by Singapore's Ministry of Education from 1997 to 2014 (Heng, 2014) have successfully brought information and communication technologies into the core of the education system in Singapore, policymakers have not yet fully considered the implications that Internet technologies have on testing and assessment. Specific to reading in CL2, the following questions arise: How different is offline reading from online reading? How do Internet technologies challenge conventional thinking about reading assessment? How might the new skills required to comprehend online content be reliably measured? These are significant issues that need to be tackled.

Evidence from the data collected from semi-structured interview and document analysis suggests that reading comprehension on the Internet is not isomorphic with traditional offline reading comprehension although there are multiple similarities. Beta corroborates:

A traditional pen and paper reading comprehension [like the GCE 1162 reading examination] requires students to read a number of common texts, answer items, often multiple-choice questions or short constructed response questions, about the content and main ideas of the texts. In contrast, when students read online, they have to generate appropriate search requests using Google or other search engines, sift through copious amounts of information [...] synthesize and critique the most relevant and reliable information [...] Each student typically follows a unique informational path, selecting a unique sequence of links to information and sampling unique segments of information from each source.

When interviewees are further exercised about the differences between offline and online reading, they list the following examples. First, online reading places greater demands on critical thinking and analysis than traditional offline reading. Adolescents when reading online need to evaluate the level of accuracy, reliability and information bias. As Kappa cautions:

The Internet is growing exponentially; we live in the era of data explosion. Faced with an abundance of online information, we must educate students to be critical consumers of information [...] they need to be equipped with the necessary skills to analyse and evaluate information. For example, they might be reading an online article without realizing that it is a paid advertisement.

Alpha elaborates:

Adolescents need to ask themselves when they surf the Internet: Is the author presenting factual information or opinions? Can the information be verified against another source? Can it be trusted? [...] What is the author's purpose and how might the purpose influence the site's claims? [...] Our adolescents may be digital natives, highly skilled at Internet games, applications and social networking [...] but they are not always as skilled at evaluating online information critically.

Although critical evaluation skills have always been necessary to engage deeply with texts in offline reading, the proliferation of unsubstantiated and even fraudulent information on the Internet poses additional challenges that are qualitatively quite different from those associated with traditional print and media sources. Fabos (2008) forewarns that as the Internet becomes increasingly harnessed for commercial purposes, educators need to understand what they are up against and provide students with evaluative skills to survive, thrive and engage in tomorrow's Internet. In the field of educational testing, it means that test designers need to reconfigure summative assessment to better capture the skills that influence online reading performance.

A second difference between online and offline literacy is that the act of reading on the Internet is perceived as a more active process than traditional offline reading. Students often read on the Internet to solve problems and answer questions. Initiated by a specific purpose, they sift through disparate sources to locate the information that meets their needs. They are constantly navigating, making choices about what to read and then taking physical action by clicking on links or scrolling up and down

the page. Successful Internet reading also requires students to actively incorporate vast amounts of data from a nearly unlimited set of sources, often presented in multiple media forms.

Eta draws attention to the need for online reading processes to inform reading assessment and instruction:

In contrast with traditional offline reading, when students read on the Internet, scanning through a vast field of sources and synthesizing the information garnered become integral to the reading task [...] Students are essentially in a labyrinth of texts, hypertexts, multiple forms of media and unlimited navigational pathways [...] it's a reading experience so different from, and perhaps much more seductive than, traditional text sources [...] Consequently, as more adolescents turn primarily to the Internet for their information, our reading curriculum and assessment need to take into consideration the differences between online and offline reading processes [...] Reading skills such as scanning and synthesizing become doubly important to becoming a successful online reader.

Last, interviewees point out that reading online is often a more collaborative and integrated process. When adolescents engage in online reading and research, they usually work collaboratively or solicit help from others online. Student Tau shares the following thoughts:

As the Chinese idiom *jisiguangyi* (集思广益)³ says, we tap into each other's knowledge when we research and read online. We usually work in groups [...] and we would share website addresses and useful resources we found. We would often post queries on our Facebook group chat [...] I think reading becomes much more integrated with listening, speaking and writing this way. When we read online, we

³ *Jisiguangyi* (集思广益) is a Chinese idiom that means drawing on collective wisdom.

listen to related news and audio clips too [...] and we share our thoughts and views on what we've read through voice or text messages.

Unfortunately, adolescents' skills at collaborative online inquiry are rarely captured with traditional reading assessments, including the GCE 1162 examinations, which assess reading performance individually and without online assistance. Reading as a component of the four tested language skills is essentially still being measured separately, though adolescents often use reading and other skills in tandem when researching online. Some interviewees mention the possibility of assimilating measures of online reading ability into Singapore's CL2 reading assessment. When I outlined the Online Reading Comprehension Assessment project,⁴ a research initiative in the United States of America to develop valid and practical assessment of online reading comprehension for schools, interviewees seemed to welcome the inclusion of online reading comprehension assessment in school-based formative assessment. Interviewees were, however, much more hesitant about incorporating it into the GCE 1162 reading examination, deeming this, in the words of Omicron, a 'monumental transformation that requires rigorous research and careful planning'.

6.4 Conclusion

This chapter has explored the defining test-takers characteristics of Singaporean adolescents sitting the GCE 1162 reading examination. Specifically, evidence gathered through semi-structure interview and document analysis has been organized around the following three issues to construct the IA and VA. First, adolescence and adolescent literacy; second, motivation and authentic assessment; and third, new literacies. Perhaps more than any other age group, adolescents signal important generational shifts in language use, mind-set and culture (Leu et al., 2009). One pattern of change observed in the study is that Singaporean adolescents are generally becoming less motivated to read extensively in the Chinese language, seemingly the direct result of digital distractions and heavy homework loads. It also appears that motivation to read is correlated with the relevance and appeal of texts available as well as the perceived value of reading in securing better education and job

⁴ More information can be found at the Online Reading Comprehension Assessment homepage (University of Connecticut, 2015).

opportunities. Another pattern is that the Internet is the defining technology for literacy and learning for Singaporean adolescents. The proliferation of the Internet necessitates additional reading skills and strategies for successful online reading comprehension and important implications can be drawn from these shifts to inform test design.

The design and development of an effective examination is necessarily iterative and cyclical. Development starts with planning and moves on to designing and trialling, followed by administration. The process continues by trialling, monitoring and reviewing test performance which then feeds back to the planning phase. In conclusion, a thorough understanding of potential test-takers needs to be established in the planning phase, failing which, policy makers and test designers risk facing threats to validity. While interviewees commend the GCE 1162 reading examination for being generally fair and free of bias to test-takers—by not favouring one gender or privileging students from higher socio-economic backgrounds and accommodating students with special needs—much more could be done to gain a clearer picture of the motivation, literacy needs and challenges of Singaporean adolescents sitting the examination. Questionnaires and information sheets could be developed and used alongside the GCE 1162 reading examination to gather valuable information about test-takers, such as their reading exposure, habits and strategies. Feedback from test-takers on the examination could also be elicited through post-examination surveys, focus groups and protocol analysis, and used in modifying passages and items where necessary. The examination could also be more sensitive to new literacies arising from the spread of the Internet which alter and extend the reading experience of Singaporean adolescents. The evaluation status (Shaw & Crisp, 2012) assigned to each assumption underlying the test-taker characteristics inference are as follows (Figure 6b):

Assumption	Provisional evaluation status based on semi-structured interview and document analysis data
1. The examination is supported by knowledge of adolescence and adolescent literacy.	Accepted with concerns
2. The examination appeals to the reading interests of Singaporean adolescents.	Plausible rejection
3. The examination is relevant and authentic to Singaporean adolescents, paralleling their real life needs.	Plausible rejection
4. The examination takes into account new forms of reading literacy.	Plausible rejection

Figure 6b: Provisional evaluation status of the assumptions underpinning the test-taker characteristics inference relating to Singapore’s GCE 1162 reading examination

As attention is turned in the next and subsequent chapters to the cognitive and contextual aspects of the GCE 1162 reading examination, it is well to bear in mind that in a socio-cognitive framework of test development and validation, there are clear links between test-taker characteristics and both cognitive and contextual parameters (Khalifa & Weir, 2009). Test-taker characteristics will have a strong influence on how a test-taker processes the passages and items and interacts with the contextual features of the examination. The dimensions of reading, cognitive skills and reading approaches that the GCE 1162 reading examination elicits from its test-takers will be investigated next in Chapter 7.

Chapter 7 Cognitive parameters

7.1 Introduction

Considerable discussion concerning both specifications and administration, and test-taker characteristics inferences of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162), was provided in Chapters 5 and 6. In this chapter, the focus shifts to the examination paper itself, where the cognitive parameters of the examination paper are delineated. Cognitive parameters, forming a critical component of Weir's (2005) socio-cognitive validity framework, are linked directly to the definition and conceptualization of the reading construct in Chapter 2. Drawing on knowledge of the reading purposes, approaches, processes and models reviewed in Chapter 2, this chapter evaluates the following three cognitive parameters of the GCE 1162 reading examination: the variety of the dimensions of reading assessment, the comprehensiveness of test items in terms of cognitive demands and the range of reading approaches demanded.

As previously stated, the GCE 1162 test specifications and the *Secondary Chinese Language Syllabus 2011* (Syllabus 2011) are in many ways limited in their explanation of the cognitive parameters of the GCE 1162 reading examination. To extend understanding, the following procedures were carried out. First, information was amassed from interviewing interviewees and examining relevant documents. Subsequently, subject matter experts (SMEs) were appointed to analyse 22 sets of GCE 1162 reading examination papers from the past decade. The purpose of so doing was to triangulate the data from multiple sources to discover similar and divergent perspectives to form a more objective judgement of the cognitive parameters of the GCE 1162 reading examination. Further analysis of the cognitive parameters of the GCE 1162 reading examination through observation, survey and scrutiny of test-taker verbal protocols is beyond the scope of this study.

7.2 Interpretive argument

The sub research question addressed in this chapter, ‘Are the cognitive requirements of the GCE 1162 reading examination appropriate and do the reading constructs sampled indicate broader competence beyond the examination?’, relates to the cognitive parameters inference. The claim for the cognitive parameters inference is justified when the cognitive requirements of the examination are appropriate and the reading constructs sampled indicate broader competence beyond the examination. The assumptions underlying the cognitive parameters inference are as follows:

1. The examination takes into account the different dimensions of reading assessment (e.g. text comprehension, knowledge and application of language and literature, multiple text reading for problem-solving, and reading volume and interest).
2. There is adequate representation of lower-order thinking items (LOT).
3. There is adequate representation of higher-order thinking items (HOT).
4. There is adequate representation of items at each specific cognitive level (remember, understand, apply, analyse, evaluate and create).
5. The examination takes into account different reading levels (local and global).
6. The examination takes into account different reading types (expeditious and careful).
7. Statistical analyses are employed in field testing to refine items in the actual examination.
8. There is alignment between the measurement objectives of the examination and the learning objectives in the syllabus.
9. The examination assesses constructs that are relevant to real-life reading contexts beyond the syllabus.

The interpretive argument (IA) constructed in this section forms the first layer of Kane’s (2009, 2006) argument-based approach (ABV) to validation by laying out the claim and assumptions. The second layer of the ABV is the validity argument (VA). Set out in the next section, the VA entails the gathering and analysis of supporting evidence and rebuttals to determine the plausibility of the IA. The VA is organized

using three headings—dimensions of reading assessment, cognitive demand of items and reading approaches.

7.3 Validity argument

7.3.1 Dimensions of reading assessment

Interviewees generally agree that measures of reading should be multi-dimensional and should attempt to encompass varied aspects of reading, such as literary appreciation and reading volume. The GCE 1162 reading examination's single way of assessing reading, namely, by testing comprehension of individual passages in a timed pen-and-paper format, inevitably has its inadequacies (Alderson, 2000). Interviewee Lambda illustrates this view:

I think the amount of time secondary school students spend reading [in the Chinese language] in the average classroom is about 20 minutes a day, and that's being optimistic [...] Not many schools have sustained silent reading periods for mother tongue languages, that's mainly for English [...] The amount of independent reading students are engaged in outside the classroom is [an] important [determinant of reading proficiency] [...] We could try incorporating assessment of extensive reading and literary appreciation into the existing GCE 1162 reading examination [...] maybe in the form of a graded Chinese language dossier?

Zhu (2015) highlights four dimensions of reading assessment. The first dimension is text comprehension. Test designers often approach summative reading assessment from the dimension of text comprehension. Test-takers are tasked to read individual passages and attempt items often in the form of multiple-choice and short-answer questions. Another common test dimension relates to test-taker knowledge and application of language and literature. Examples cited by Zhu (2015) include items that require test-takers to familiarize themselves with important quotes, verses, plots and characters from works in the literary canon. Items on vocabulary and grammar

knowledge, such as the multiple-choice gap-filling test in the GCE 1162 reading examination, also fall into this category.

The two dimensions that the GCE 1162 reading examination may have overlooked are multiple text reading for problem-solving, and reading volume and interest. Chapter 2 briefly addressed how connecting content and ideas across multiple texts involves additional cognitive skills as compared to processing a single text. Gathering and synthesizing information from various sources to reach an effective solution is a way of assessing reading comprehension that is gaining popularity (List & Alexander, 2017). Similarly, reading volume and interest is central to the development of reading proficiency. Extensive voluntary reading is verifiably a hallmark of fluent reading. In mainland China, a supplementary reading list consisting of poetry, prose and essays is included in the *Language Syllabus* (Ministry of Education of the People's Republic of China, 2011). Students are tested on their understanding of these works in the national Chinese language examinations. In addition, students are expected to have read at least four million words independently outside the classroom in their nine years of compulsory elementary and middle school education. Items in the national Chinese language examinations include writing a short analysis of any character from one of the four great Chinese classical novels¹ and applying knowledge of literary devices to critique a short poem. In Singapore, however, reading volume and interest go unmonitored in the national Chinese language examinations, including the GCE 1162 reading examination.

The predominant view expressed by interviewees is that the dimensions of multiple text reading for problem-solving, and reading volume and interest are integral to reading assessment. There is, however, less agreement on how they can be assessed, especially under high-stakes examination settings. Interviewee Omicron indicates that the:

CPDD [Curriculum Planning and Development Division, of the
Ministry of Education, Singapore] could come up with a reading list or

¹ The four great Chinese classical novels are widely deemed to be: *The Water Margin* (《水浒传》), *Romance of the Three Kingdoms* (《三国演义》), *Journey to the West* (《西游记》) and *Dream of the Red Chamber* (《红楼梦》).

bring back those abridged supplementary readers that we used to have when I was a secondary school student, like the *Rickshaw Boy* (《骆驼祥子》)² [...] but I am not sure how we could include these in GCE 1162 [...] We could possibly have one or two questions which require students to show understanding and appreciation of these texts, like the examination questions in mainland China [...] however, teachers might get students to memorize standard answers instead of encouraging them to read the texts. This not only causes unnecessary stress but also defeats the purpose [...] As for multiple text reading for problem-solving, the International Baccalaureate examination papers could be excellent sources of reference.

The very act of taking an examination may activate different sorts of reading processes from real-world reading. As some interviewees remark, it may simply be the case that certain dimensions of reading, for example, literary appreciation, enjoyment, problem-solving and creative thinking, simply cannot be measured in a timed pen-and-paper examination and need to be assessed and reported in alternative ways. This shared interviewee view resonates with that of the Chinese Language Curriculum and Pedagogy Review Committee (CLCPRC, 2004: 21)³ that felt ‘merit in extending school-based assessment to Chinese language national examinations [...] [as it is] an effective way of testing a student’s Chinese language proficiency authentically [...] [which] requires students to demonstrate skills and competencies that realistically represent problems and situations likely to be encountered in daily life’.

The CLCPRC proposed the inclusion of a school-based element in the GCE 1162 reading examination, such as a reading portfolio. To date, SEAB has not taken up this proposal from CLCPRC. Interestingly, in comparison, Hong Kong Examinations

² *Rickshaw Boy* (《骆驼祥子》) is a novel by the Chinese author *Laoshe* (老舍). First published in 1937, it is considered one of the most critically acclaimed novels of modern Chinese literature. In the 1990s, CPDD published a set of abridged secondary Chinese language supplementary readers which included *Rickshaw Boy*, *The Family* (《家》) and *Tears of Yangtze* (《一江春水向东流》) among other titles.

³ The CLCPRC was formed in February 2004 by the Ministry of Education to conduct a comprehensive review of the teaching and learning of the Chinese language in schools in Singapore.

and Assessment Authority (2013) introduced a reading portfolio and presentation component to its Middle School Standardised Examination in 2007.

7.3.2 Cognitive demand of items

Interviewees are by and large confident that the GCE 1162 reading examination has a breadth of items catering to test-takers of varying language proficiencies. Providing items of a wide range of cognitive demand is consistent with MOE's long-standing policy of customized learning for mother tongue languages to meet the needs of students from different home backgrounds, as maintained by the Mother Tongue Languages Review Committee (MTLRC, 2011). Most interviewees are, however, of the impression that there is an over-representation of lower-order thinking items that assess literal comprehension. As interviewee Alpha remarks:

There are quite a number of giveaway questions, literally, unless the student cannot make sense of the passage *at all* (original emphases). The answer is explicitly stated in the passage, [...] students need not even paraphrase [...] For the weakest students, I would get them to look at the keywords in the question stem, locate these words in the passage and just lift the sentence or sentences and voilà, they score at least 2 out of 3 points [...] Yes, there were changes made to the reading paper [in 2012 and 2016] but I think there needs to be more higher-order thinking questions, questions that require students to analyse, critique and create.

Interviewee sentiments corroborated CLCPRC opinions. In its 2004 report, the CLCPRC (2004: 22) recommended:

Chinese language examinations should do more to test thinking skills. The current Chinese language comprehension component generally assesses lower-order thinking skills such as factual recall and comprehension. There is a need to include more questions that assess higher-order thinking skills, such as application, analysis, synthesis and evaluation.

In a nationwide report published by the MTLRC (2011) after extensive consultation with teachers, students, parents, language professionals and community leaders, it is stated that Chinese language national examinations could be made more authentic to ensure closer alignment between curriculum and assessment and that there need to be more items that allow for authentic application of language skills. To gain further insights into the coverage and spread of items across different cognitive levels, 660 items from 22 sets of reading examination papers were extracted and analysed by a panel of four SMEs in this study. The total composite score for each reading examination paper is 70 marks, yielding a grand total of $70 \times 22 = 1,540$ marks. Each item was reviewed by a pair of SMEs and all incongruities in categorization were subsequently resolved by discussion with the other pair of SMEs. Findings were tabulated and the quantitative data are used in a descriptive manner to provide a point of reference for the qualitative interviews and document analysis. Figure 7a shows the cognitive levels measured by the GCE 1162 reading examination.

Cognitive level	Item		Score	
	Frequency	Percentage	Frequency	Percentage
Lower-order thinking (LOT)	602	91.21%	1,288	83.64%
Remember	127.5	19.32%	373	24.22%
Understand	465	70.45%	877	56.95%
Apply	9.5	1.44%	38	2.47%
Higher-order thinking (HOT)	58	8.79%	252	16.36%
Analyse	10	1.52%	27	1.75%
Evaluate	48	7.27%	225	14.61%
Create	0	0.00%	0	0.00%
	Total number of items: $127.5 + 465 + 9.5 + 10 + 48 + 0 =$ 660		Total score: $373 + 877 + 38 + 27 + 225 + 0 =$ 1,540	

Figure 7a: Cognitive level examined by items in the GCE 1162 reading examination (May 2006-May 2016)

The SMEs agree that an overwhelming 91.21% of items in the GCE 1162 reading examination focus on lower-order thinking skills (LOT), accounting for 83.64% of the total score of the reading examination paper. Only 8.79% of items assess higher-order thinking skills (HOT), accounting for 16.36% of the total score (see Figures 7a and 7b).

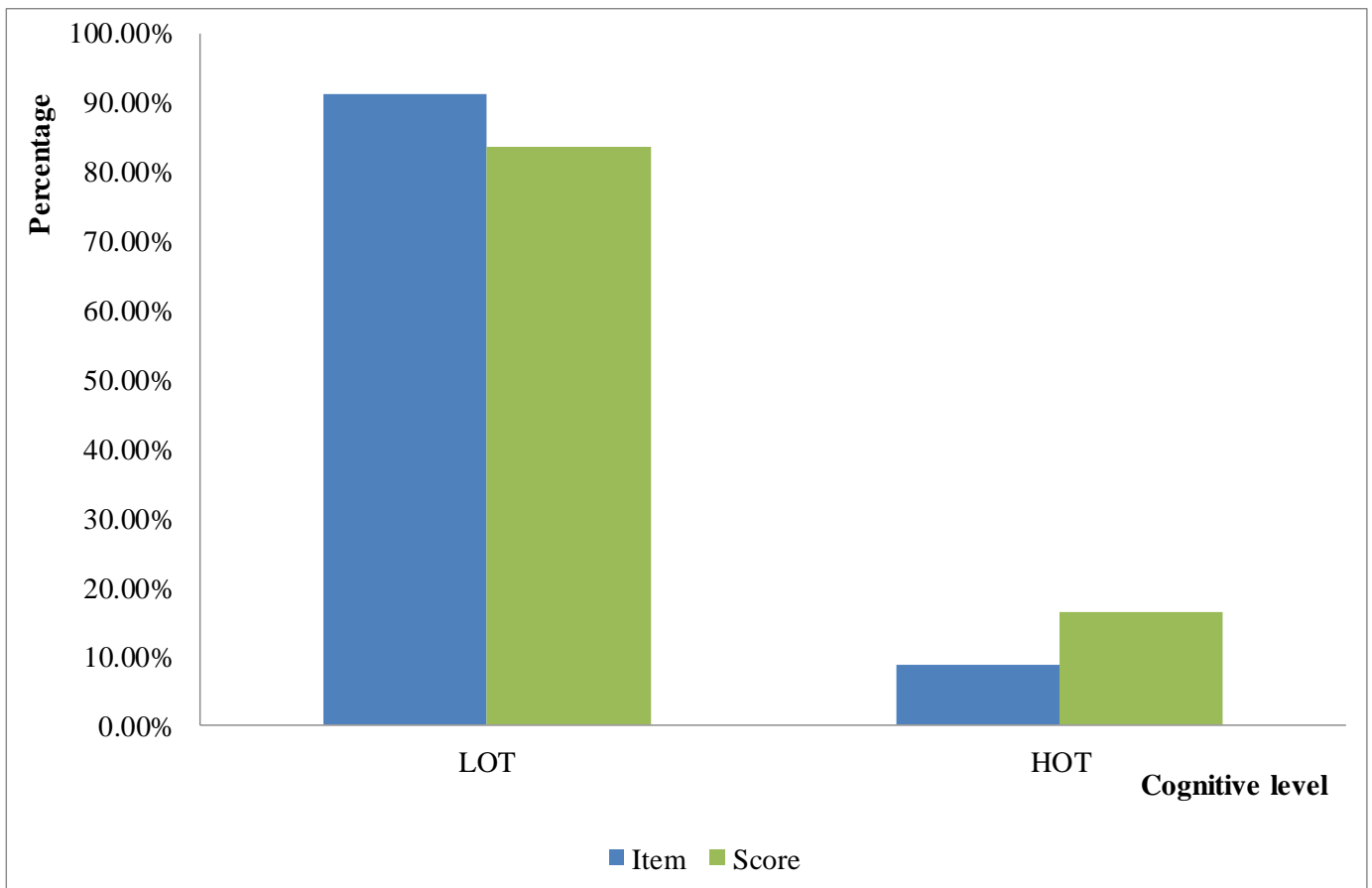


Figure 7b: Breakdown by cognitive level in the GCE 1162 reading examination (May 2006-May 2016)

Figures 7a and 7c further illustrate the occurrence of items across the six specific cognitive levels. The understand level accounts for 70.45% of all items, followed by remember items (19.32%). These two groups make up 56.95% and 24.22% of the total score respectively. In contrast, the analyse level (1.57%) and apply level (1.44%) account for the fewest items. Items at these two levels make up 1.75% and 2.47% of the total score respectively. There are no items involving creating.

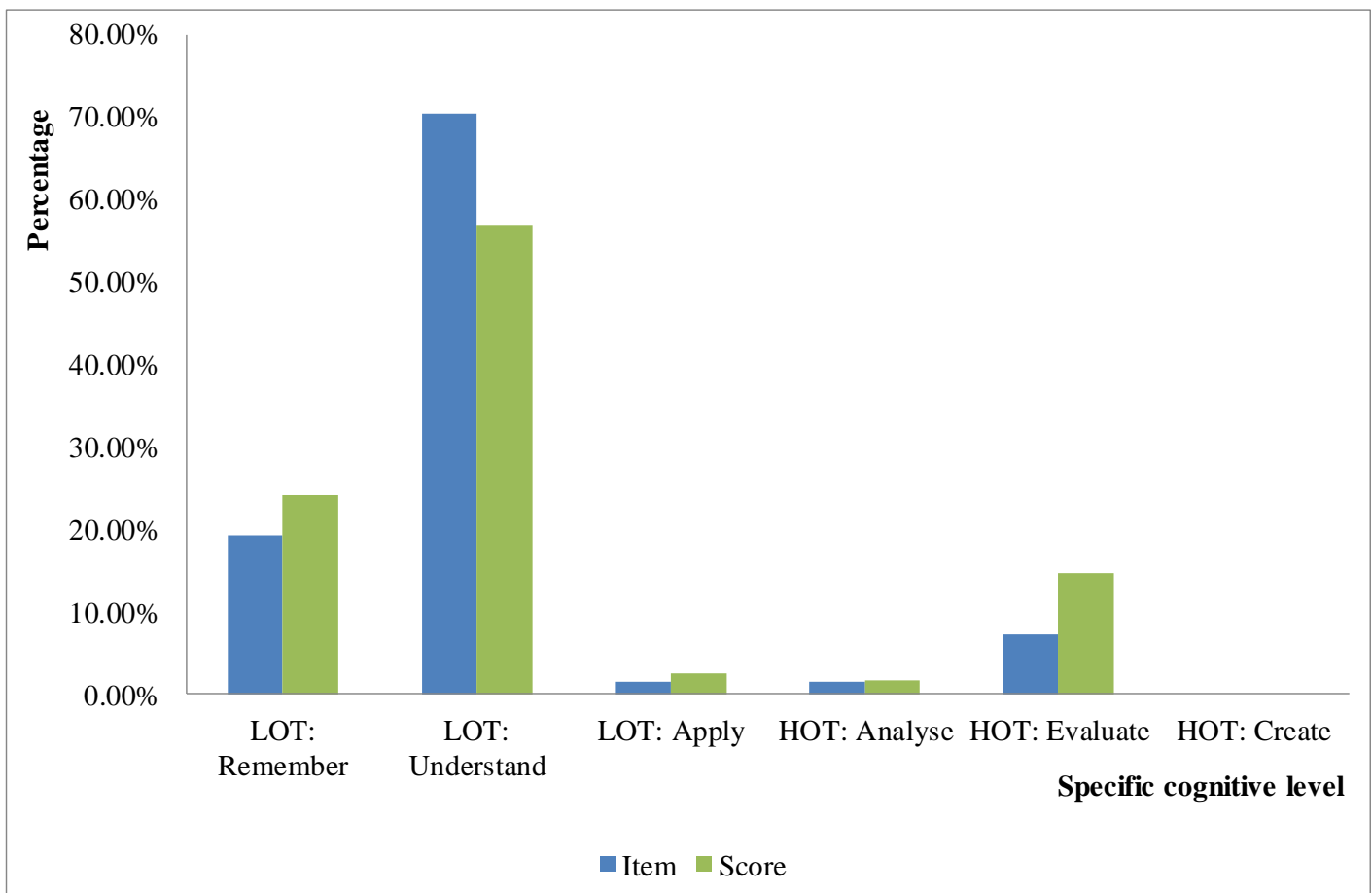


Figure 7c: Breakdown by specific cognitive level in the GCE 1162 reading examination (May 2006-May 2016)

Figures 7d and 7e document changes in the cognitive demands of items after the revisions made to the GCE 1162 reading examination in 2006, 2012 and 2016. The percentage of higher-order thinking items are 9.17%, 8.75% and 6.67%, accounting for 16.79%, 16.79% and 12.14% of the total score, for the old (May 2006-November 2011), new (May 2012-November 2015) and latest (May 2016 onwards) examination formats respectively. Using the Chi-square test of independence, a Chi-square value of 2.00 is obtained (see Figure 7f). The P-value is 0.37. The difference between cognitive demands across examination formats is therefore not statistically significant at the 5% significance level. In other words, the percentage of total higher-order thinking and lower-order thinking item scores remains basically unchanged despite the three revisions. A comparison of specific cognitive levels across the three different examination formats is also presented below (see Figures 7g and 7h).

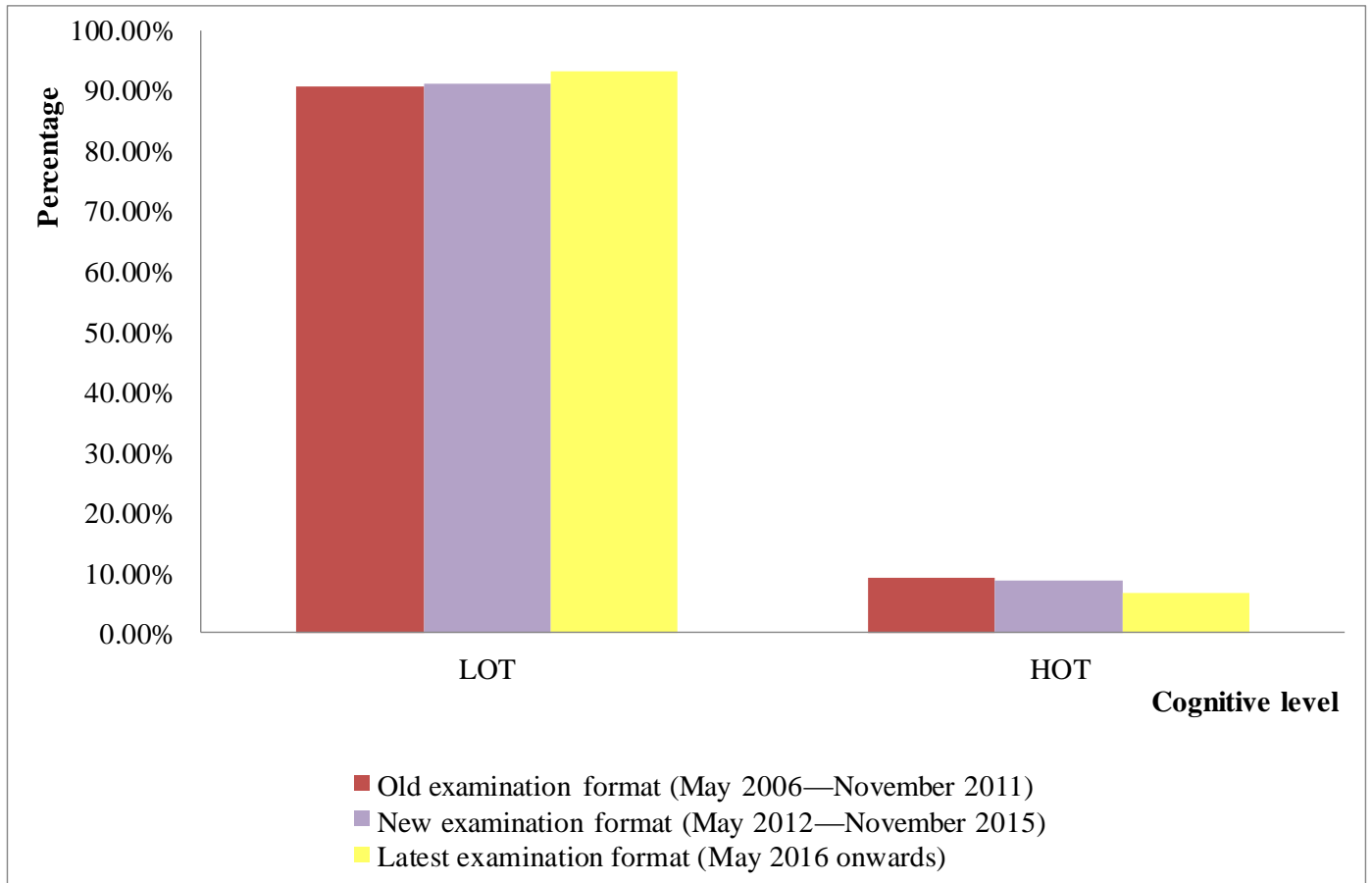


Figure 7d: Breakdown of reading items by cognitive level across examination formats

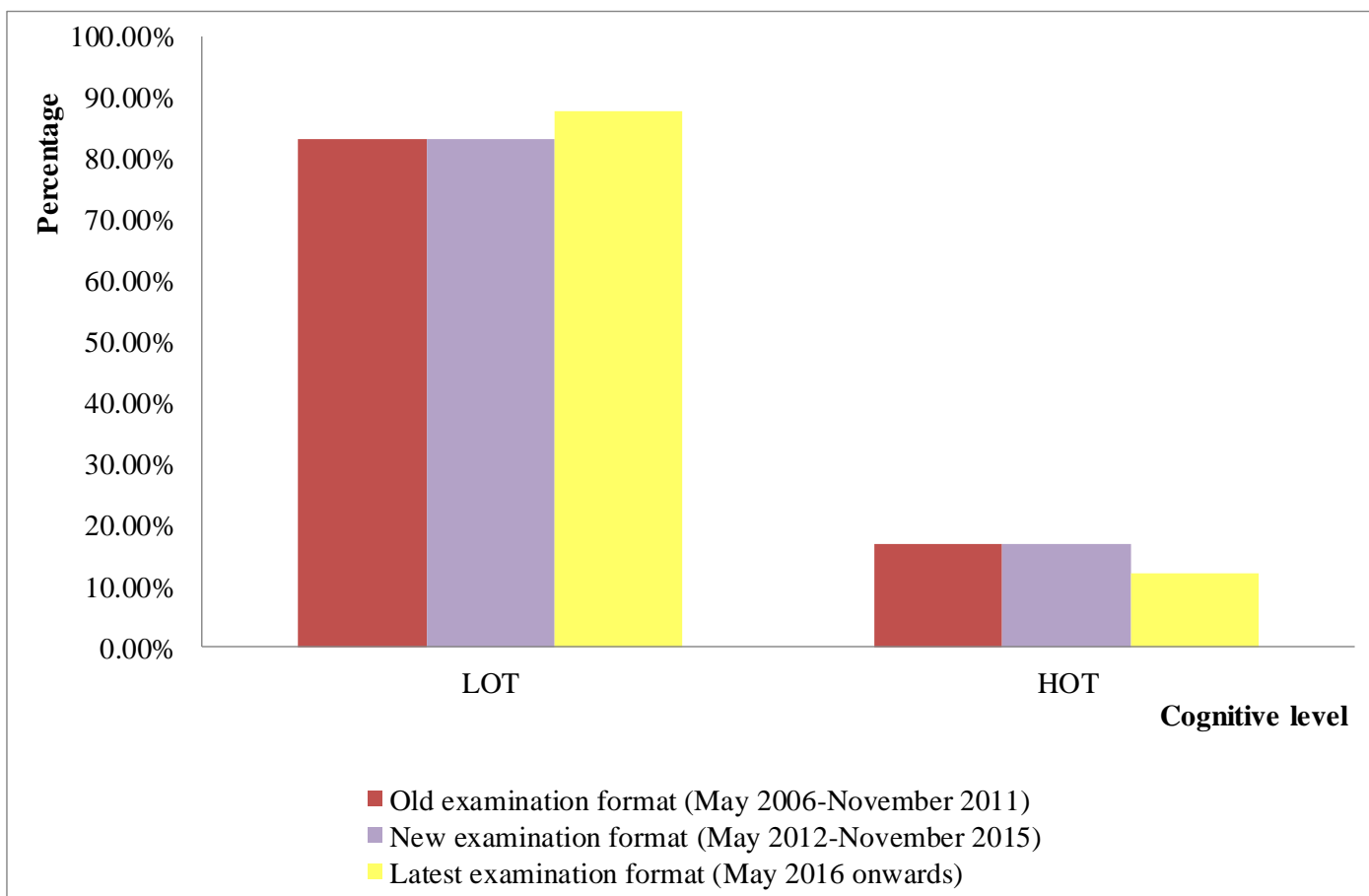


Figure 7e: Breakdown of reading scores by cognitive level across examination formats

	Old examination format (May 2006- November 2011)	New examination format (May 2012- November 2015)	Latest examination format (May 2016 onwards)	Row total
Score for lower-order thinking items (LOT)	699	466	123	$699 + 466 + 123 = 1,288$
Score for higher-order thinking items (HOT)	141	94	17	$141 + 94 + 17 = 252$
Column total	$699 + 141 = 840$	$466 + 94 = 560$	$123 + 17 = 140$	$1,288 + 252 = 1,540$ (Grand total)
Chi-square	2.00			
P-value	0.37			

Figure 7f: Reading scores for LOT and HOT items across examination formats

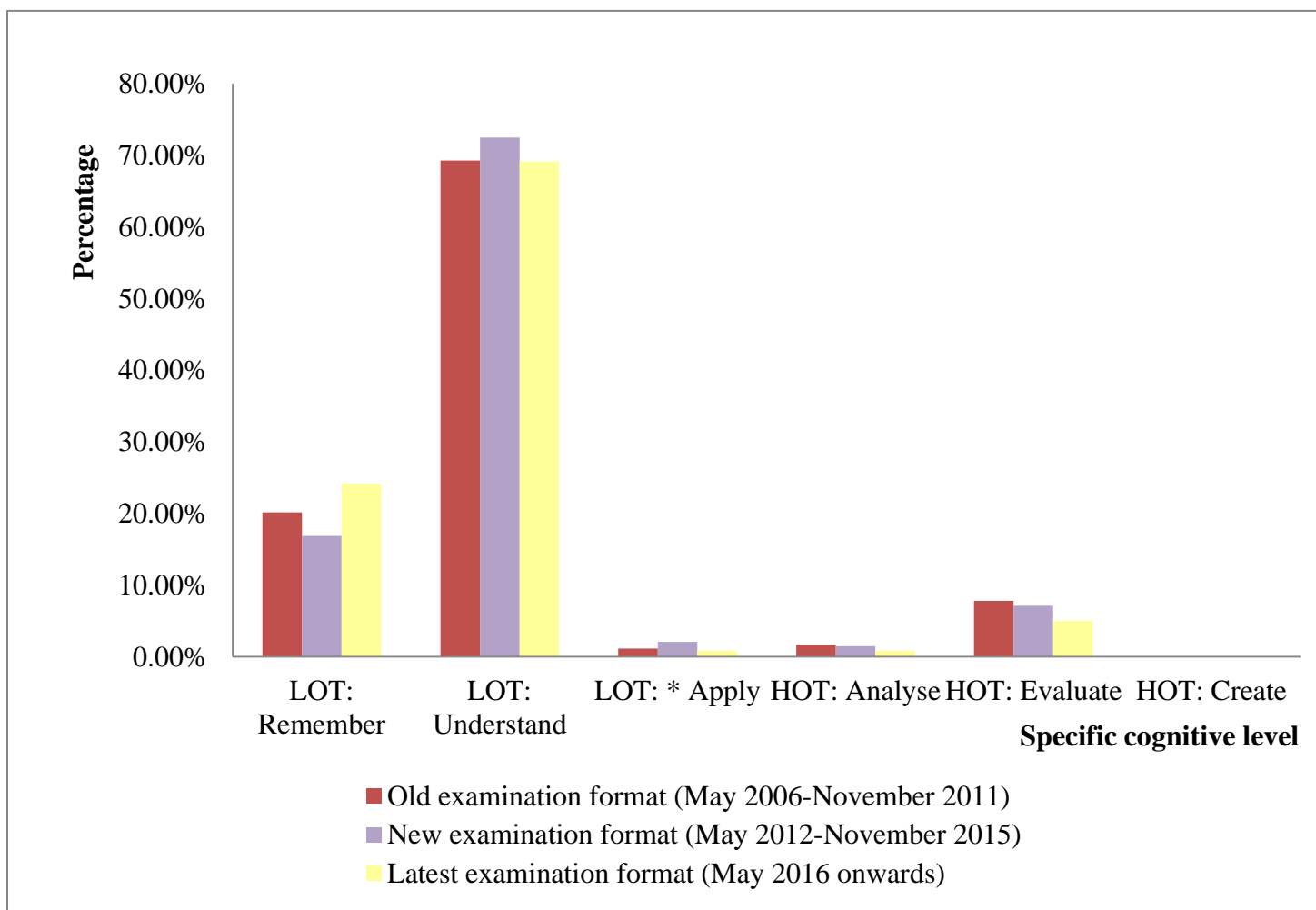


Figure 7g: Breakdown of reading items by specific cognitive level across examination formats

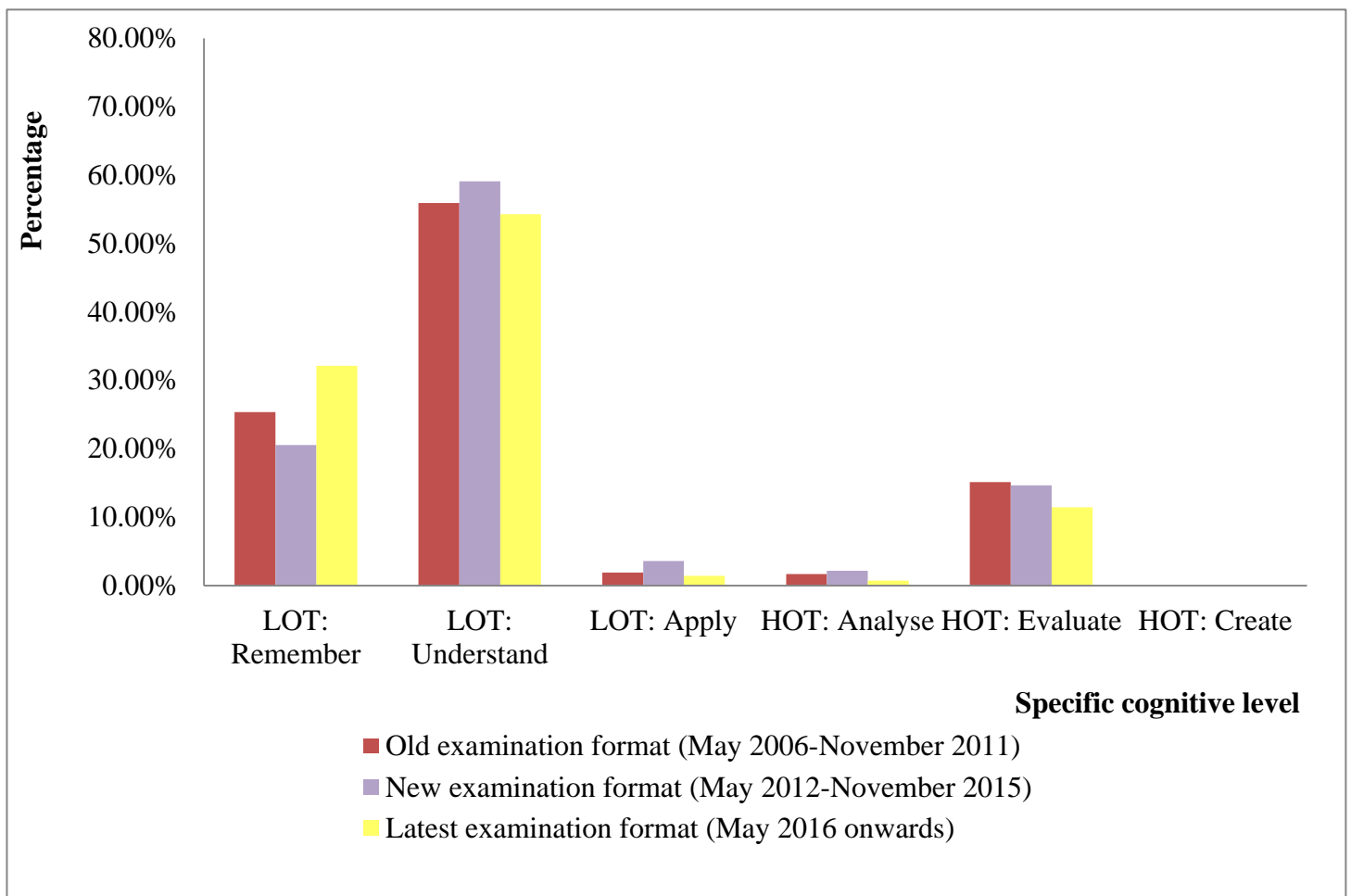


Figure 7h: Breakdown of reading scores by specific cognitive level across examination formats

The opinions of the SMEs accord with the observations from the interviews and document analysis. In 2006, following the CLCPRC's (2004) recommendation to abolish rote learning, three components were removed from Paper 2, fill in the blank, grammar and sentence construction (see Figure 4g in Chapter 4). Testing a student's knowledge of words and phrases in discrete isolated sentences, the CLCPRC claims, is not an effective way of assessing reading competency. The CLCPRC (2004: 20) also cited 'the memorisation of words and phrases to prepare for examinations [...] as the top reason that students across most levels dislike learning CL'. Interviewees are, however, considerably less certain as to whether there has been an increase in the proportion of higher-order thinking items advocated by the CLCPRC. Changes in

the proportion of higher-order thinking items are ‘not substantial’ and ‘almost non-apparent’ to some, even more than a decade after the proposal was put forward.

It is discernible that, the overall cognitive difficulty of the reading examination could have fallen slightly, as several interviewees speculate, with the introduction of more functional texts in 2012.⁴ The multiple-choice items that follow the functional texts are ‘generally too simple’, ‘testing predominantly lower-order thinking skills of identifying and summarizing explicitly stated information’. Another revision was made in 2016.⁵ Although the later revision mainly concerns the oral component of the GCE 1162 examination, there are interviewees who foresee more higher-order thinking reading comprehension items that require students to ‘appreciate and respond aesthetically to literary devices’ in future examinations. The inclusion of a greater number of higher-order thinking items is intended to better align the reading examination with shifts in curriculum objectives. It is too early to draw any conclusions as only one official specimen paper issued by SEAB exists along with the first reading examination paper administered in May 2016 at the time of writing. There is cause for concern among interviewees, however, since the official specimen paper consists almost entirely of passages and items from the May 2006, May 2007 and November 2007 examination papers, signalling that there might be no fundamental changes in the cognitive demands of the examination, among other parameters, for the coming years.

Given the focus of this chapter on cognitive parameters, I will now present a qualitative description of the items at each cognitive level across the 22 sets of selected GCE 1162 reading examination papers. An outline of the underlying reading theories and models has been presented in Chapter 2. I recognize, of course, that the

⁴ In view of *The Mother Tongue Languages Review Committee report* (MTLRC 2011), SEAB introduced more functional texts to Section 2 of the examination in 2012 to increase the authenticity of the assessment. Examples include an air purifier flyer (Passage 2C, Sample 2016) and a newspaper article on longer green man road crossing time for the elderly and disabled pedestrians (Passage 2A, May 2013).

⁵ The latest revision made to the GCE 1162 examination was in 2016. Changes were mainly reflected in the oral examination. Prior to 2016, a picture consisting of a particular scene, for example an airport or school canteen, was used. This section was known as picture description and a separate component, named conversation followed. The new format retains only the latter and uses a video clip instead of a picture as stimulus.

influence of items, passages and mark schemes on the cognitive load of a reading examination is equally important and the difficulty of the examination is dependent upon all three factors. In Chapter 8 which addresses contextual parameters, I provide a detailed analysis of the passages and mark schemes of the examination.

The first level in the 2001 revised Bloom's taxonomy (Anderson & Krathwohl, 2001) is remembering. Items at this foundation level test literal comprehension. These items offer a quick route to check for basic understanding of texts. Test-takers are only required to identify information in basically the same form in which it is presented in a passage. Reading processes involve mainly word recognition as well as syntactic parsing and extracting propositional meaning at sentence level. The items are cognitively less demanding as there is no necessity to make connections to build a mental or text model. Implicit in the remember items is the assumption by item setters that expeditious reading is encouraged for the average and above average test-taker as information can be quickly retrieved. There are five types of remember items found in the GCE 1162 reading examination:

Of the types of remember items, the first measures recognition of facts and details, often the who, what, when and where of a text. Examples in the examination include the traits and feelings of a character or the functions of an advertised product:

空气净化器能除臭、去毒和净化空气，靠的是_____

The advertised air purifier is able to remove unwanted odours, eliminate harmful gases and freshen the air because it _____

(Q15, Specimen Paper 2016)

A second remember items type is the recognition of main ideas. Test-takers are expected to identify the main idea, moral or theme of a paragraph or entire passage which is explicitly stated, for example:

这段文字的重点是什么？

What is the main idea of this short passage?

(Q11, November 2015)

A third remember items type is the recognition of a list or sequence. Test-takers are asked to recall a list of items, actions or events or the sequence in which they appear or take place. This includes multiple-choice items that require test-takers to choose from a list of options, an item, action or event that is not mentioned in the text, for example:

以下哪一点是文中没有提到的？

Which of the following is not mentioned in the passage?

(Q14, Specimen Paper 2016; Q18, November 2007)

A fourth remember items type is the recognition of comparisons, such as the similarities and differences between characters and events that are clearly presented in the text, for example:

今年的植树节活动和往年有什么不同？

What is the difference between the Plant-a-Tree Event held this year and those held in previous years?

(Q12, November 2015)

The fifth and final remember items type is the recognition of cause and effect relationships. This includes explicit reasons and outcomes of actions and events, and associations between concepts, for example:

作者认为宽容和烦恼有着怎样的关系？

According to the author, what is the relationship between forgiveness and distress?

(Q22, May 2015)

The second level in the 2001 revised Bloom's taxonomy is understanding. Items at the understand level go one step beyond simple recognition or recall of information. The information is for the most part explicitly stated but requires additional processing, thus increasing difficulty. Establishing a text model at the global macro-level is often not necessary though test-takers will benefit from reading the entire text to see how it fits together. Understand items are the most dominant item type in the

GCE 1162 reading examination (70.45%). There are five types of items that test-takers have encountered at the examination:

The first type of items assesses the ability to paraphrase. Although the information is readily available in the text, test-takers have to express it in their own words to achieve greater clarity and to answer the question fully, for example:

为什么蔡耀星在第一次得到金牌时，很多痛苦的思绪都涌上心头？

Why was *Yao-Xing Cai* overwhelmed with grief when he first won the gold medal?
(Q27, May 2008)

The first three paragraphs of the passage depict a disabled swimmer's (*Yao-Xing Cai*) perseverance in the face of ridicule and external barriers. The swimmer had overcome huge obstacles to succeed and was therefore overwhelmed with grief when reflecting upon the past. In order to answer this question, test-takers will have to identify and reword relevant information instead of directly lifting chunks of text.

A second type of items assesses the ability to reorganize. Reorganization is based on literal understanding of a text. Information is retrieved from various parts of the text and combined to show deeper understanding, for example:

“我”对岛国的看法有怎样的转变？

How has the author's opinion of the island city-state changed?
(Q23, May 2010)

At the beginning of the passage the author complains of boredom. At the end, after having travelled abroad, the author starts to appreciate the convenience of living in Singapore. These pieces of information from various parts of the passage have to be organized and linked with conjunctions to form a cogent answer.

A third type of items focuses on the ability to summarize. Test-takers must take larger selections of text and focus on the heart of the matter—the gist and the key

ideas. Summarizing involves the distillation and condensing of a text into its primary notions, for example:

以下哪一句话最适合作为这个广告的标题?

Select the most suitable caption for this advertisement.

(Q15, November 2012)

A fourth type of items focuses on the ability to explain. The meaning of a word, phrase or sentence is sought in the context given, for example:

试解释(这句话)在文中的意思:

该放手时就放手, 植树与育人, 道理相同。

Explain the meaning of the following sentence:

Growing trees or raising children—sometimes we just have to let go.

(Q29b, May 2011)

Items in the multiple-choice gap-filling section also fall into the category of explain.

The final type of items assesses the ability to make inferences. Test-takers have to deduce or conclude from information and evidence in the passage, although at this cognitive level, the process is linear and straightforward, for example,

“风筝聚会”是怎样形成的?

How was the kite flying interest group formed?

(Q14, May 2012)

In the passage, it is written that ‘no one knows when the interest group was formed’, ‘the group is not managed professionally’ and ‘students and residents come to the grass patch to fly kites’. From these pieces of information, test-takers have to infer that the interest group was formed voluntarily and spontaneously by students and residents who share an interest in kite flying.

The third level in the 2001 revised Bloom's taxonomy is application. Application refers to the ability to use information, concepts, methods or principles in a new but related situation to answer a question, solve a problem or to perform a task. For example, students may be required to select and then transfer what has been read in the text into a similar situation. The lower-level apply items that test-takers encounter in the GCE 1162 reading examination lack variety, predominantly asking test-takers to relate their personal experiences to the passage and provide real-life examples to support their interpretations and views:

作者认为学会原谅，最大的受益人士自己。你同意他的看法吗？试举一个你生活中的例子加以说明 (emphases added)。

'Forgiveness is not something we do for other people. It's something we do for ourselves.' Do you agree with the author's point of view? Why? *Use a specific example from your personal experience to support your opinion* (emphases added).

(Q30, May 2013)

Analysing represents the fourth level in the 2001 revised Bloom's taxonomy. Of the three levels of higher-order thinking skills, namely, analyse, evaluate and create, an analyse item is one that assesses a test-taker's ability to break down information into its component parts so as to identify the parts, study the relationship between the parts and recognize the organizational principals. It represents a higher cognitive level than understanding and application because test-takers are expected to grasp both the content and structure of the text. Integrating information across long stretches of text and forming macro mental and text models are often required. Examples of *analyse* items include forming hypotheses about the author's perspectives, detecting logical fallacies in reasoning and determining the relevancy of the information presented.

Analyse items make up an exceptionally small percentage (1.52%) of the 660 items reviewed. Half of the 22 sets of GCE 1162 reading examination papers do not contain analyse items. This appears to be somewhat anomalous in an examination targeted at students who have been learning the Chinese language for ten years or

more. The analyse items found in the examination can be grouped into one of two categories:

The first category comprises items that require test-takers to determine the intentions, feelings or viewpoints of the author, for example:

文中结尾所说的“区别”指的是什么？作者为什么会有这种感叹 (emphases added)?

From paragraph 6: What does ‘the difference’ refer to? *What explains the author’s argument* (emphases added)?

(Q30, November 2010)

The passage given is about problem-solving. A hotel spends tens of thousands of dollars removing snow and ice from a slope; a nearby inn solves the same problem with only thirty dollars. Advanced surveying, comparing and inferencing skills are needed to arrive at the answer to the analyse item (as emphasized above), namely, that most complex problems have a simple solution and all that is needed is a paradigm shift.

The second category consists of items that require test-takers to determine the style of writing or use of literary effects, for example:

作者说“每个人都像长颈鹿那样在眺望”，他是用了什么写作手法？

The author says that ‘the people stood looking like giraffes stretching their necks’. Which literary device is being used here?

(Q13, May 2012)

Answering the following question involves deducing the layers of meaning in an allegory:

本文传达了什么重要的信息？

What message is the passage trying to convey?

(Q18, November 2010)

The passage is a short Zen story about a man who found his way out of a labyrinth using the rock that tripped him. Test-takers need to analyse the story correctly to reveal the hidden meaning of turning opposition into opportunities in life.

Evaluating represents the fifth level in the 2001 revised Bloom's taxonomy. Evaluation items require test-takers to relate information in a text to their own knowledge and experience to form reasoned judgements of various kinds or to articulate emotional and aesthetic responses. At this cognitive level, test-takers are actively engaged in the process of deep reading and critical thinking. There appears to be only one type of evaluate item in the examination, which involves test-takers forming judgements of acceptability and worth based on their value systems and beliefs, for example:

作者认为只有怀着积极心态的人，才能在与人竞争中立于不败之地。你同意吗？试举一个生活中的例子加以说明 (emphases added)。

'Only people with a positive outlook can succeed in a fiercely competitive society.' Do you agree with the author's point of view? Why? Use a specific example from your personal experience to support your opinion (emphases added).

(Q30, November 2011)

“受苦的人，没有悲观的权利”这句话带给你什么启示？

'A sufferer has no right to pessimism.' What can we learn from this statement?

(Q30, November 2009)

There seem to be no evaluate items in the examination papers reviewed that call for judgements of fact or opinion, adequacy and validity, appropriateness and importance, or literary value and significance as specified in the Syllabus 2011.

The final level in the 2001 revised Bloom's taxonomy is creating. At the highest level of thinking, test-takers are encouraged to assemble parts to form and generate a new whole. Create items invite test-takers to go beyond limitations and be original and fresh in their ideas, solutions and perspectives. Test-takers not only need to read the lines and read between the lines, they need to read beyond the lines. These items enable test-takers to find real value in the information they are reading. Create items

in reading comprehension can include modifying the plot and ending, introducing new characters and catalysts, and proposing solutions and alternatives. Although creating is one of the key skills documented in the Syllabus 2011, there are unfortunately no items in the GCE 1162 reading examination set at this cognitive level.

From the judgement of the SMEs, it can be cautiously concluded that there is an under-sampling of higher-order thinking skills in the GCE 1162 reading examination. The bulk of items are set at the lower-levels of remember and understand, posing a threat of construct over-representation. There are too few analyse, evaluate and create items accounting for too small a percentage of the total score to support legitimate inferences in these domains. If, as the Syllabus 2011 envisions (CPDD, 2011), Singaporean Chinese students are to become lifelong independent learners actively engaged with ambiguous and unfamiliar problems, including those drawn from real life, then more emphasis needs to be placed on higher-order thinking items that require complex reasoning, judgement and creativity. It might also be prudent to ensure that the examination is eliciting data on test-taker ability to form intertextual representations. After all, as many interviewees point out, students take their cues about what is important from what is being assessed, ‘if you want to change the way students learn, then change the way they are being assessed’.

7.3.3 Reading approaches

On the topic of reading approaches, interviewees in general are less than certain as to whether the GCE 1162 reading examination sufficiently elicits responses to and assesses each of the four categories of local and global and expeditious and careful interpretation. As Iota reasons: ‘we do not have control over the time test-takers spend on each part of the reading paper. Hence, we cannot be sure how test-takers approach individual items’. Test-takers, for example, when faced with an expeditious local item may feel the need to read the whole passage with a high level of attention in seeking reassurance that they have given the correct answer, which seems especially true for weaker test-takers, as the interview data would suggest. Nevertheless, interviewees and SMEs are of the view that test-takers have to cope with both expeditious and careful reading types at both local and global levels if they

are to complete the paper, and score reasonably well, within the one and a half hour time limit. More precisely put, test-takers have to demonstrate the ability to approach texts in the following ways:

The first approach is expeditious reading at the local level, namely, scanning. Scanning is reading a text quickly and selectively at the local level in order to find a specific piece of information, such as particular words, names, figures and facts. Only word recognition and a limited amount of syntactic parsing are involved, building up a macro text or mental model is not required. The cognitive level demanded seldom exceeds remembering. Sentences are often not read in full and a low level of attention is accorded until a potential match is found. From glancing through a television listing for a favourite show to checking the time a train leaves in a train schedule—scanning is essential in everyday life as it allows a reader to save time and effort.

The second approach is expeditious reading at the global level, namely, skimming. Skimming involves reading rapidly to extract the gist and purpose of a text or to discover an author's tone and intention. Readers attempt to build a macro-structure of the entire text, at the same time linking the information with their existing knowledge and experience, with as few details from the text as possible. Their metacognitive mechanisms monitor whether the information gleaned is useful and appropriate. Skimming is particularly helpful when carrying out research under time constraints—a reading speed of 700 words per minute and above means that large amounts of materials can be read quickly for general understanding.

Interviewees and SMEs comment that test-takers of the GCE 1162 reading examination are often advised by their teachers to use skimming at the pre-reading stage in order to establish preliminarily the main idea of a passage; and for reviewing when they have completed the items. Skimming alone, however, is less than ideal when complete comprehension of the text is the main objective. To arrive at an accurate answer for global level items in the examination necessitates careful, rather than merely expeditious, reading.

The third approach is careful local reading. This refers to the approach where a reader attempts to extract complete propositional meanings at the local level, from within sentences to a short paragraph, effected by close attention to individual words, syntax and syntactic structure of clause and sentence. Some local inferencing may also be required. Unlike careful reading at the global level, careful local reading does not entail integrating individual pieces of local information into a larger meaning representation.

The fourth approach is careful global reading. This type of reading is defined by a sustained meticulous understanding and interpretation of the text. The majority of information in the text is processed and utilized to construct a macro-structure. Propositions are analysed and organized into a hierarchy of meaning. Careful global reading draws upon most if not all of the reading processes discussed in Chapter 2. In the real-world context, careful global reading is an approach readers commonly adopt when reading to learn and collating propositional information across texts for academic writing.

Using the above information on reading approaches, SMEs evaluated 660 items from 22 sets of GCE 1162 reading examination papers. 76.52% of items, accounting for 68.05% of the total score, require only reading at the local level (see Figure 7i). An example of a local-level item is as follows:

国际口足画艺协会是个怎样的组织？

What is the mission of the Association of Mouth and Foot Painting Artists?

(Q26, November 2014)

The association's mission is clearly stated in the second and third sentences of the first paragraph.

For 23.48% of the items, accounting for 31.95% of the total score, test-takers have to cope with reading at the global level (see Figure 7i). For example:

本文对你的学习有什么启发？试写出你的看法。

Using your own words, what lessons can we draw from this passage that can be extrapolated to learning?

(Q30, Specimen Paper 2016; Q25, November 2007)

Test-takers answering this item which was set in both the Specimen Paper 2016 and November 2007 examination paper needed to process the entire passage on rock climbing and relate it to their own learning experiences.

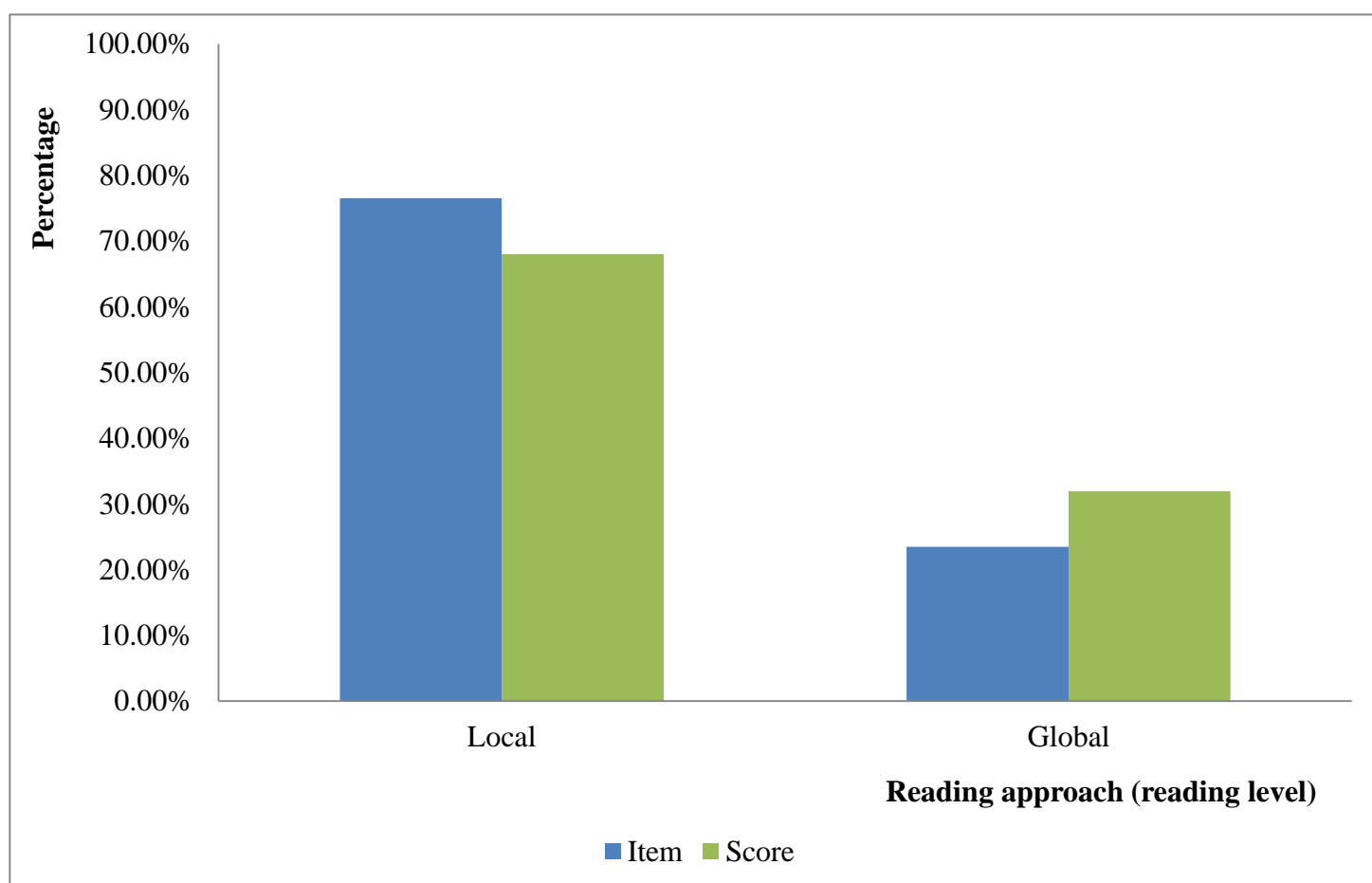


Figure 7i: Breakdown by reading level in the GCE 1162 reading examination (May 2006-May 2016)

Figures 7j and 7k show the difference in the breakdown of local and global level items for the old (May 2006-November 2011), new (May 2012-November 2015) and latest (May 2016 onwards) examination formats. The percentage of global-level items are 23.61%, 24.17% and 20.00%, accounting for 31.90%, 33.75% and 25.00% of the total score for the old, new and latest group of examination papers respectively. Using the Chi-square test of independence, a Chi-square value of 3.95 is obtained (see Figure 7l) with a P-value of 0.14. The difference between reading levels across examination formats is therefore not statistically significant at the 5% significance level.

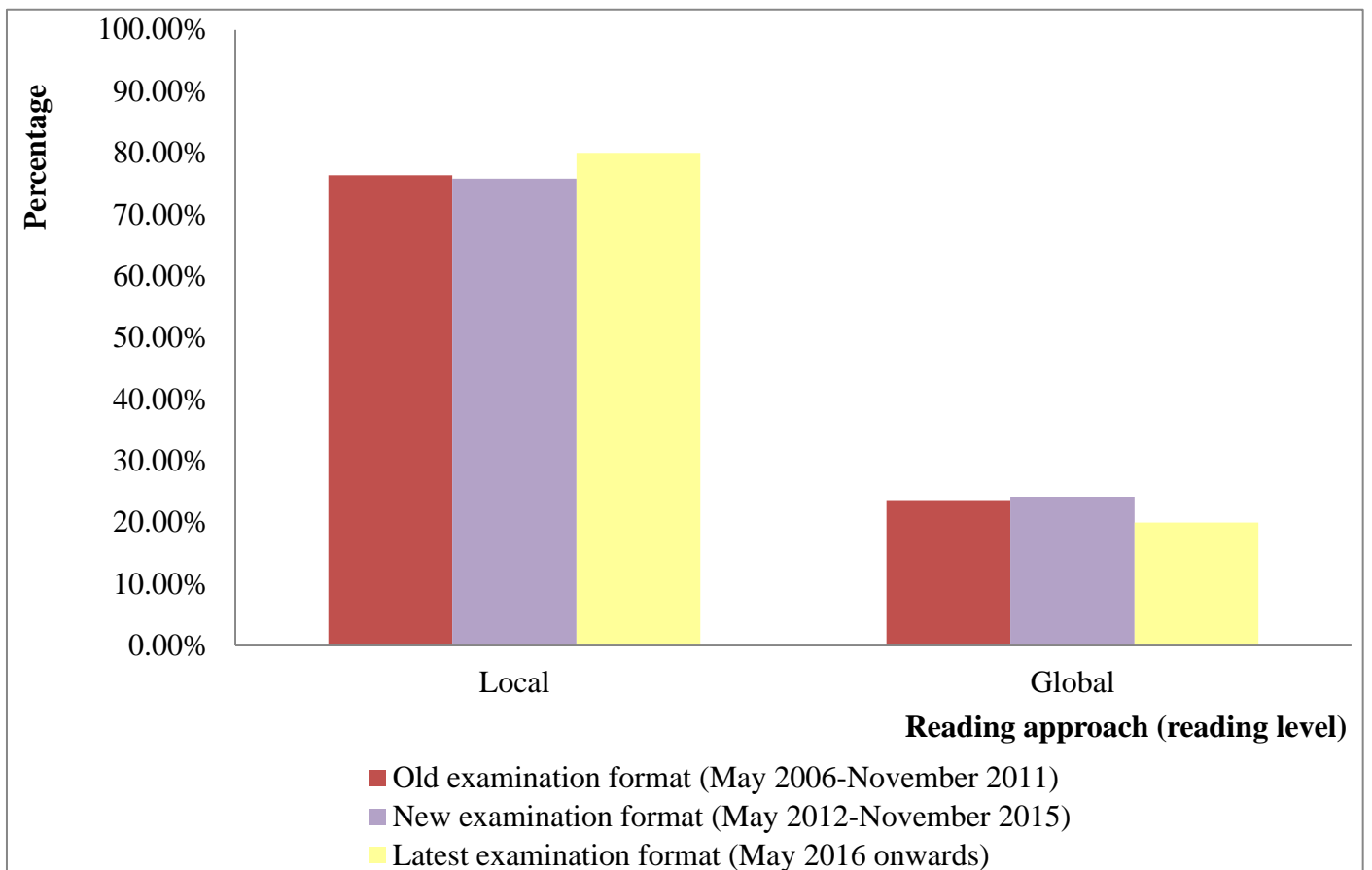


Figure 7j: Breakdown of reading items by reading level across examination formats

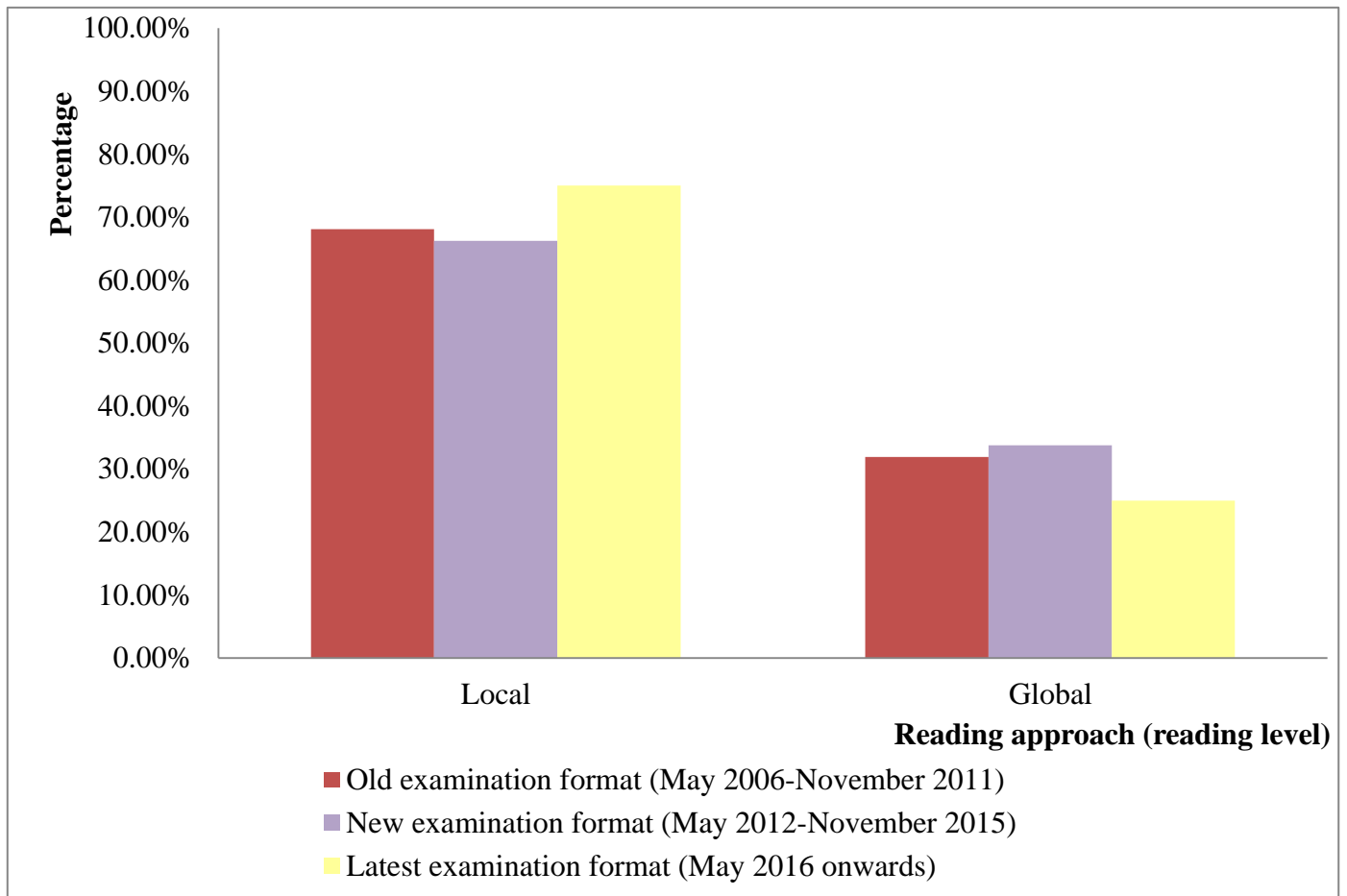


Figure 7k: Breakdown of reading scores by reading level across examination formats

	Old examination format (May 2006-November 2011)	New examination format (May 2012-November 2015)	Latest examination format (May 2016 onwards)	Row total
Score for local reading items	572	371	105	572 + 371 + 105 = 1,048
Score for global reading items	268	189	35	268 + 189 + 35 = 492
Column total	572 + 268 = 840	371 + 189 = 560	105 + 35 = 140	840 + 560 + 140 = 1,540 (Grand total)
Chi-square	3.95			
P-value	0.14			

Figure 71: Reading scores for local and global items across examination formats

In relation to reading types, 17.73% of items, accounting for 22.53% of the total score, require only expeditious reading (see Figure 7m). For example:

李总理提出了哪些学好华语的建议？怎样才能确保这些建议取得成效？

What suggestions did Prime Minister Lee offer for mastering the Chinese language?

What could be done to ensure the effectiveness of these suggestions?

(Q22, May 2006)

The wording of this item allows test-takers to match item prompts to the passage directly. By scanning the passage for the phrases ‘suggestions for mastering the Chinese language’ and ‘effectiveness of suggestions’, test-takers would be able to locate the answer which is explicitly stated in the first and second sentences of the

third paragraph. Lack of control on the time spent on each item of the examination, however, may mean that some test-takers use careful reading rather than expeditious reading when completing such items.

With regard to careful reading, 82.27% of the items, accounting for 77.47% of the total score, demand this type of reading (see Figure 7m), for example:

试解释(这个短语)在文中的意思:

从即将被对手淘汰的边缘拉回来 (第三段)

Explain the meaning of the following phrase:

Pulled the opponent clear back from the brink of elimination (Paragraph 3)

(Q24a, November 2006)

Answering this item involves careful reading to understand the propositional meaning at clause and sentence level.

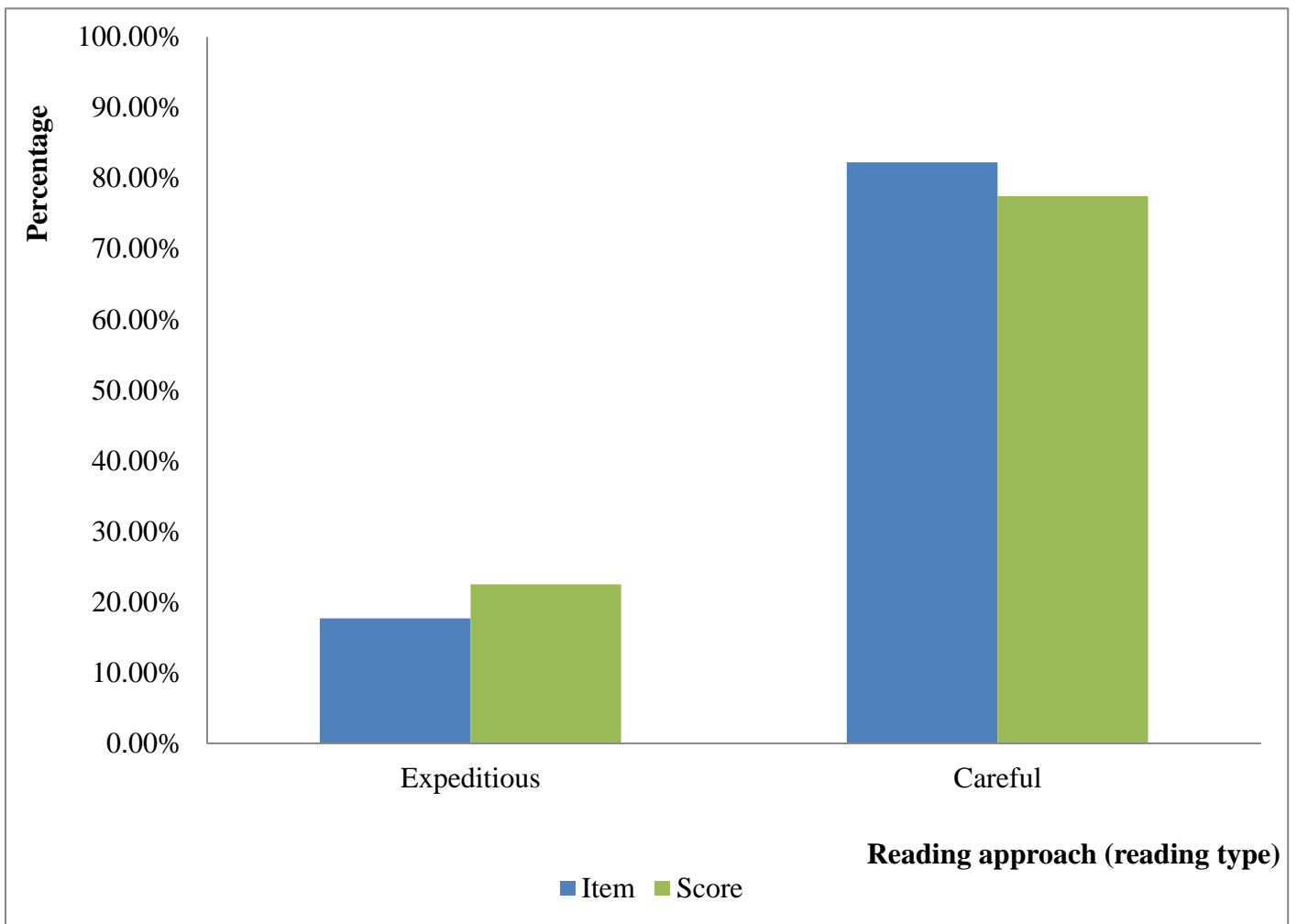


Figure 7m: Breakdown by reading type in the GCE 1162 reading examination
(May 2006-May 2016)

Figures 7n and 7o show that the breakdown of careful items is 81.67%, 83.33% and 81.67%, accounting for 76.67%, 79.82% and 72.86% of the total score for the old, new and latest group of examination papers respectively. Using the Chi-square test of independence, a Chi-square value of 3.79 is obtained (see Figure 7p) with a P-value of 0.15. The difference between reading types across examination formats is therefore not statistically significant at the 5% significance level. In other words, the reading types of items after the revisions made to the GCE 1162 examination remain relatively unchanged.

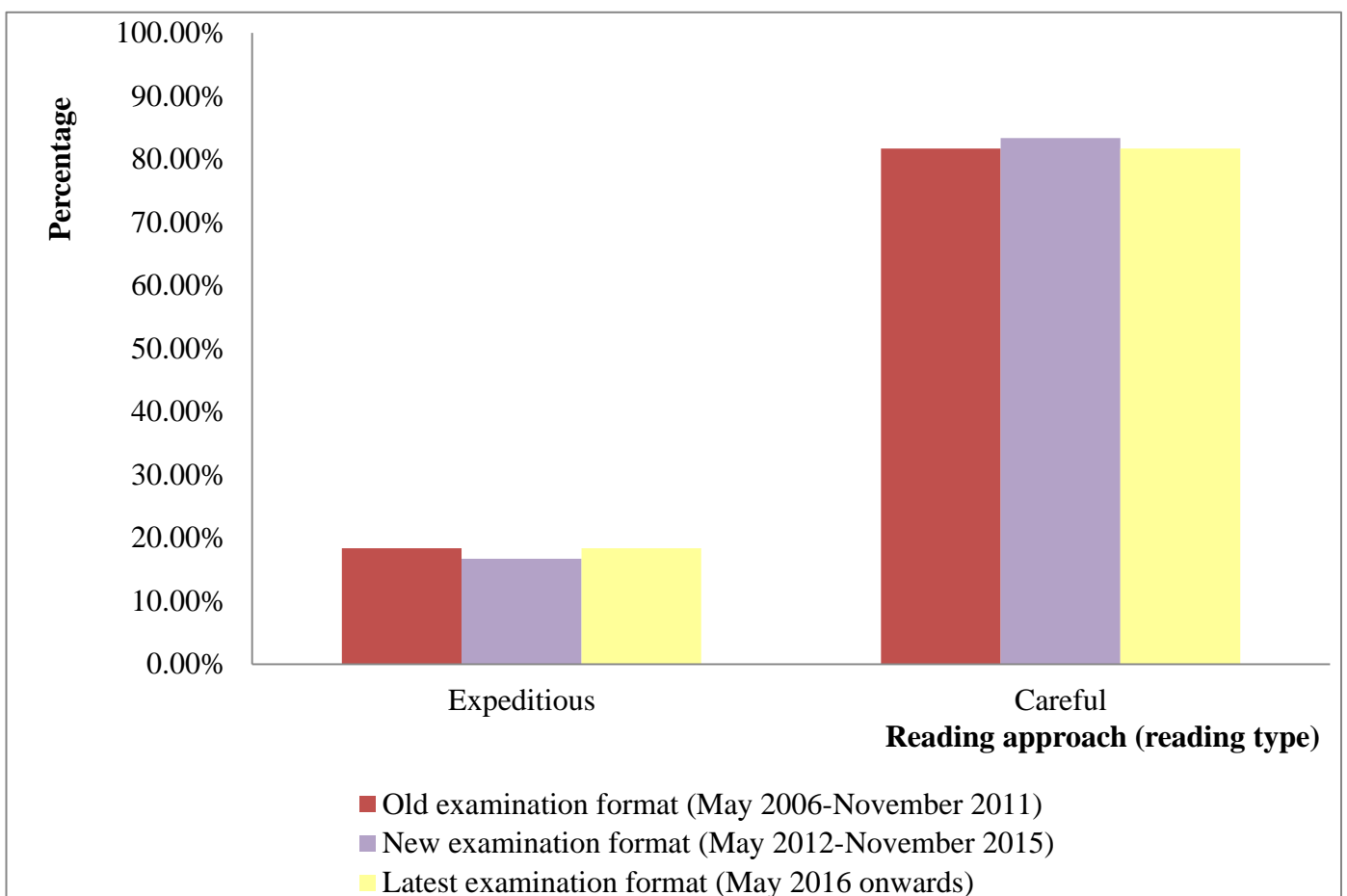


Figure 7n: Breakdown of reading items by reading type across examination formats

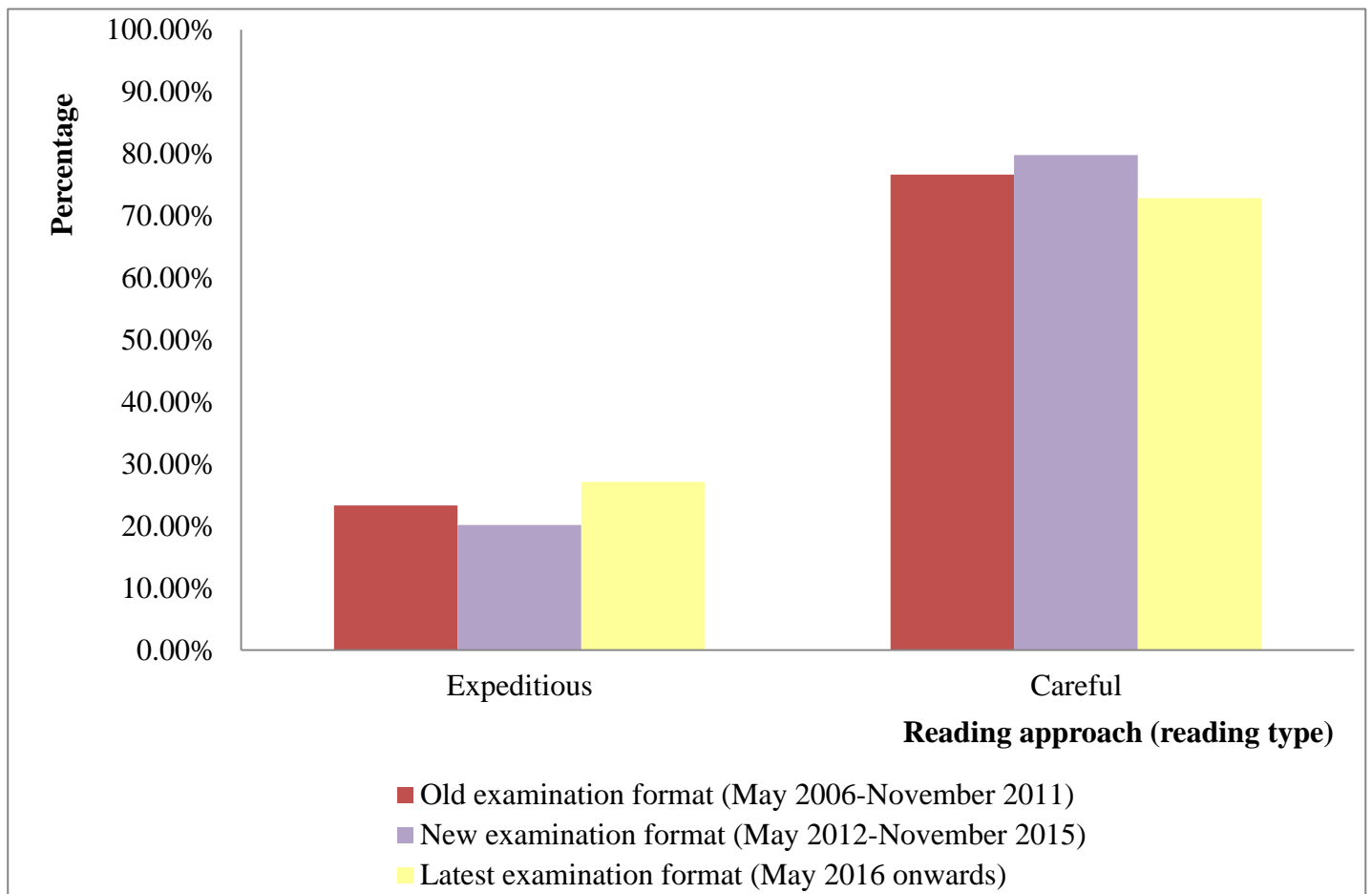


Figure 7o: Breakdown of reading scores by reading type across examination formats

	Old examination format (May 2006- November 2011)	New examination format (May 2012- November 2015)	Latest examination format (May 2016 onwards)	Row total
Score for expeditious reading items	196	113	38	196 + 113 + 38 = 347
Score for careful reading items	644	447	102	644 + 447 + 102 = 1,193
Column total	196 + 644 = 840	113 + 447 = 560	38 + 102 = 140	840 + 560 + 140 = 1,540 (Grand total)
Chi-square	3.79			
P-value	0.15			

Figure 7p: Reading scores for expeditious and careful items across examination formats

It is not always easy to determine the reading approach used by test-takers for each item unequivocally and reliably. As indicated previously in this section, under intense examination settings, test-takers may be compelled to read the whole text with a high level of attention, even several times, to answer a local expeditious level item. This is especially plausible when there is no time constraint for each item. In other words, while there is a representation of local, global, expeditious and careful level items, further research is needed to understand the optimal combination for the GCE 1162 reading examination. The inclusion of a separate computer-based expeditious reading paper, which will be performed under strict time constraints, has been suggested by some interviewees, including Xi who held high office at the MOE. Faced with the sheer volume of information to be absorbed in the Internet age, adolescents need to be able to skim, scan and search read to increase their speed of

reading and the setting of an expeditious reading paper is deemed useful by some interviewees. Practical implementation issues of having such a component in the examination still need to be addressed.

7.3.4 Item difficulty and discrimination

I have studied in detail the cognitive parameters of the GCE 1162 reading examination. Analysis of semi-structured interviews, documents and expert judgements were employed to evaluate whether the cognitive requirements of the examination are appropriate. This subsection provides a supplementary overview of the statistical analysis employed by SEAB during the GCE 1162 reading examination field tests. Although Lado (1961: 5) asserts that linguistic and not statistical analysis should be the major determinant of the content of a language examination, ‘statistical treatment has its place in the refinement of the test’.

Data gathered in the study suggest that SEAB uses both Classical Test Theory (CTT) and Item Response Theory (IRT) based methods to ascertain item difficulty and discrimination during field testing in order to refine the items in the actual examination. CTT and IRT are applied again at the a posteriori stage to assess the suitability of items tested in the GCE 1162 reading examination. CTT considers two main statistics, Facility Index (FI) and Discrimination Index (DI). The FI is an indication of the difficulty of an item, with a high facility index indicating an easy item and a low facility index indicating a difficult item. The statistics for FI are given by the formula:

$$FI(p) = \frac{\overline{X}}{X \text{ max}}$$

where $FI(p)$ = the facility index of an item

\overline{X} = the mean score obtained by all test-takers attempting the item

$X \text{ max}$ = the maximum score available on the item

It is desirable that items have an FI of close to 0.5 to provide the widest scope for variation among test-takers. Items that are too easy, for example with an FI of 0.8 or

above, or too challenging, for instance with an FI of 0.2 or below, will not provide much information since they reveal little about the varying levels of proficiency. The closer the item is to an FI of 0.5, the more it contributes to the measurement of the test-takers. It is, however, acceptable and common practice to have a few items with a high FI at the beginning of an examination to help test-takers build confidence and a few items with a low FI at the end to allow test-takers with the strongest abilities to distinguish themselves.

The DI is a measure of how test-takers perform on an item as compared to another measure of performance. Examination boards tend to use a measure of discrimination known as the point-biserial correlation or Pearson r , which is the correlation coefficient between the scores for an item and the scores for the total examination. The possible range of the DI is from 1.00 to -1.00. A DI of 1.00 indicates a perfect positive correlation between those who score high marks in the item and those who score high marks in the examination. Interpretations can be made based on the range of DI values—very good (>0.40), good (<0.39 to >0.30), fair (<0.29 to >0.20), non-discriminating (<0.19 to 0.00) and needs attention (<0.00) (Alagumalai, Curtis & Hungi, 2005). When items discriminate negatively, test-takers with the highest abilities overall are shown to be getting the items wrong while the weakest test-takers are getting them right. Removing these items, together with those which are non-discriminating, will improve test validity. It is worth noting the effect of extreme FI values on DI statistics, where reduced variance of these items necessarily lowers the ceiling values for item discrimination. DI can be calculated using the formula:

$$DI (r_{xy}) = \frac{\sum_{xy}}{NS_xS_y}$$

where $DI (r_{xy}) =$ the correlation between the item (x) and the test total (y)

$\sum_{xy} =$ the sum of the products of the deviations of the items and the totals

$N =$ the number of observations

$S_x =$ the standard deviation of the item

$S_y =$ the standard deviation of total score

In the formula above, y represents the test total. In some situations, y is replaced by the total score minus the score of item x . By excluding the item's value from the total score, the corrected value is used to mitigate the problem of overestimating the DI of an item. Other measures such as the total score for a section or an external score can also be similarly substituted.

Under CTT, FI and DI statistics aid the elimination of those items which, during field testing, are shown to have a very high success or failure rate or do not discriminate well. These indices also help in identifying non-functioning distractors in selected response items. The use of FI and DI statistics thereby improve the validity of test scores by increasing the number of items which adequately sample the identified constructs. The main shortcoming of CTT is that FI and DI statistics are sample-dependent—the values of the indices vary according to the level and spread of ability in the group of test-takers from which they have been obtained.

IRT-based models, on the other hand, provide an analysis of test items that are sample-independent so that the measurement of test-takers can be adequately equated across examination papers from different years. IRT-based models, however, assume uni-dimensionality, where there is only one underlying ability or trait being tested, and local independence of items, where test-takers' responses to one item are not dependent on a previous or subsequent item. The various IRT-based models express the relationship between item performance and test-taker ability using a one, two or three parameter logistic function. Rasch analysis, for example, applies one parameter analysis where item difficulty is the only parameter. Other models take into consideration additional parameters such as item discrimination, item location and a guessing factor for selected response items. It is unclear which IRT-based model SEAB uses for the GCE 1162 reading examination and whether a bank of calibrated items is assembled. Xi, a retired key member of personnel in the MOE, assures that while many of SEAB's operations remain confidential, SEAB does have a team of competent measurement and analytics officers to generate FI, DI and other statistics to assess item suitability using a number of in-house and external applications during field testing. In addition, officers from SEAB routinely interact with representatives from Cambridge Assessment to share best practice and to acquire the latest statistical and psychometric methods and tools.

7.4 Conclusion

This chapter, comprising IA and VA sections, has built an ABV of the cognitive parameters of the GCE 1162 reading examination. The VA section was divided into three subsections. The first subsection outlined the various dimensions of reading assessment and argued that the GCE 1162 reading examination focuses primarily on the dimension of text comprehension, possibly overlooking other dimensions such as multiple text reading for problem-solving, and reading volume and interest. This discussion was followed, in the second subsection, by a qualitative and quantitative evaluation of the cognitive demands elicited by items in the GCE 1162 reading examination. The overall findings suggest that although the examination has a range of items catering to test-takers of varying language proficiencies, there is an inadequate representation of HOTS items. In addition, items from the apply, analyse and create cognitive levels are underrepresented, making up less than 5% of the total score of the GCE 1162 reading examination paper. The third subsection on reading approaches established that the different reading levels, namely local and global, and reading types, namely expeditious and careful, are covered appropriately in the examination, although further research needs to be undertaken to understand the optimal combination of reading approaches for the examination. The fourth subsection provided a short account of the CTT and IRT-based methods used by SEAB in field testing to ascertain item difficulty and discrimination before the actual examination. Comparisons across the old, new and latest group of examination papers indicate that changes in cognitive parameters across examination formats are not statistically significant, calling into question the alignment between the measurement objectives of the examination and learning objectives in the syllabus.

Drawing on evidence in the three subsections, another conceivable threat to validity was revealed. It was found that the reading constructs measured by the examination are limited by the reading dimensions, cognitive levels and reading approaches sampled, implying that scores in the examination might not be generalizable to real-world reading contexts. In summary, the evaluation status (Shaw & Crisp, 2012) assigned to each assumption underlying the cognitive parameters inference are listed in Figure 7q.

Assumption		Provisional evaluation status based on semi-structured interview, document analysis and expert judgement data
1.	The examination takes into account the different dimensions of reading assessment (e.g. text comprehension, knowledge and application of language and literature, multiple text reading for problem-solving, and reading volume and interest).	Plausible rejection
2.	There is adequate representation of lower-order thinking items (LOT).	Accepted with concerns
3.	There is adequate representation of higher-order thinking items (HOT).	Plausible rejection
4.	There is adequate representation of items at each specific cognitive level (remember, understand, apply, analyse, evaluate and create).	Plausible rejection
5.	The examination takes into account different reading levels (local and global).	Accepted with concerns
6.	The examination takes into account different reading types (expeditious and careful).	Accepted with concerns
7.	Statistical analyses are employed in field testing to refine items in the actual examination.	Accepted with concerns
8.	There is alignment between the measurement objectives of the examination and the learning objectives in the syllabus.	Accepted with concerns
9.	The examination assesses constructs that are relevant to real-life reading contexts beyond the syllabus.	Plausible rejection

Figure 7q: Provisional evaluation status of the assumptions underpinning the cognitive parameters inference relating to Singapore's GCE 1162 reading examination

It is useful to remember that all cognitive parameters of the GCE 1162 reading examination are mediated by the contextual parameters of the passages and items used. Contextual parameters are the performance conditions under which reading assessment takes place. Examples of contextual parameters include the type of items, the complexity of selected texts and the design of mark schemes, which both individually and in combination, are likely to affect the product and process of reading. A full discussion of these factors is presented next in Chapter 8.

Chapter 8 Contextual parameters

8.1 Introduction

The last decade of the twentieth century saw a growing interest in the importance of context in the field of testing and assessment (Weir, 2005). An extensive discussion of the wider contexts of Singapore and the specifications and administration of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) has been provided in Chapters 3 and 5. This chapter focuses instead on the contextual parameters pertaining to the GCE 1162 reading examination paper per se. Reading, during an examination, takes place within predetermined contextual parameters and not in a vacuum or an item and text neutral position as purely theoretical works sometimes assume (Khalifa & Weir, 2009). It is thus crucial that the GCE 1162 reading examination is operationalized within the contextual parameters deemed suitable by stakeholders and test-takers.

The chapter illustrates five contextual parameters which interviewees and subject matter experts (SME) in the study have foregrounded as the most pertinent in influencing reading performance in the GCE 1162 reading examination. The contextual parameters to be explored can be categorized into item-related and text-based parameters. Item-related parameters include item type and mark scheme; while text-based parameters comprise discourse mode and text purpose, propositional content, and readability. Altering these contextual parameters will affect the performance of test-takers and possibly the way they construe the examination. The five contextual parameters that form the assumptions in the Interpretive Argument (IA), and which are defined in the next section, inform the structure of the validity argument (VA).

8.2 Interpretive argument

The literature review on validity in Chapter 1 proffers the contention that recent progress in thinking about validation involves the organizing of evidence into a persuasive argument to validate an examination. Kane's (2009, 2006) argument-

based approach to validation (ABV), consisting of an IA and VA, is now widely influential in the field of testing and assessment (Newton, 2017a). The IA centres on an inference, which in this chapter refers to the contextual parameters. The contextual parameters inference addresses the fourth sub research question in the study, namely, ‘Are the characteristics of the test items and passages appropriate and fair?’ The accompanying claim would be that the characteristics of test items and passages in the GCE 1162 reading examination are appropriate and fair. The IA pivots on the following assumptions:

1. A variety of suitable item types is employed to assess reading constructs.
2. Mark schemes are well defined for differentiating the quality of answers.
3. There is adequate representation of passages of different discourse modes and text purposes.
4. There is adequate representation of passages with different propositional content.
5. Passages in general possess literary merit.
6. Passages in general are of a suitable readability level.
7. The contextual parameters of the examination support the assessment of constructs that are relevant to real-life reading contexts beyond the syllabus.

The list of assumptions is based on the five contextual parameters of item type, mark scheme, discourse mode and text purpose, propositional content, and readability. In the VA section below, an outline of the main research and literature related to each parameter is provided before each parameter is exemplified in relation to the GCE 1162 reading examination. Supporting evidence and rebuttals derived from the examination of semi-structured interviews, documents and expert judgements are used to strengthen the VA.

8.3 Validity argument

8.3.1 Item type

It is conceivable that different item types permit the measurement of different aspects of a reading construct (Alderson, 2000). It is therefore necessary that test designers

are aware of what various item types are capable of assessing. Such awareness would aim to ensure that the item types chosen for an examination do not unduly constrain the range of reading dimensions, approaches and cognitive levels which the item setters test. Some item types, for example, summary which involves a writing component, might also conflate reading proficiency with writing ability. Further, the item types and answer requirements for an examination must be made known to all test-takers to ensure fairness and equity. Regardless of the item type or types chosen, there is no one best method for testing reading; hence, employing a number of suitable item types in a high-stakes examination, in so far as it is practical, helps reduce threats to validity. Interviewees maintain that a suitable range of item types is currently adopted in the GCE 1162 reading examination although as touched upon in Chapter 6, the inclusion of item types such as information transfer items could better equip test-takers with the reading abilities they will need for performing in a real-world context.

A variety of item types has been employed by test designers to assess second language reading comprehension. The more commonly used item types include multiple-choice questions (MCQs), matching, true/false items, information transfer tasks, cloze and gap-filling, summary and short-answer questions (SAQs). Item types can involve a selected response or a constructed response. For a selected response item, test-takers choose a response from a list of options provided by the test developer. As the answers are pre-determined, selected response items can be scored automatically by a machine. Examples include MCQs and matching items. A constructed response item, in contrast, requires test-takers to supply the answer themselves, for example by writing a word, sentence or short paragraph onto an answer sheet. The strengths and limitations of the four item types employed in the GCE 1162 reading examination, specifically the MCQs, gap-filling test, SAQs and open-ended questions are expounded below.

MCQs, which form a type of selected response item, are frequently used for assessing students' reading comprehension abilities in large-scale summative testing. MCQs rapidly gained popularity in the mid-twentieth century when data processing machines and scanners enabled MCQ test items to be marked automatically. MCQs consist of a question stem, the key, which is the correct answer, and distractors. In

the GCE 1162 reading examination, test-takers choose from four answers, of which only one is correct. A correct answer earns points toward the total mark. Test-takers receive no credit but are not penalized for a wrong choice. There are 10 MCQs accounting for 20 marks which represent 28.57% of the total score for Paper 2 (see Figure 4g in Chapter 4).¹

MCQs are often favoured by test designers not least for their ease of marking. They also exhibit almost complete inter- and intra-marker reliability as the marking process is totally objective. When well-constructed, MCQs can test a wide range of cognitive levels in a more controlled way than is possible through constructed response items. MCQs also lend themselves to statistical analysis, whether it is the calculation of Facility and Discrimination Indices in Classical Test Theory or the Bookmarking of items through Item Response Theory.

Despite the ubiquity of MCQs in reading assessment, their drawback cannot be overlooked. Even for experienced item setters, developing a sufficient number of high quality MCQs for a passage is a skilled and time consuming process. To design plausible but incorrect distractors that will differentiate between the weaker and more proficient reader is far from simple. Moreover, while marking MCQs is an objective process, setting and selecting items and deciding on the key can be a matter of subjective judgement. Although interviewees and SMEs acknowledge that MCQs in the GCE 1162 reading examination are generally adequately designed, they detect the following three problems.

The first problem is flawed distractors and keys. Some distractors are likely to be non-functioning,² hence diminishing the difficulty of the item. There are also examples of possible mal-functioning distractors³ that may penalize more competent readers. In addition, there are MCQs with more than one possible key and a small

¹ In addition to the MCQs in Section 2, there are 10 gap-filling items in Section 1 of the GCE 1162 reading examination that provide multiple choices for test-takers to select from.

² Non-functioning distractors can be defined as options that are chosen by fewer than 5% of the test-takers (Haladyna & Downing, 1993). In the current study, however, I rely on expert judgement to gather examples of non-functioning distractors.

³ Mal-functioning distractors can be defined as those with a negative Discrimination Index, that is, when low-performing test-takers answer a specific item correctly more often than the high scorers.

number where test-takers may be able to determine the key without recourse to the text. For example:

自信是前进的动力，主要在于它能_____

- (1) 避开外来的障碍
- (2) 激发奋斗的勇气 (正确答案)
- (3) 使人认清前进的目标
- (4) 令别人对自己有信心

Self-confidence is the key ingredient for progress, mainly because it_____

- (1) helps you overcome external barriers
- (2) spurs you on to greater heights (Key)
- (3) leads to clearer objectives
- (4) inspires confidence in others

(Q19, Specimen Paper 2016; Q14, May 2007)

The passage is about a boy who gave up on himself and constantly misbehaved because the people around him labelled him as a wilful and disobedient child. He finally regained his self-confidence with his stepmother's encouragement, proving others wrong and achieving success. SMEs agree that there is no unequivocal answer—options 1, 2 and 4 all seem acceptable. Failure to understand the text as the item setter intended can result in an 'incorrect' response, even if the test-taker's interpretation is completely reasonable; and unlike constructed response items, MCQs do not allow test-takers to explain their answers and potentially receive credit.

A second problem detected is items that do not adhere to the general principles of test design. Urquhart and Weir (1998) summarize a list of guidelines for setting items. As stated in the list, 'questions should not contain harder vocabulary than the text' and that 'if the candidate understands the text he should be able to answer the question' (Urquhart & Weir, 1998: 152). SMEs highlight some MCQs that may be in conflict with these requirements, for example:

这篇短文的用意是什么？

(1) 鼓励公众踊跃出席“海啸摄影展”

[...]

(4) 呼吁公众支持“分享愉悦，捐赠玩具”的活动 (正确答案)

What is the main purpose of the passage?

(1) To encourage *ebullient participation* in the Tsunami Photography Exhibition

[...]

(4) It makes an *earnest appeal* to support the Share a Joy, Donate a Toy Drive (Key)

(Q11, Specimen Paper 2016; Q20, May 2006, emphases added)

As seen above, a test-taker's wrong answer may be due to a lack of comprehension of the item options rather than a lack of comprehension of the passage.

Furthermore, some stems and options unwittingly provide clues to the correct answer, such as when options are not parallel in format; the key is the longest option with the most difficult words; or the key contains words or phrases from the text. For example:

这篇短文的主旨是什么？

(1) 爱是一种神奇的力量。(正确答案)

What message does the passage convey?

(1) Miracles happen in the presence of love. (Key)

(Q20, November 2012)

This MCQ appears to be an understand item that requires test-takers to condense the text into its primary message; the key, however, undermines the effectiveness of the item by lifting the concluding sentence ‘The most astonishing miracles happen in the presence of love’ verbatim from the text. Given the challenges of writing good quality MCQs, SMEs stress the need for items to be validated through a number of trialling phases. The selection and training of item setters must also be a rigorous and transparent process.

A third problem is issues with authenticity. Answering MCQs, by nature, is not an authentic task. A reader is rarely presented with a text in real life and made to choose

from among four options to indicate understanding. MCQs are problematic also because under normal circumstances item setters and markers are unable to discern why test-takers respond the way they do. Test-takers may arrive at the correct answer by eliminating wrong options instead of choosing the right one in the first place; or they may simply select a random answer out of the four options and still have a 25% chance of getting the item right. Similarly, test-takers may get an item wrong for the ‘right’ reason. These problems could be mitigated by requiring test-takers to state their reasons for selecting an option but then the practical advantage of MCQs in terms of marking would be greatly diminished.

Next, I examine gap-filling tests which form another item type used in the GCE 1162 reading examination. Gap-filling tests are closely related to cloze tests. Cloze tests feature a passage where every n^{th} word is deleted after allowing a few sentences of introduction. The deletion rate is mechanically set, with n usually between every 5 and 11 words. Test-takers are required to restore the words that have been deleted, although credit is often given if a word provided is not the one originally omitted but makes sense in the gap. Cloze tests were initially used to determine text readability for English as a first language of instruction. They soon became prevalent in the 1970s in second language curriculum and assessment. In gap-filling tests, also referred to as rational cloze tests or selective deletion cloze tests, item setters choose which words to delete on a rational basis. These words can be both content and function words. A common variant of the gap-filling test—the multiple-choice gap-filling test is used in the GCE 1162 reading examination. Four options are given for each of the ten gaps in the given passage with each correct answer being allocated one mark, accounting for a maximum of 10 marks (or 14.29% of the total score for Paper 2, see Figure 4g in Chapter 4). The advantages and disadvantages of test-takers not supplying but selecting the missing words are similar to those explored earlier under the MCQ item format.

With cloze tests, as an item setter has no control over which individual words are deleted once the starting point is chosen, it is not possible to predict with confidence what aspect of the reading construct each gap and item will measure. In contrast, an item setter has the freedom to decide which words to remove for gap-filling tests. It would appear, therefore, that the item setter could not only design gap-filling items to

measure vocabulary mastery but also items that target the understanding of the entire passage, by, for example, omitting words which are essential to the main propositions. SMEs, however, are of the view that gap-filling tests essentially assess lexis and syntax, predominantly involving careful reading at local level. This SME view is mirrored by the data collected from interviews with student interviewees. All four student interviewees describe that they focus largely on decoding at the word or immediate constituent level when attempting gap-filling tests and that they complete items without paying much attention to the text as a piece of connected discourse. As investigations carried out by Bernhardt (2011), Khalifa and Weir (2009) and Alderson (1978) suggest, comprehension ability is not as critical as lexical and grammatical sensitivity in order to succeed at the gap-filling test. As Khalifa and Weir (2009: 90) maintain, ‘on its own, therefore, a test of the ability to replace single words is likely to be an insufficient indicator of a candidate’s reading ability because of the restricted processing involved’.

Although the gap-filling tests rarely seem to involve macro-level reading processes, interviewees notice that GCE 1162 reading examination test-takers are especially weak when it comes to the gap-filling items. Interviewee Lambda explains:

Weaker students may have a vague idea of what the words mean [...] but they are unlikely to get the answer right as they are unable to grasp the nuances among the four given options [...] and because they hardly read they are also unaware that some words tend to co-occur with others [...] They lack the language sensitivity or intuition that competent and native readers possess.

Besides measuring the test-takers’ vocabulary size or breadth, gap-filling items also provide an indication of their vocabulary depth, the dimension of vocabulary knowledge that pertains to the quality of the knowledge that individuals have about words (Schmitt, 2014). Test-takers, as interviewee Lambda points out, may have what has been described as a ‘limited unclear idea of what a word means’ (Read, 2004: 211) in that they are unable to appreciate the connotations, level of formality and collocation pattern the word has and thus fail to select the correct response. For example:

逢年过节，甘榜里就充满着热闹的节庆气氛，摊贩的生意也特别 (1 兴盛 2 繁荣 3 兴旺 (正确答案) 4 繁盛)。

During the festive season, the kampongs are filled with mirth and merriment and business is especially () for the hawkers and peddlers.

(Q6, November 2011)

All four options (1) 兴盛 (*xingsheng*), (2) 繁荣 (*fanrong*), (3) 兴旺 (*xingwang*) and (4) 繁盛 (*fansheng*) can be loosely translated as ‘prosperous’. Test-takers, however, need to have the specific knowledge that in Chinese, options 1, 2 and 4 are habitually used to describe countries, economies and large-scale businesses and thus, option 3 is the most suitable word.

Taking the apparent difficulty of gap-filling tests into consideration, interviewees agree that the words tested in the GCE 1162 gap-filling tests should remain largely those taught in the curriculum. According to interviewee Iota, a curriculum specialist, the testing of words used in the curriculum also ‘serves to motivate students to look through their textbooks before the examination’. Some interviewees are receptive to the inclusion of one or two words beyond the curriculum being tested in the gap-filling tests to reward test-takers who read extensively outside the classroom. The fast growing field of corpus linguistics will greatly facilitate the selection of which words outside the curriculum should be included. Test designers are now able to efficiently determine, using new computer corpus software, the frequency and range of words that appear in the Chinese newspapers, magazines and books which Singaporean adolescents are encouraged to read.⁴

I have examined comprehensively the two types of selected response items utilized in the GCE 1162 reading examination, namely MCQs and the gap-filling test. Attention is now turned to two types of constructed response items, namely, the SAQ and the open-ended question. The SAQ, which Bachman and Palmer (1996) classify

⁴ The *Frequency Dictionary of Daily Chinese Words Encountered by Singapore Students* (Goh, Lin & Zhao, 2013) published by the Singapore Centre for Chinese Language is a notable example of a frequency dictionary based on a corpus of words that appear in the Chinese newspapers, magazines and books which Singaporean adolescents are encouraged to read.

as a limited production response type, is a semi-objective item type which requires test-takers to construct a brief response. The length of an answer may vary from a word or phrase to a few sentences. In the GCE 1162 reading examination, SAQs, which carry 2 to 4 marks each, can usually be answered in a couple of sentences. The mark value of each SAQ and the number of lines provided on the answer script provide indications of the expected length of response. For example:

父亲为什么会改掉喝米酒的习惯？（2分）

Why did the author's father quit drinking rice wine? (2 marks)

(Q21, November 2012)

The answer required is a single sentence stating that the author has written his father a letter with a list of compelling reasons as to why he should quit drinking and encourages him to do so.

Unlike MCQs, the answers for SAQs need to be sought rather than chosen, making it harder for test-takers to get credit from guessing. SAQs can be designed so that a fairly large number of items covering a broad range of cognitive processes can be included within a relatively short testing time. By identifying the central ideas and information of a text through expert judgement or recall protocols, and mapping SAQs onto them, item setters can avoid testing trivial details. A cause of worry with SAQs, however, is marking reliability. For large-scale assessments like the GCE 1162 reading examination, a large group of markers is often involved. As test-takers' responses become more complex it is difficult to determine the quality of a response and to assign marks with zero variation. The problem of objectivity grows more noticeable with open-ended questions as seen below. Marker training, standardization practices and detailed mark schemes go a long way to mitigating the issue. Another main concern is that answering SAQs involves writing and this may contaminate the construct being measured. Test-takers may be able to comprehend the texts but are less capable of expressing themselves in writing. Grammar, spelling and punctuation mistakes may also cause test-takers to be penalized.

In comparison with narrow SAQs, open-ended questions often yield answers that are more varied and less predictable. Zeta terms these open-ended questions, which are

usually the last one or two items for each of the passages in Section 3, as ‘mini-essay questions that are each worth 4 to 5 marks’. For example:

“和其他国家的人比较起来，我们真的好幸福啊！”你同意这种说法吗？为什么？（5分）

‘We are very fortunate compared to people from other nations!’ Do you agree with the above statement? Why? (5 marks)

(Q25, May 2010)

Together with the SAQs, open-ended questions make up 40 marks in the GCE 1162 reading examination (or 57.14% of the total score for Paper 2, see Figure 4g in Chapter 4). An attraction of this item type is that it easily allows for the testing of higher-order thinking skills such as analysing, evaluating and creating. Open-ended questions enable interaction between test-takers and the given texts, encouraging the integration of different propositions, experiences and knowledge. The main drawback, as some interviewees argue, is that a significant amount of difficulty is added when answering open-ended questions as test-takers have to write in their own words rather than use language supplied in the text. It is worth noting however that although writing tends to be traditionally viewed as a source of construct contamination in reading examinations, a recent trend in assessment is to integrate the two skills by asking test-takers to respond in writing after reading the given text or texts (Weigle, 2004).

The rationale behind the movement toward skill integration is to enhance authenticity. This objective is commendable but the plethora of possible answers from test-takers will necessarily complicate the scoring procedures. For the GCE 1162 reading examination, all open-ended questions and SAQs are judged by at least two markers. The scores are then written in the margins of the answer scripts; the second marker will be able to see the first marker’s score. The mean score is taken as the final mark awarded if the difference between the two scores is 2 marks or less. If there is a major disparity in the marks awarded, a third marker will be assigned whose decision is final. Not all interviewees, though, agree with the extent to which the procedure is efficacious. Open-ended questions are not items that demand a single correct

response, nor are they items where responses are all acceptable or of comparable quality. Interviewees speak of the need to develop a more detailed mark scheme. Mark schemes from the GCE 1162 reading examination will be analysed in the next subsection alongside Ahmed and Pollitt's (2011) general taxonomy of mark schemes.

8.3.2 Mark scheme

The next contextual parameter that is drawn on in this chapter is the mark scheme. Markers often encounter a wide range of answers and a mark scheme provides instructions, advice and support on how marks are to be awarded. Mark schemes ensure a more objective, consistent and reliable assessment of test-takers' responses. Findings from a study commissioned by the Qualifications and Curriculum Authority (QCA) in England (Pollitt, Ahmed, Baird, Tognolini & Davidson, 2008: 6) suggest that 'as a priority, training in how to write mark schemes will probably lead to more immediate improvement in exam validity than will any other measure.' In a similar vein, interviewee Zeta states:

Item writers convey their requirements to markers through a mark scheme. In a large-scale national examination [like the GCE 1162 reading examination], there are many markers involved [...] there is no opportunity for communication between the markers and item writers [...] thus, a mark scheme is the primary way in making certain that marking is done consistently across scripts and across markers. A detailed mark scheme is especially helpful for the open-ended questions in Section 3 [...] a vague mark scheme is likely to lead to impressionistic and subjective marking of these open-ended questions.

Eta makes a similar observation:

I would say that the mark scheme, especially for the open-ended questions, is open to interpretation [...] I suppose I wasn't the only one facing this situation during [the] marking [of the GCE 1162 reading examination scripts] [...] We informed the presiding examiner of our group when there were doubts about the model answer provided [...]

and there was a review process but I don't know what happened with the scripts that we had already marked or the scripts that other groups had marked [...] I wonder how widespread this problem is.

First, the mark scheme design for the GCE 1162 reading examination should ideally begin with a clear understanding of what it means to be proficient in Chinese as a second language (CL2) reading. Syllabuses, conventionally, contain a clear delineation of what students are expected to attain during their course of study which in turn defines the constructs of the examination. In Chapter 5 on specifications and administration, I argued that there is a considerable degree of ambiguity surrounding the constructs of the GCE 1162 reading examination. The lack of a clear and detailed explanation of the reading constructs is a weakness in the validity chain which in turn affects the quality of the mark schemes.

Second, to appraise the effectiveness of the GCE 1162 reading examination mark schemes, test designers need to determine how accurately they predict and describe the *Outcome Space* (Ahmed & Pollitt, 2011). The Outcome Space represents all responses produced by test-takers to an item, both anticipated and unexpected. A good mark scheme written in advance assists markers in distinguishing the quality of anticipated responses and how to award marks accordingly. Even then, the mark scheme may still need to be revised throughout the marking process to take into account unexpected answers that warrant credit. This subsection will now examine in detail the GCE 1162 reading examination mark schemes with reference to the Outcome Space.

Sections 1 and 2 of the GCE 1162 examination comprise a multiple-choice gap-filling test and short passages with corresponding MCQs, which naturally restrict the number of possible responses. Although a mark scheme stating the correct options will suffice for MCQs, the Outcome Space must still be considered carefully. Distractors should be based on common errors selected from the predicted Outcome Space had the questions been set as constructed response items. Moving to short-answer (SAQ) and open-ended items in Section 3, the emphasis shifts from whether the answer is right or wrong to the *quality* of the response. Additionally, as the items in Section 3 range from 2 to 5 marks each, markers may have to decide if an answer

demonstrates full understanding of the text or deserves only partial credit. Hence, the mark scheme needs to be well thought-out. The following example⁵, a Section 3 SAQ, has been chosen to illustrate possible inadequacies in the GCE 1162 reading examination mark schemes.

试解释（这句话）在文中的意思：

一切荣耀，都是短暂的，最后都敌不过死亡。（2分）

参考答案：所有的光荣和成就都是一时的，（1分）因为人终究会死去（1分）。

Explain the meaning of the following sentence:

All glory is fleeting as we are equal in the presence of death. (2 marks)

Model answer: Material success and achievements are ephemeral (1 mark) as we all have to die someday, taking nothing with us (1 mark).

(Q28b, May 2010)

Looking at the mark scheme, it is not apparent how much of this model answer a test-taker must write to gain full marks. Would the response: ‘Fame is ephemeral as we all have to die’ deserve 2 out of 2 marks? In a points mark scheme like this, every point is of the same value and hence of equal importance. Interviewees notice that test-takers, therefore, often copy or write as much as they can for there is no credit for summarizing or stating the most crucial points and that high ability test-takers are sometimes penalized for condensing information. It is also unclear what are considered acceptable substitutes for ‘glory’ and ‘fleeting’. Would ‘riches’ and ‘impermanent’ be rejected? The mark scheme would benefit from stating the criteria for awarding marks and listing examples of good and poor responses.

Perhaps more problematic are the model answers for the open-ended items:

“如果每个人都能把反省提前几十年，便有 50% 的人可以让自己成为一个了不起的人”，你同意这个说法吗？为什么？（5分）

参考答案：我同意这个说法。每个人都有潜力去取得成功，不过不是每个人最终能做到。我觉得反省自我和正确的生活态度是关键，及时反省可让我们在未

⁵ Official mark schemes for the GCE 1162 reading examinations are not available to the public. Examples used in this study are written by SEAB-approved publishers based on official mark schemes.

来避免曾犯的过错。我听过一则故事，有个前囚犯在狱中反省自己的过错，出狱后不但不再犯错，还成功创业。可见，他有成功的潜力，如果在年少时便懂得反省，可能会成为更了不起的人。（答案合理即可）

‘If we reflect on our goals and actions earlier, half of us could become remarkable people’. Do you agree with the author’s point of view? Why? (5 marks)

Model answer: I agree with the author. Everybody has the potential to be extraordinary but not everyone eventually succeeds. A lack of self-reflection and positivity can stop us from achieving our goals. Being constantly introspective prevents us from committing the same mistakes. There is a story about an ex-convict who reflected on his mistakes when he was in prison. When he was released, he turned his life around and even started a flourishing business. It is evident that he has the capacity for success, had he reflected earlier he could have accomplished even greater things. (Or any suitable answer)

(Q30, May 2014)

Schemes of this kind offer markers little assistance. It is left almost entirely to the markers to decide what constitutes a ‘suitable answer’. Omicron comments:

The mark scheme fails to make clear how the 4 or 5 marks are awarded. Does stating a claim get 1 mark? Does an example get 1 or 2 marks? [...] Do markers look out for the strength of an argument, the relevance of an example or how well an answer is articulated? [...] [Other than the mark scheme] markers are shown a few good and poor responses selected from actual scripts but marking open-ended items can still be rather subjective [...] and as markers mark really fast, issues may not get flagged up.

A more adequate mark scheme could be more specific about what gains credit. Consider, for example, student Tau’s suggestion:

For reading comprehension worksheets and tests, our teachers at school mark our answers [for the open-ended questions] using PEEL [Point, Evidence, Explanation and Link], usually a mark is awarded for each of

these components [...] We must show understanding of the key phrases in the item stem, evidence given must be appropriate and all components must form a cogent whole [...] we are told that our answers will be judged using these criteria. Perhaps the [GCE 1162] reading examination mark scheme could adopt this method too.

Tau's recommendation is essentially a points mark scheme. A points mark scheme, though not without its problems, is more functional than the available model answer listed above, which fails to make clear how much of the model answer a test-taker must provide to gain the available marks. According to Ahmed and Pollitt's (2011) general taxonomy of mark schemes, a better mark scheme would not only try to list acceptable answers but also unacceptable responses. At the highest level of the taxonomy, test designers give due consideration to all possible answers and state the governing rules and principles for distinguishing between good and poor responses. For instance, Pi who has marked GCE 1162 reading examination scripts speaks of 'awarding marks not only to content but also quality', 'answers that are poorly expressed with several grammatical and spelling errors will be penalized'. The mark scheme for open-ended items must hence first contain a statement or a rule for awarding marks for expression. Next, referring again to Q30 from the May 2014 paper as an example, the best mark schemes should state explicitly the principles by which markers must abide, as shown in the mark scheme that follows:

The first principle is to credit answers that state opinion (agree/disagree) and display a clear understanding of 'reflection' as giving serious consideration to life and its meaning, and one's actions and mistakes.

Point and explanation (2 marks). No credit for opinion without elaboration.

A second principle is to look for one relevant example. An example in support of the statement must demonstrate how reflection can have a life-altering effect on a person. An example against the argument must undermine the significance of reflection. Example (2 marks), partial credit (1 mark) for an example that proves reflection is a good/unproductive habit but does not highlight/discredit its life-changing quality. Link (1 mark).

As demonstrated by the examples above, marks schemes, especially those for SAQs and open-ended items, if inadequately designed can greatly compromise the quality of marking, thereby increasing threats to validity. Ahmed and Pollitt (2011) contend that it is futile to design test items of quality if an equal amount of care is not invested in the design of mark schemes to ensure that items are marked consistently and fairly. An effective mark scheme offers a good prediction of the Outcome Space which a real group of test-takers will produce. In addition, the mark scheme is not only specific about what gains credit but also provides markers with clear governing principles for discriminating between good and poor responses.

In sum, I have presented in this section supporting evidence and rebuttals pertaining to two core item-related parameters, namely, item type and mark scheme. The following subsections continue to construct the VA with reference to text-based parameters, the focal points being discourse mode and text purpose, propositional content and readability.

8.3.3 Discourse mode and text purpose

Texts are written under different circumstances and for a variety of purposes. To obtain a meaningful level of analysis, texts can be classified according to their discourse mode. Common discourse modes include narrative, expository, argumentative, functional and descriptive forms. Discourse modes ‘have a particular force and make different contributions to a text’ (Smith, 2003: 7), affecting it linguistically, both structurally and stylistically. Another classification relates to the overall text purpose or dominant intention (Weigle, 2002). Texts can be written primarily for metalingual mathetic (intended to learn) referential (intended to inform), conative (intended to persuade or convince), emotive (intended to convey feelings or emotions), poetic (intended to entertain, delight, please) or phatic (intended to keep in touch) purposes (Weigle, 2002: 9).

In relation to discourse modes, Figure 8a below illustrates the proportion of passages in the GCE 1162 reading examination from five categories of discourse mode, specifically, narrative, expository, argumentative, functional and descriptive.

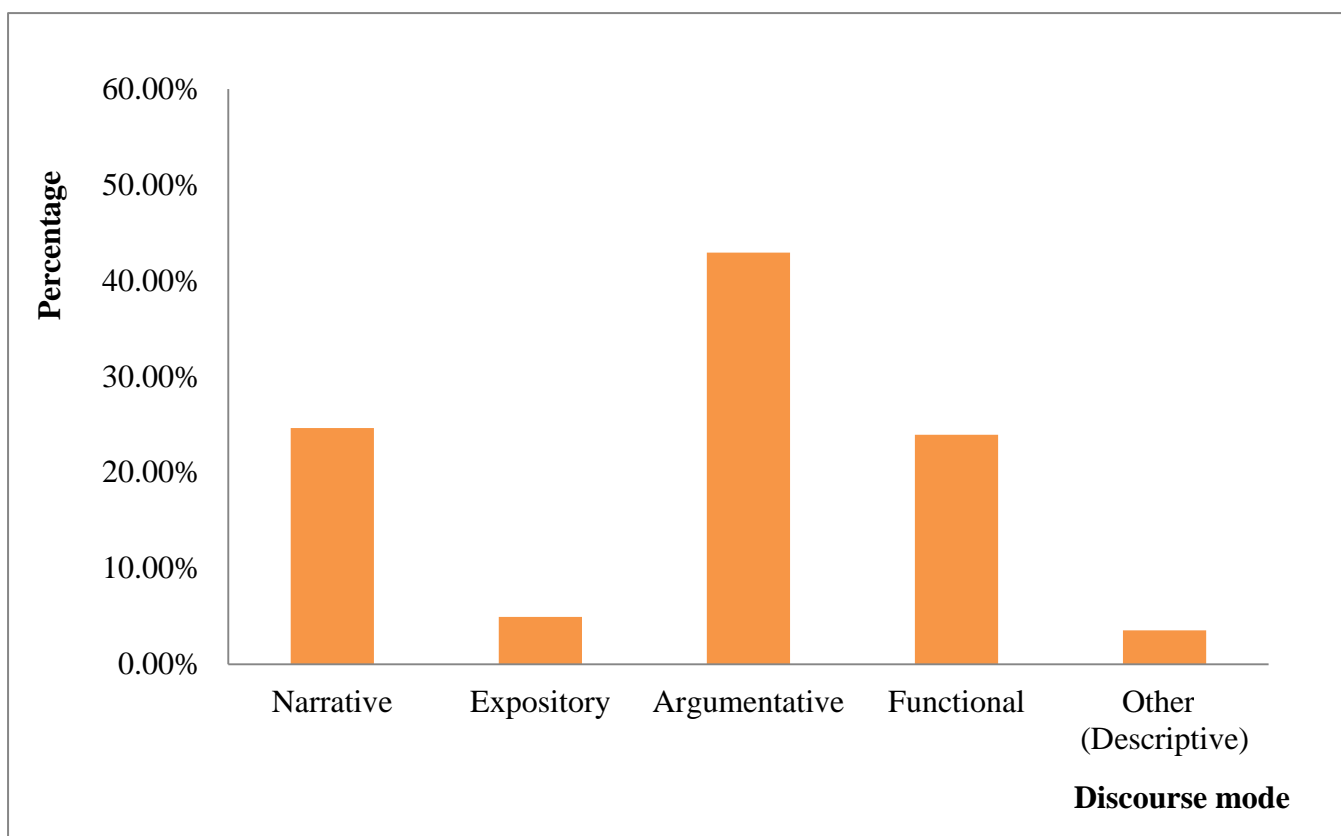


Figure 8a: Breakdown of passages by discourse mode in the GCE 1162 reading examination (May 2006-May 2016)

Of the 142 passages studied, 61 passages (42.96%) are argumentative texts. Also known as persuasive texts, argumentative texts form the largest group by discourse mode. Argumentative texts present the writer's viewpoint in the hope that the reader will accept the particular assertion being made. Well-written argumentative texts are structured and organized, consisting of the position, reasons, supporting evidence and counter arguments held together by sound reasoning. The main purpose of argumentative texts is conative and items usually assess the test-taker's understanding of the author's attitude and opinion, text organization and implications. Examples from the examination include passages persuading adolescents to seize the day and work hard (Passage 2A, May 2011) and convincing them to persevere and succeed in life (Passage 3B, November 2009).

Narrative texts tell a story with the intention of entertaining and engaging the reader or to convey the emotions and feelings of the author. Narrative texts can also be

written to inform, inspire or persuade. Narratives can be either fictional or based on facts. They are characterized by temporal organization: beginning (an orientation that sets the scene and introduces characters), middle (complications and climax), and end (resolution and coda). Events unfolding over time constitute the plot. Other generic features of narratives include characters, settings and themes. 35 passages (24.65%) from the examined GCE 1162 reading papers are denoted as narratives. Examples include a short story about an old lady waiting for her sons to return home for a Chinese New Year reunion dinner (Passage 2A, May 2009) and an anecdote about the American novelist Nathaniel Hawthorne (Passage 2C, November 2008). Identifying the 5W1H (who, what, where, when, why, how), summarizing the plot, comparing the characters and identifying the themes are all typical items designed for narrative texts in the examination.

Functional texts constitute an equally sizable group in the 22 sets of GCE 1162 reading examination papers (34 passages, 23.94%). Functional texts are texts written to provide support, directions and other useful information to help readers accomplish everyday tasks. They can range from instruction manuals and recipes to TV schedules, posters and directories. Examples from the examination include an advertisement for foldable handbag hooks (Passage 2C, May 2012) and a toy donation drive announcement (Passage 2A, Specimen Paper 2016; Passage 2C, May 2006). Items generally test lower-order thinking skills such as locating explicit information and identifying the overall gist. Following revisions to the CL2 curriculum and assessment in 2012, there has been a surge in the proportion of functional texts which adopt a more communicative approach. The percentage of functional texts has increased from 15.28% (old examination format) to 32.14% (new examination format) and 35.71% (latest examination format) (see Figure 8b). Using the Chi-square test of independence, a Chi-square value of 6.10 is obtained (see Figure 8c). The P-value is 0.047. The difference between the proportion of functional texts and examination formats is therefore statistically significant at the 5% significance level.

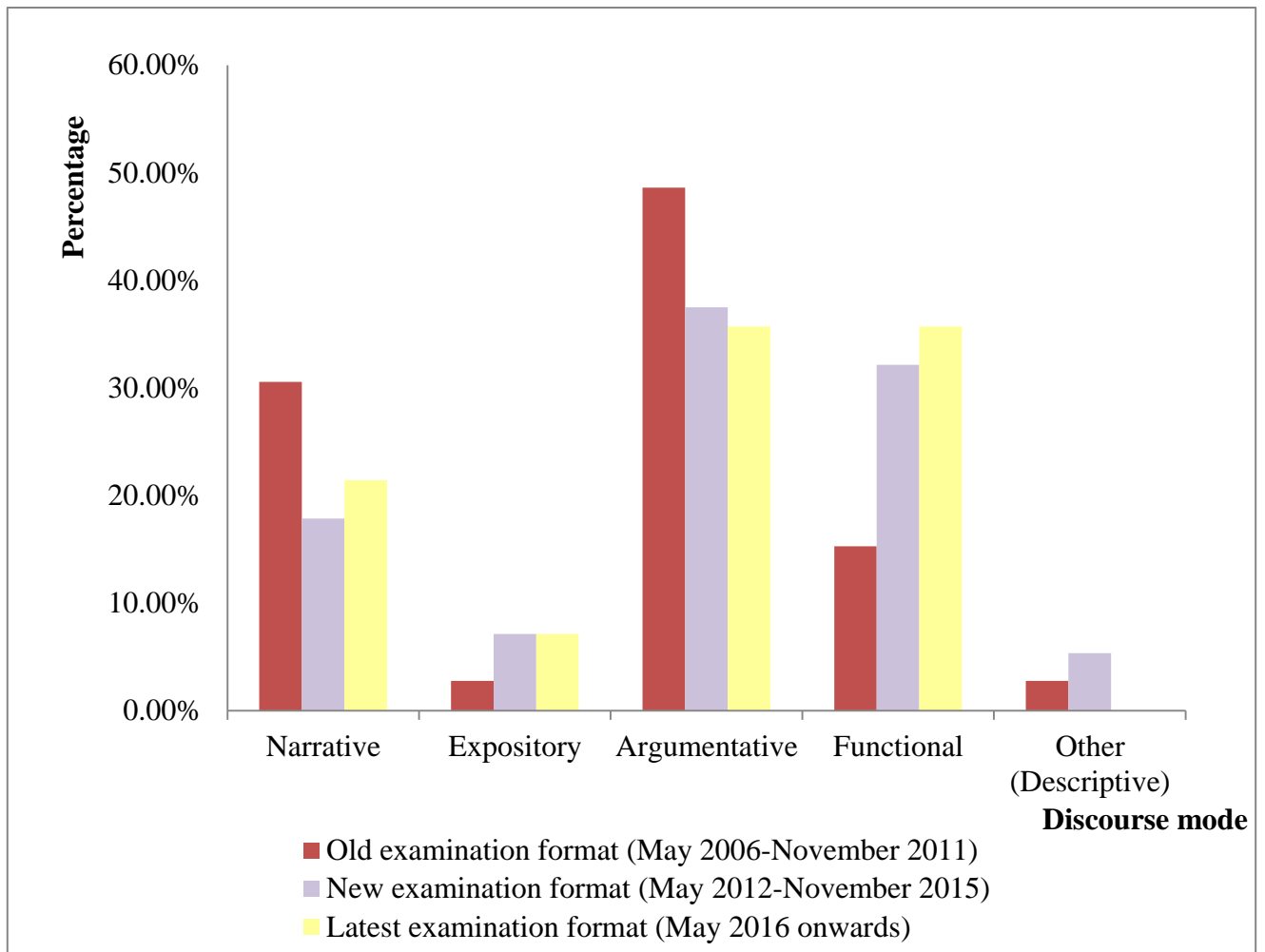


Figure 8b: Breakdown of passages by discourse mode across examination formats

	Old examination format (May 2006-November 2011)	New examination format (May 2012-November 2015)	Latest examination format (May 2016 onwards)	Row total
Functional texts	11	18	5	11 + 18 + 5 = 34
All other discourse modes	61	38	9	61 + 38 + 9 = 108
Column total	11 + 61 = 72	18 + 38 = 56	5 + 9 = 14	72 + 56 + 14 = 142 (Grand total)
Chi-square	6.10			
P-value	0.047			

Figure 8c: Frequency of passages of different discourse modes by examination format (May 2006-May 2016)

Expository texts form an under-represented group of 7 passages (4.93%). To distinguish between expository and functional texts, the narrower definition of expository texts has been adopted, that is, texts that focus on educating the reader. These texts serve a referential purpose, introducing facts, analysing the information and presenting an appropriate discussion in a clear and concise manner. Examples from the examination include passages on sustainable urban development in Europe (Passage 2D, May 2013) and the Mid-autumn Festival in Singapore (Passage 1, November 2008). There is also a rare popular science expository text on bees and their vital role in the food chain in a recent examination (Passage 2D, May 2016).

Although the descriptive text, for reasons unclear to the SMEs, is not listed as one of the discourse modes to be assessed in the *Secondary Chinese Language Syllabus 2011* (Syllabus 2011), there are a number of descriptive passages in the GCE 1162 reading examination (5 passages, 3.52%). Descriptive texts recreate, through careful observation, the specific and distinctive features of a person, place or event.

Examples from the examination include passages describing a local bazaar (Passage 2B, November 2013) and kampongs (Passage 1, November 2011).

In looking at discourse mode and text purpose in the GCE 1162 reading examination, it becomes apparent that argumentative texts aimed primarily at convincing readers to adopt values such as filial piety, humility and industriousness make up a significant proportion. This emphasis on mores is in step with the role of mother tongue languages as a vehicle for the dissemination of virtues and culture, as detailed in Chapter 3 on the Singaporean context. Test designers may also be influenced by the central tenet in classical Chinese writings that literature is subservient to a system of morality, that the larger purpose of literature is to teach morality (文以載道). SMEs opine that even within many of the narrative passages in the examination, there are chunks of persuasive writing. As one SME remarks:

Some of these stories become half narrative and half persuasive [...] The moral lessons are not embedded, rather they are stated explicitly and at great length [...] sometimes for several paragraphs. They are like a kind of *awkward hybrid text* (original emphases) [...] not only leaving little room for test-takers to contemplate and discover the underlying message themselves [...] but also limiting the higher-order thinking items item setters can formulate, since the message is made so conspicuous.

8.3.4 Propositional content

Interviewees and SMEs reveal three main principles that guide the selection of propositional content. The first principle is topic familiarity. Even though background knowledge is not tested per se, the relationship between a test-taker's existing schemata and the content of selected texts is one that is crucial. Texts at both extremes of the familiarity continuum should be excluded. Arcane texts that are inaccessible and texts with an inappropriate level of specificity will inevitably penalize test-takers. For example, while the inclusion of a popular science expository text on bees in the GCE 1162 reading examination (Passage 2D, May 2016) is appropriate, a passage on the biology and external morphology of bees for the same

audience would be obscure. Topic familiarity is needed to help test-takers allocate attention, direct interest and judge the importance of information, all of which are necessary for deriving meaning. As Alderson (2000: 29) asserts, 'every attempt should be made to allow background knowledge to facilitate performance, rather than allowing its absence to inhibit performance'. Conversely, the content should not be so familiar that most items designed can be answered without recourse to the text itself. Care should be exercised to avoid unnecessary repetition of similar or identical topics within a set of examination paper and also across time.

The second principle in the selection of texts relates to their degree of authenticity and appeal. Test-takers should not, as far as possible, be faced with texts that are constructed for the purpose of tests and examinations. Given that genuine and unmodified texts can afford a reading experience much closer to real life, it can be argued that the conclusions extrapolated from test results are more valid and reliable. Consideration should also be given to whether the texts are of interest and value to the broad range of test-takers, which in the case of the GCE 1162 reading examination, is a relatively heterogeneous group of adolescents. Undoubtedly, engaging passages are more likely to facilitate interaction between test-takers and text. In addition, given the curriculum goal of enhancing literary awareness and appreciation, selected passages should ideally be texts of literary merit. Studies conducted indicate that both topic familiarity and interest correlate with text comprehension under examination settings (e.g. Rahman & Mislevy, 2017; Bray & Barron, 2004; Artelt, Schiefele & Schneider, 2001).

Third, topics that are considered potentially distressing or biased are deemed unsuitable. These include texts about religion, politics, terminal illness, severe family or social problems and unethical behaviour. Texts that adopt offensive or condescending attitudes towards other nations, cultures and beliefs are not appropriate either. Test designers must also be aware if any topic favours test-takers of a particular background, age or gender.

Bearing in mind the three main principles that guide the selection of propositional content, I next describe how the GCE 1162 reading examination has addressed the parameter of propositional content. Based on the thematic concerns indicated in

Syllabus 2011, SMEs recognized six main topics into which the 142 passages could be grouped: values and attitudes, traditions and festivals, local news and culture, global awareness, aesthetic appreciation and advertisement and lifestyle (see Figure 8d).

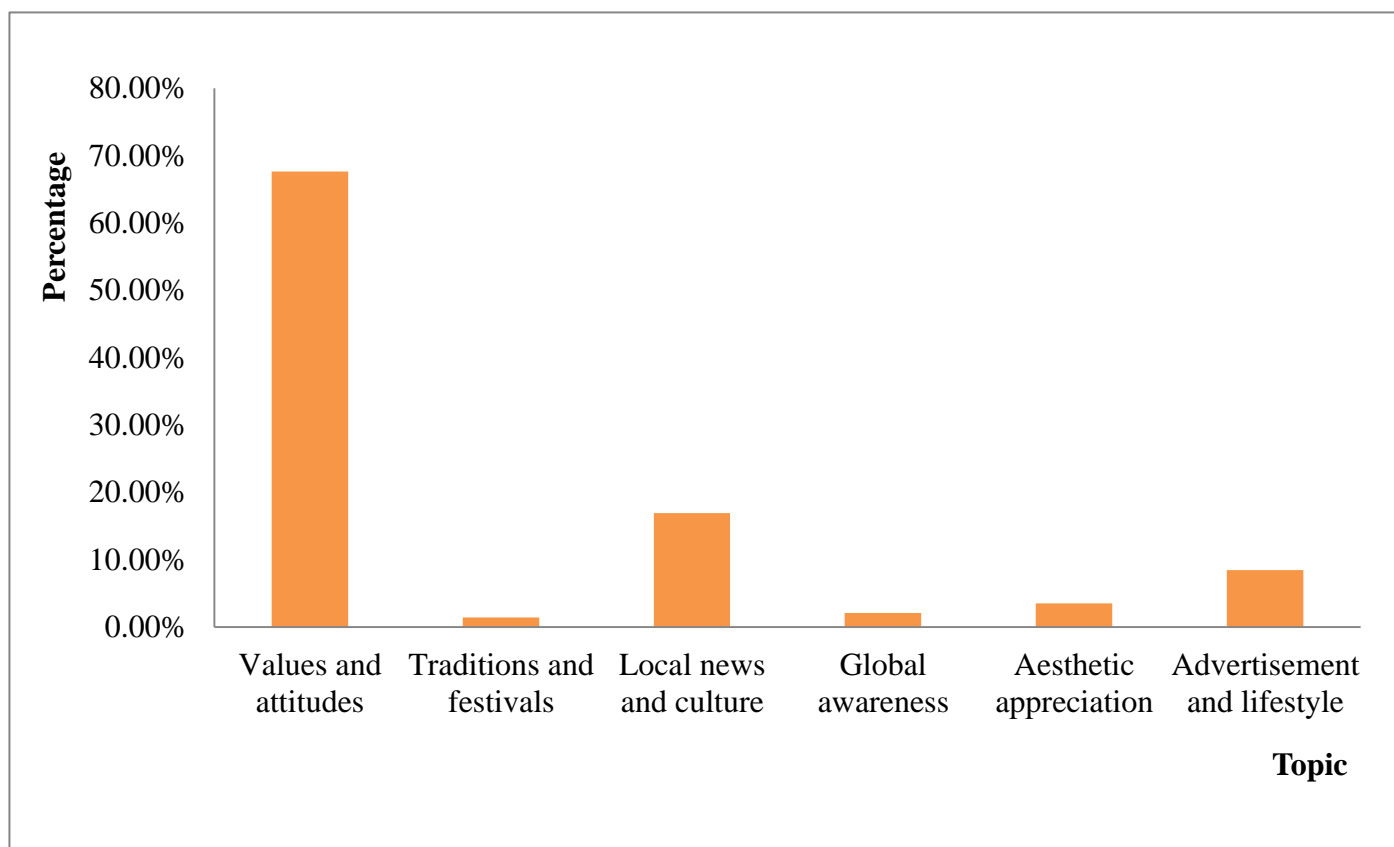


Figure 8d: Breakdown of passages by topic in the GCE 1162 reading examination (May 2006-May 2016)

Figure 8d above shows that a substantial percentage of passages (67.61%) centre on values and attitudes, such as determination (Passage 3B, November 2015), kindness (Passage 2C, May 2014), gratitude (Passage 3B, May 2013) and punctuality (Passage 1, May 2006). The GCE 1162 reading examination, being skewed towards passages that expound values and attitudes, presents a multitude of problems. To begin with, these passages generally score low on authenticity and appeal, as evinced by data from semi-structured interviews and expert judgement. As previously mentioned, interviewees express their doubts that these passages will interest adolescents. Omicron comments that ‘it is unlikely that an adolescent will voluntarily pick up a

text on *Liyilianchi* (礼义廉耻)⁶ to read’, ‘even as an adult, most of these passages [...] which seem to be contrived for testing purposes [...] do not pique my interest’. An SME adds that ‘too many of these “preachy” passages make the examination appear unnecessarily “dated”’, and that they ‘may have a negative washback effect on the curriculum, teacher and learner behaviours’ and ‘could be better aligned with the new curriculum’s communicative approach’.

The issue, for SMEs and interviewees, is not so much the transmission of values and attitudes as how they are being transmitted. Mu indicates that ‘moral lessons could be delivered more subtly [...] [and woven] more seamlessly into the passage’, giving test-takers the opportunity to analyse and validate for themselves. This naturally encourages the activation of higher order cognitive processes. Furthermore, repetition leads to predictability. Some teachers, as interviewees reveal, provide model answers for open-ended evaluate and apply questions on commonly included values and attitudes such as perseverance, fortitude and love for the Chinese language, and make students memorize and regurgitate these prepared answers in the examination. This rote learning could erode the validity of the test scores obtained.

Passages on Singapore’s local news and culture represent the second largest group (16.90%). Examples include extracts from newspaper articles on the local Plant-a-Tree Event (Passage 2B, November 2015), the Singapore Kite Day at Marina Barrage (Passage 2B, May 2014) and a phone application developed by Singapore polytechnic students for visually impaired commuters (Passage 2A, November 2013). The initiative to include local elements in the examination is laudable, as interviewees comment, and may spur students to read more to gain awareness of what is happening around them. Next, advertisements and lifestyle articles account for 8.45% of the passages. Advertisements on air purifiers (Passage 2C, Specimen Paper 2016), spectacles (Passage 2C, November 2015), ornaments (Passage 2C, May 2013) and electronic book readers (Passage 2C, November 2013) can be found in recent examination papers. Lifestyle articles include sports features, for example rock-climbing (Passage 3B, Specimen Paper 2016; Passage 3A, November 2007),

⁶ *Liyilianchi* (礼义廉耻) are four basic social bonds, namely the senses of propriety, justice, integrity and honour.

and updates on technological advancements, such as how touch-screen devices affect the way we use our fingers (Passage 2B, November 2008).

Texts on the remaining three topics constitute only a small fraction of the 142 passages surveyed. A handful of passages (3.52%) are about aesthetic appreciation, for instance, a love for music (Passage 2B, Specimen Paper 2016; Passage 2B, November 2007) and unique stamp designs (Passage 2B, May 2015). There are also a few passages (2.11%) that inspire global awareness, commenting on, for example, the book-crossing project in America (Passage 2C, May 2016) and sustainable urban development in Europe (Passage 2D, May 2013). Additionally, there are a couple of passages (1.41%) on traditions and festivals—one discusses the need to preserve the essence of traditional festivals while adapting their forms to changing times (Passage 2B, May 2013) and another raises the topic of the Mid-autumn festival (Passage 1, November 2008). There are also passages that touch on traditions and cultures but spotlight the values embedded in them; these passages would be more aptly placed under the values and attitudes category. The noticeable absence of passages on contemporary Chinese culture deserves attention from all involved in developing the GCE 1162 reading examination. In a bid to stay relevant to adolescents, comparable Chinese language examinations, such as the Cambridge International General Certificate of Secondary Education and the International Baccalaureate suite of Chinese as a second language papers have given more prominence to passages on contemporary Chinese culture and contemporary culture at large. Following suit, texts on popular contemporary culture including film, art, literature and even the practices, mindsets and concerns of different Chinese communities could also be integrated into the GCE 1162 reading examination.

Finally in this subsection, the literary merit of the passages selected for the GCE 1162 reading examination needs to be considered. Literary merit refers to the quality of writing. While literary merit can be a highly contentious topic, there seem to be some common yardsticks against which the quality of a piece of text can be measured (The College Board, 2010). Works by authors such as *Lu Xun* (鲁迅), *Shen Congwen* (沈从文) and *Eileen Chang* (张爱玲) have stood the test of time. The almost complete unanimity of opinion that their works are part of the Chinese literary

canon must have had some common basis of formation. The SMEs, all with a first degree in Chinese language and literature, drew up a list of criteria for measuring the literary merit of the 142 passages assessed (see Figure 8e).

Criteria	Question
Plot (narrative texts)	Does the plot demonstrate the writer's originality and creativity? Is it deft, interesting and true to life?
Argument/objective (argumentative and expository texts)	Does the text display depth of thought? Is it convincing and well-researched?
Characters (narrative texts)	Are the characters robust? Do they exhibit emotional complexity? Are their dialogues sharp and realistic?
Evidence/facts (argumentative and expository texts)	Are the texts in alignment with the argument/objective? Are they persuasive and credible?
Form and style	Are the writer's words crafted and vivid? Is the text well-structured? Does it achieve its overall text purpose?
Appeal	Does the text have the potential to resonate with test-takers?

Figure 8e: Criteria for measuring the literary merit of the GCE 1162 passages

The list of criteria is then used by the SMEs to score each passage holistically, with 0 being not applicable or of low literary calibre, 1 being moderate calibre and 2 being high calibre. SMEs arrive at their decisions independently. A mean score of 0.23 was obtained for the 142 passages, indicating a general lack of literary merit. Attaching a quantitative value to literary merit, as SMEs discern, can be in some way arbitrary. Nevertheless, collectively these scores further our understanding of the examination. An SME asserts:

Most of the passages lack lustre [...] they do not leave much of an impression. I didn't come across any extracts from canonical texts, I think. There's no indication of text sources, unlike in the O level English examination [...] There are also few passages that adolescents will enjoy reading [...] The issue is not with the word limit, flash fiction stories can be masterful works of literature despite their brevity, even advertising slogans, like the iconic *Ni pai he ma*(你怕黑吗).⁷

8.3.5 Readability

Readability is the ease at which a reader can understand a piece of writing. It is what makes some texts more complex and challenging to read than others. *The Literacy Dictionary* defines readability as 'the ease of comprehension because of style of writing' (Harris & Hodges, 1995: 203). Hargis et al. (2004: 6) state that readability is 'the ease of reading words and sentences'. In a broader sense, readability can be understood as 'the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting' (Dale & Chall, 1949: 23). The readability of a text is one of the prime determinants of whether a text is suitable for use in an examination. In order to select texts that are geared to the test-takers' reading proficiency level, test designers generally rely on two approaches: an intuitive or a reading formula approach.

An intuitive approach relies on the test designers' natural instinct regarding text comprehensibility, or the 'gut feel' that interviewees identify. Some characteristics of texts with high readability that interviewees list include a large proportion of high frequency words, simple sentence structure with few subordinate clauses, shorter text and sentence length and greater text cohesion and coherence. Determining the difficulty of a text using the intuitive approach naturally has its limitations—it is

⁷ *Ni pai he ma* (你怕黑吗) or *Darkness* is an influential Guinness Stout advertisement fronted by Hong Kong celebrity George Lam. Created by a Singaporean advertising designer *Sau-Hoong Lim* in 1992, the tagline 你，怕黑吗？黑有什么好的？怕黑，那你不是白白地活着吗？(Are you afraid of the dark? If so, wouldn't all life be futility?) is a clever play on the Chinese word 白 (*bai*) which can mean *white* or *futile*.

highly subjective and time consuming and even then, it is often difficult to reach a consensus. In a recent study conducted by the National Taiwan Normal University, SMEs initially failed to agree on 83.46% of the texts they were asked to assign levels to (Sung, Lin, Dyson, Chang & Chen, 2015).⁸

To mitigate the problems associated with an intuitive approach, a readability formula approach can be adopted. Classical readability formulae originated from research in English as a first language. In 1921, Thorndike published *The Teacher's Word Book* which contains an extensive list of words in English by frequency (Thorndike, 1921). This book became a basis for the work of Lively and Pressey who developed the first readability formula in 1923. By the 1980s, there were more than 200 readability formulae and over a thousand relevant studies (DuBay, 2004). Of the numerous readability formulae available, the Flesch-Kincaid Grade Level (Flesch, 1951, 1943), Dale-Chall (Chall & Dall, 1995; Dale & Chall, 1948) and Fry Graph (Fry, 1977) are probably the best known. Classical formulae provide an approximation of readability through surface textual features such as sentence length, word length and word difficulty. The revolution in cognitive psychology in 1975 brought about a deeper understanding of reading and gave rise to new readability formulae that attempted to account for more complex textual as well as cognitive features. Concerns about the suitability of readability formulae created from first language texts for establishing the readability of second language assessment and educational materials have arisen. Subsequently, new second language readability formulae emerged such as the Coh-Metrix Reading Index developed by Crossley, Greenfield and McNamara (2008), described as 'a computational tool that measures cohesion and text difficulty at various levels of language, discourse and conceptual analysis' (Crossley, Allen & McNamara, 2011: 88).

While European and American academics have developed increasingly sophisticated algorithms for measuring English language readability, research in Chinese language readability is lagging behind substantially (Sung et. al, 2015; Wang, 2008). Chinese, being a logographic language, is significantly different from alphabetic languages,

⁸ SMEs were required to grade selected CLF texts using the Common European Framework of Reference (CEFR) global scale, namely texts suitable for level A1 (Breakthrough), A2 (Waystage), B1 (Threshold), B2 (Vantage), C1 (Effective operational proficiency) or C2 (Mastery).

preventing the direct application of any English readability formulae. An early study (Yang, 1971) which factors in the complexity of characters in addition to the difficulty of vocabulary and number of sentences reveals how Chinese words are made up of characters which are themselves composed of strokes (笔画). Characters with a greater number of strokes are more difficult to remember and identify and are thus likely to increase text complexity. Subsequent Chinese readability formulae (Shen, 2005; Zhang, 2000; Jing, 1995; Sun, 1992) involve similar sets of variables. In short, most Chinese readability formulae tend to include only a few shallow textual features and are therefore unlikely to represent accurately the full readability of a text. Furthermore, none of these formulae has gained widespread usage (Wang, 2008).

In light of the drawbacks of the aforementioned Chinese readability formulae, the Chinese Readability Index Explorer for Chinese as a Foreign Language (CRIE-CFL) has been chosen in this study to examine the readability of passages used in the GCE 1162 reading examination. The CRIE-CFL, currently undergoing refinement, was devised by a group of academics at the National Taiwan Normal University. Building on the Coh-Metrix Reading Index, the CRIE-CFL aims to eventually take into consideration an extensive range of indicators when calculating the Chinese readability index (Sung et. al, 2015). Although the CRIE-CFL is not developed with Singapore's context in mind, it has the capacity to provide a baseline for further exploration into the readability evaluation of CL2 assessment materials in Singapore. At present, the CRIE-CFL has a readability mathematical model that sorts 30 indicators by importance and sequentially integrates them into the CRIE-CFL to gauge the difficulty of a text. In relation to the GCE 1162 reading examination, I now proceed to examine ten of the pertinent CRIE-CFL indicators of readability.

The first indicator of readability offers a basic numerical break down of texts. The total number of paragraphs and sentences and the average number of sentences in each paragraph are listed in Figure 8f.

	Number of passages	Number of paragraphs	Number of sentences	Average number of sentences in each paragraph (Standard deviation SD)
Old examination format (May 2006-November 2011)	72	335	954	2.85 (0.27)
New examination format (May 2012-November 2015)	56	257	713	2.77 (0.27)
Latest examination format (May 2016 onwards)	14	74	199	2.69 (0.28)
May 2006-May 2016 examination papers	72 + 56 + 14 = 142	335 + 257 + 74 = 666	954 + 713 + 199 = 1,866	1,866 ÷ 666 = 2.80 (0.26)

Figure 8f: Basic numerical break down of passages in the GCE 1162 reading examination (May 2006-May 2016)

The second indicator of readability is the lexical level. Figure 8g displays information on characters and words in the examination. Longer texts are deemed more demanding for test-takers. A reading rate of 34.47 characters per minute is obtained for the GCE 1162 reading examination.⁹ In comparison, the *Hanyu Shuiping Kaoshi* (HSK 汉语水平考试) Level 5 reading examination demands an approximate reading rate of 128 characters per minute (The Office of Chinese Language Council International, 2016). There appears to be a sizable disparity in the

⁹ To obtain the reading rate for the GCE 1162 reading examination, the number of characters for all items, including stems and options (18,592 characters) is added to the total text length (49,663 characters), averaged across the 22 sets of papers from May 2006 to May 2016 $(18,592 + 49,663) \div 22 = 3,102.5$ and divided by the allocated test time of 90 minutes $(3,102.5 \div 90 = 34.47$ characters per minute).

reading rate required for the two examinations, however, as the GCE 1162 reading examination partly comprises constructed-response items, unlike the latter which consists entirely of MCQ items, time for writing has to be taken into account.

	Number of characters	Number of words	Average number of characters in each passage (Standard deviation SD)	Average number of words in each passage (Standard deviation SD)
Old examination format (May 2006-November 2011)	25,263	15,283	$25,263 \div (6 \times 12) = 350.88$ (16.94)	$15,283 \div (6 \times 12) = 212.26$ (12.51)
New examination format (May 2012-November 2015)	19,490	11,769	$19,490 \div (7 \times 8) = 348.04$ (21.48)	$11,769 \div (7 \times 8) = 210.16$ (14.18)
Latest examination format (May 2016 onwards)	4,910	3,021	$4,910 \div (7 \times 2) = 350.71$ (22.43)	$3,021 \div (7 \times 2) = 215.79$ (25.96)
May 2006-May 2016 examination papers	49,663	30,073	$49,663 \div 142 = 349.74$ (18.16)	$30,073 \div 142 = 211.78$ (13.55)

Figure 8g: Numerical breakdown of character and word counts in the GCE 1162 reading examination (May 2006-May 2016)

The third indicator is character complexity (see Figure 8h). 85.03% of the 49,663 characters have a low stroke count (10 strokes and below) and 14.92% are characters with a medium stroke count (11 to 20 strokes). Only 0.04% have a high stroke count (21 strokes and above). Characters with a high stroke count increase the time needed for text comprehension, not unlike long words in alphabetic languages. The GCE 1162 reading examination has an average character stroke count of 7.32, which falls within the range of a low stroke count.

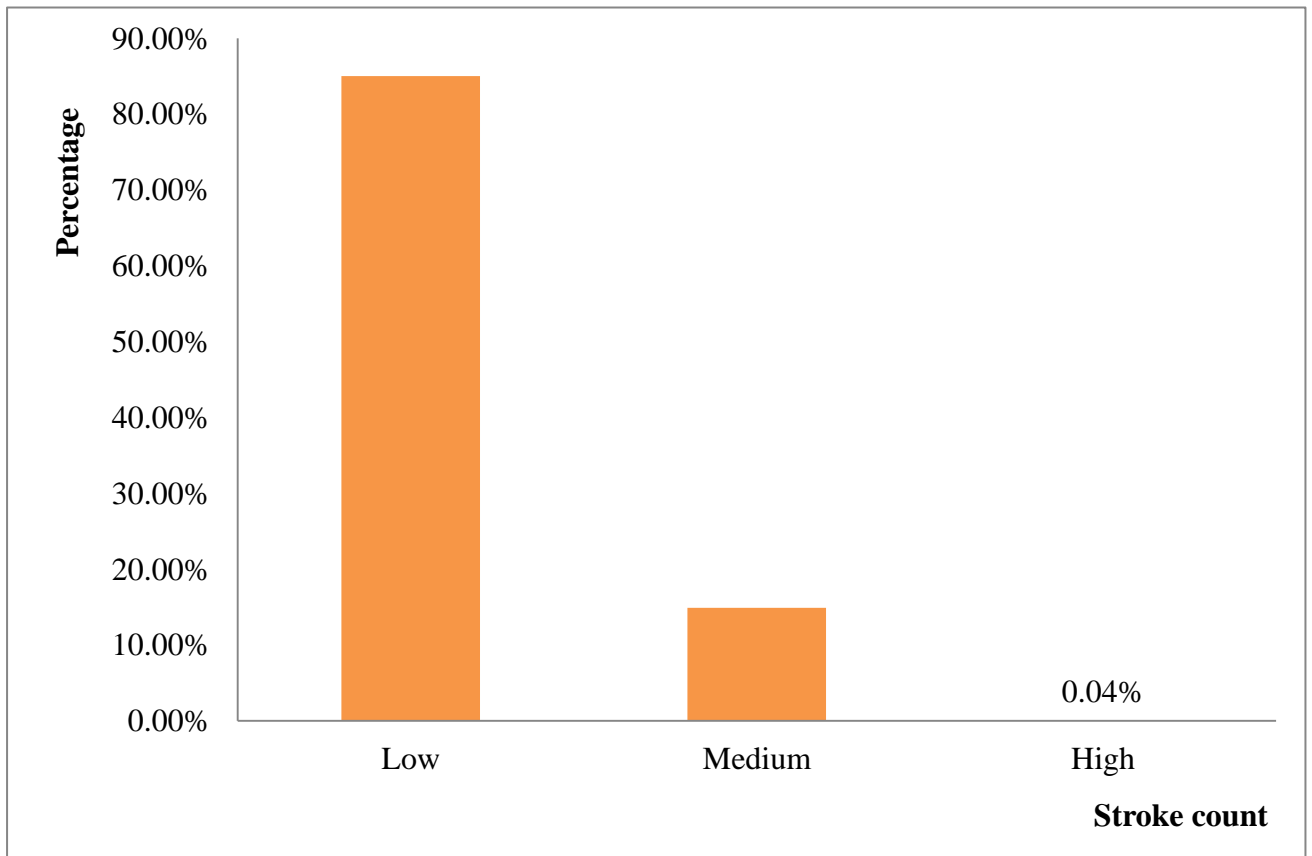


Figure 8h: Numerical breakdown of character stroke counts in the GCE 1162 reading examination (May 2006-May 2016)

The type-token ratio (TTR) is a fourth indicator commonly used to establish the readability of a text. When calculating the TTR, the count of all unique words in a text (types) is divided by the number of total words (tokens). Hence, the higher the TTR, the richer the lexical variety. A low TTR of below 0.50 denotes frequent lexical repetition. The TTR across the 142 passages sampled is relatively stable, with an average ratio of 0.69, indicating moderate lexical variation.

The difficulty of individual words in a passage is a fifth indicator of text readability. It has long been recognized that vocabulary knowledge plays a key role in reading comprehension (Nation 2001; Laufer, 1992; Anderson & Freebody, 1983; Tuinman & Brady, 1974). Nation (2001) proposes that 80% coverage of a text, meaning that four words in every five need to be comprehended by the reader, is the minimum for basic understanding. Building on Nation's findings, subsequent research on English reading among second language learners shows that a vocabulary size of 3,750-4,500

and 4,500-5,000 is demanded at levels C1 and C2 respectively on the Common European Framework of Reference for Languages (CEFR) (Meara & Milton, 2003). In CL2 reading, academics such as Chang (2012), Cheng (1998), and Liu and Song (1992), suggest that 3,000 high frequency words account for 86% of non-academic texts and with a vocabulary of 8,000 words, a reader can understand 95% of most texts.¹⁰ Chang (2012) then goes a step further by ranking the 8,000 words by frequency and classifying them according to the CEFR levels. It is estimated that at C2 level, CL2 learners should have a grasp of 8,000 words although the Office of Chinese Language Council International sets a lower benchmark of 5,000 words (Chang, 2012). These numbers are useful guidelines in that they represent what could be expected of test-takers of the GCE 1162 reading examination.

Based on the premise that there is a strong inverse correlation between word difficulty and frequency, the CRIE-CFL algorithm matches all words in a text to Chang's 8,000 Chinese words list and assigns a difficulty level to each word (Sung et al., 2015). As the presence of a difficult word can cause a sentence to be considerably more difficult, the algorithm calculates the mean square of the vocabulary level to increase the influence of difficult words. The number of difficult words, as defined by words in the CEFR B2 to C2 range, is also provided. The 142 passages from the GCE 1162 reading examination have in sum 10,371 difficult words, accounting for 34.49% of the total number of words. As a caveat, word frequency lists become less indicative at higher levels of language proficiency (Khalifa & Weir, 2009; Nation 2001), mainly because at higher levels, discourse can range across a wide variety of topics, making it impossible to specify a particular set of vocabulary.

The sixth indicator of readability is the semantic-propositional level of the passage. As all words can be broadly classified into content or function words, the first semantic-propositional feature is the proportion of content words to function words. Content words are words with lexical meaning. They include nouns, verbs, adjectives

¹⁰ The *Secondary Chinese Language Syllabus 2002* states that students are expected to recognize the 3,000 Chinese characters in the MOE stipulated character list at the end of their secondary education. The number of words, however, is not stated in either the 2002 or 2011 syllabuses.

and most adverbs, all of which belong to open-class words. Function words are close-class words that express grammatical relationships, such as conjunctions, auxiliary verbs, particles and prepositions. Function words are a finite set of words—they are very resistant to the introduction of new items; whereas new content words may be added readily, such as technical terms, cyber-slang and adoptions of foreign words. Content words form 79.62% of all the words in the 142 passages examined. In general, more content words imply more concepts, thus requiring readers to expend more time and effort to understand the text.

Semantic fields of a word is a seventh indicator of readability (Cheng, 2005). A semantic field is a lexical set of words with closely interrelated meanings. For example, under the concept of furniture, words like table, wardrobe, sofa and bed (Passage 2A, May 2014) could form a semantic field. In this semantic field, furniture is the genus or general concept and table, wardrobe, sofa and bed are species or specific concepts. When the semantic field represents a concrete concept, such as furniture, it is generally easier to comprehend than if it denotes an abstract concept, such as regret (Passage 2A, May 2011). Concrete concepts are easier to process as test-takers can tap into both verbal and imagery systems, while they could only draw on the verbal system when decoding abstract concepts. In addition, the more semantic fields a single word can be categorized under, the more onerous it is to make meaning of the sentence that contains it. When sentences in the GCE 1162 reading examinations are compared with those in the CRIE-CFL corpus of more than 1,500 CL2 texts, 62.09% of the GCE 1162 sentences contain words that are semantically more complex than the average of the corpus.

The eighth indicator of readability is a semantic-propositional feature based on the number of idioms and sub-technical words. Idioms are fixed expressions that are typically used in a figurative sense. As many Chinese idioms are steeped in culture and history, their meanings are often undeducible from their constituent words. For instance, 独树一帜 (*dushuyizhi*; Passage 1, November 2009): in the literal meaning, this can be construed as erecting a separate flag when the actual meaning of this idiom is to be distinctive and unique. Sub-technical vocabulary covers a wide range of academic words which are relatively independent of any specific discipline but are

less general than everyday words. Examples include 息息相关 (*xixixiangguan*, correlated; Passage 1, May 2008), 档案 (*dangan*, archives; Passage 1, November 2011) and 趋势 (*qushi*, trend, Passage 2A, November 2014). A text featuring little or no sub-technical words is presumed to be easier to read. 671 idioms and sub-technical words were picked out from the total text input, accounting for 2.23% of the 30,073 words.

The ninth indication of readability is the syntactic level. From the beginning of readability studies, it has been recognized that the difficulty of a sentence involves elements beyond the difficulties of the words in it. For many years, researchers have used average sentence length as a proxy for both syntactic and lexical load. Long sentences tend to have more modifiers and qualifiers, more embedded phrases and clauses and complex rather than simple structures, requiring students to demonstrate higher levels of reading proficiency. Although as Alderson (2000) cautions, there is substantial research which shows that adding words, instead of deleting words, may sometimes lend clarity to a sentence. Average sentence length has been included in many of the more widely used readability formulae and is relatively quick and inexpensive to compute. The sentence length across the GCE 1162 passages sampled is relatively constant, with an average of 16.12 words per sentence (see Figure 8i).

	Average sentence length (Standard deviation SD)
Old examination format (May 2006-November 2011)	16.02 (0.21)
New examination format (May 2012-November 2015)	16.51 (0.68)
Latest examination format (May 2016 onwards)	15.18 (1.39)
May 2006-May 2016 examination papers	16.12 (1.15)

Figure 8i: Breakdown of average sentence length in the GCE 1162 reading examination (May 2006-May 2016)

Besides calculating the average sentence length, the CRIE-CFL algorithm also identifies simple sentences by checking whether a sentence contains commas and coordinating conjunctions such as 和 (*he*, and), 但是 (*danshi*, but), 所以 (*suoyi*, so). A simple sentence is made up of only one independent clause, for instance, 我们经常有这样的经历。(We often experience this.; Passage 3B, November 2015) and 他终于站在眼前了。(He finally stood in front of us.; Passage 2A, May 2009), and is more comprehensible than compound and complex sentences. Simple sentences account for 37.30% of the total number of sentences. Though less refined as compared to, for example, the Revised Rosenberg and Abbeduto D-Level Scale (Covington, He, Brown, Naci & Brown, 2006) which rates a sentence from 0 to 7 based on its complexity, the CRIE-CFL algorithm nevertheless provides a useful estimate.

The tenth and final indicator of readability is the cohesion level. Cohesion relates to the interconnectedness of the various components of the surface text (Halliday & Hasan, 1976). Cohesion depends on presupposition and occurs when the understanding of one component is dependent on that of another. Conjunctions, substitution, ellipsis and reiteration are all cohesion markers although the CRIE-CFL algorithm can only recognize conjunctions at present. The numbers of conjunctions, including positive, negative and casual conjunctions are adopted as features.

Taking into consideration the ten indicators of readability above, on the CRIE-CFL gauge the GCE 1162 reading examination has a readability level equivalent to that of B2 (Vantage) to C1 (Effective operational proficiency) Levels, which in turn are comparable to Levels 4 to 5 of the HSK. The readability of passages selected for the GCE 1162 reading examination means that they are suitable for upper-intermediate to lower-advanced CL2 test-takers who can ‘understand the main ideas of complex text on both concrete and abstract topics’ and are beginning to ‘understand a wide range of demanding, longer texts, and recognize implicit meaning’ (Verhelst, Van Avermaet, Takala, Figueras & North, 2009: 24). The readability level appears to be appropriate for the GCE 1162 reading examination, the aim of which is to measure a ‘student’s ability to read narrative, expository, argumentative and functional texts of appropriate difficulty and appreciate literary texts [...] in congruence with the

secondary Chinese language syllabus' (SEAB, 2014a: 2). As investigated in Chapter 7, the GCE 1162 reading examination slants toward the eliciting of lower-order thinking skills of remember and understand instead of the higher-order thinking skills of analyse, evaluate and create.

Care must be taken, however, when comparing the GCE 1162 reading examination with other similar examinations such as the HSK or external standards such as the CEFR. An outline of external criteria is provided in the closing chapter, Chapter 9. It is also useful to remember that although the difficulty of a text is largely reliant on its lexical, semantic-propositional, syntactic and organizational attributes, it is equally dependant on text coherence too. Coherence, the configuration of content and ideas, relates to whether these elements of a text are accessible and relevant. Gauging coherence involves measuring the quality of the mental model constructed by a reader (McNamara, Ozuru, Graesser & Louwerse, 2006). The effectiveness of all quantitative readability formulae, including the CRIE-CFL, is predicated on the assumption that there is a relationship between the measurable properties of a text and its coherence. It also means that readability is subject to variation across a number of factors, such as background knowledge and experience, motivation and reading purpose. A formulaic approach that takes all of these conditions into consideration is not available and is unlikely to become so (Urquhart & Weir, 1998). Readability indices will therefore need to be complemented with qualitative feedback from SMEs and future test-takers.

8.4 Conclusion

In this chapter I have constructed an IA and VA for the contextual parameters inference, drawing on evidence which centres on five contextual parameters, specifically item type, mark scheme, discourse mode and text purpose, propositional content, and readability. In the first subsection of this chapter, the strengths and weaknesses of the item types used in the GCE 1162 reading examination, namely the MCQ, gap-filling test, SAQ and open-ended question were explored. Data affirm that suitable item types are employed in the GCE 1162 reading examination although additional item types such as information transfer items could potentially be included. The next subsection identified inadequacies in the GCE 1162 reading examination's

mark schemes. Suggestions were made on how to modify the existing mark schemes, especially for open-ended questions, so as to predict and reflect the Outcome Space more accurately. Evidence presented in the third subsection suggests that there is a fair representation of passages of different discourse modes and text purposes in the GCE 1162 reading examination, albeit concerns that argumentative texts may be over-represented at the expense of expository texts. The fourth subsection demonstrated that the range of propositional content for the GCE 1162 reading examination needs to be broadened as the examination features a disproportionately high percentage of passages manifesting values and attitudes. Passages selected for use in the examination also appear to score low in the areas of authenticity, appeal and literary merit. In the fifth subsection on readability, the CRIE-CFL was employed to assess the readability of passages in the GCE 1162 reading examination. A readability level equivalent to that of CEFR B2 (Vantage) to C1 (Effective operational proficiency) Levels was established, which appears to be appropriate for the GCE 1162 reading examination. Based on the data gathered through semi-structured interview and document analysis, the evaluation status of each assumption based on Shaw and Crisp's (2012) indicators are given below (Figure 8j):

	Assumption	Provisional evaluation status based on semi-structured interview and document analysis data
1.	A variety of suitable item types is employed to assess reading constructs.	Accepted with concerns
2.	Mark schemes are well defined for differentiating the quality of answers.	Plausible rejection
3.	There is adequate representation of passages of different discourse modes and text purposes.	Accepted with concerns
4.	There is adequate representation of passages with different propositional content.	Plausible rejection
5.	Passages in general possess literary merit.	Plausible rejection
6.	Passages in general are of a suitable readability level.	Accepted with concerns
7.	The contextual parameters of the examination support the assessment of constructs that are relevant to real-life reading contexts beyond the syllabus.	Plausible rejection

Figure 8j: Provisional evaluation status of the assumptions underpinning the contextual parameters inference relating to Singapore’s GCE 1162 reading examination

The discussion on the a priori validation of the GCE 1162 reading examination is now complete. The IA and VA of each of the four a priori inferences have been constructed, substantiated by supporting evidence and rebuttals collected through semi-structured interview, document analysis and expert judgement. Specifically, I have evaluated the specifications and administration of the examination, focusing on the purpose, construct and administrative structure. I have also portrayed the defining characteristics of test-takers sitting the examination and considered their possible implications for testing. The theory and research undergirding the cognitive and contextual parameters were reviewed and the GCE 1162 reading examination papers subjected to scrutiny. Although for descriptive purposes the various inferences are

presented as being distinct from one another, they are undoubtedly interconnected. Taken together, they define to a large extent the quality of the GCE 1162 reading examination.

It is useful to reiterate at this junction, that the four a priori inferences, derived mainly from Weir's (2005) socio-cognitive validity framework, point towards the before-the-test event parameters. Once the examination papers are marked and scores are available, the validation study enters the a posteriori or after-the-test event phase. Although the a posteriori validation is beyond the scope of this study, the closing chapter, Chapter 9 briefly outlines a posteriori parameters related to scoring, criterion, and washback effects and impact. In accordance with Kane's (2009, 2006) ABV, the four inferences with their underpinning research questions, claims, assumptions, supporting evidence and rebuttals will be revisited in Chapter 9 and consolidated to form an overall evaluation of the GCE 1162 reading examination. The contribution of this study to increased understanding of the GCE 1162 reading examination and the concepts of validity, validation and reading will also be discussed.

Chapter 9 Concluding remarks

9.1 Introduction

This study has investigated the degree to which the objectives of the reading component of the Singapore-Cambridge General Certificate of Education Ordinary-Level Chinese Language Examination (GCE 1162) are accurately measured. Comprehensive coverage of the fields of validity and validation, reading and the Singaporean context were first presented in Chapters 1 to 3. Next, an explication of the conceptual and methodological underpinnings of the study was provided in Chapter 4, specifically, its adoption of an argument-based approach to validation (ABV) (Kane, 2009, 2006), its pragmatic base guided by a mixed methods design and supported through the research methods of semi-structured interview, document analysis and expert judgement. Framed by Weir's (2005) socio-cognitive validity framework, the ABV, spanning Chapters 5 to 8, was organized around four a priori inferences, namely, specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters. In this closing chapter, the interpretive argument (IA) and validity argument (VA) for each inference is summarized and an overall evaluation of the GCE 1162 reading examination is provided to answer the main research question. The evaluation is followed by a short exploration into the a posteriori parameters and directions for future research. The chapter closes with highlights of the key implications and impact of the study.

9.2 Overall validity evaluation of the GCE 1162 reading examination

Validity is a central concern in any assessment—the objective of this study is essentially to amass and evaluate validity evidence and potential rebuttals for the GCE 1162 reading examination through an ABV. Yet, examination boards rarely set out to maximize validity, rather, they seek only to optimize it, making validity ‘as high as possible, whilst accommodating a broad profile of intended purposes and recognising a wide range of operational constraints’ (Newton, 2017b: 20). Assessment policy, design and practice are plagued with contradictions and dilemmas, and compromise is, therefore, inevitable. There is often a trade-off, as

demonstrated earlier, between validity and other considerations such as feasibility, authenticity and wider educational and political goals. Conflict also arises among stakeholders. For instance, when decisions about the purposes and constructs of an examination are made, there will be competing interests, intentions and interpretations; when the blueprint of an examination is transmitted from policy makers to test designers for enactment, it is almost always subject to multiple and sometimes selective decoding. Similarly, at the receiving end, test-takers and users of examination results may have agendas and expectations that run contrary to policy makers' intentions. The complex and dynamic nature of an assessment procedure, especially that of a high-stakes national examination like the GCE 1162, predetermines that there is no such thing as perfect validity.

The question that needs to be asked as pieces of validity evidence are drawn together in this section is, therefore, how much validity is sufficient validity? Validation processes tend to be very open-ended and 'lengthy, even endless' (Cronbach, 1989: 151). Weir's socio-cognitive framework was used in this research to provide a more systematic and manageable approach to validation. To further narrow the limitless scope of validity investigation, it is useful to consider whether there is any evidence to nullify the inferences, thereby undermining the ABV. The inference that all swans are white is a simple example that can be used to demonstrate this logic of falsification. The inference is impossible to verify as it would involve observing *all* the swans in the universe, however, it would be falsified by observing a single black swan. The focus of the ABV constructed over the previous chapters was thus to determine potential rebuttals and weak links in the GCE 1162 reading examination.

The second point to be raised concerns the nature of an assessment procedure which may be compared to building a house of cards. Any incorrectly positioned or missing card inherently weakens the structure and may cause the house of cards to collapse. A high-stakes national examination has a more complicated and intricate structure involving more cards at each level than for example, a classroom-based formative spelling assessment. Once a card has been misplaced, it cannot be substituted for by another card; each card needs to be accurately positioned, individually and as part of an ensemble, to ensure the soundness of the entire construction. Similarly, threats to validity identified at each stage of the assessment procedure cannot be compensated

at subsequent stages—this is not unlike Reid’s (2011) assertion that in every chain of reasoning, the evidence of the last conclusion can be no greater than that of the weakest link of the chain, whatever may be the strength of the rest.

Bearing the above in mind, this section revisits the inferences made by the Singapore Examinations and Assessment Board (SEAB) and Singapore’s Ministry of Education (MOE) with regard to the GCE 1162 reading examination. The assumptions underlying the four inferences were assessed in Chapters 5 to 8 and a provisional status of accepted, rejected or not investigated was awarded to each assumption (see Figures 9a to 9d). In this section an overall validity evaluation of the GCE 1162 reading examination is reached, thereby answering the main research question, ‘To what degree have the intended measurement objectives of the GCE 1162 been achieved?’ In order to attain an answer, two final steps are needed (Shaw & Crisp, 2012) and these are taken in the rest of this section.

Step one involves holistically assigning an evaluation status to each inference as a whole: justified, defeated or unevaluated. Step two relates the evaluation statuses to a comprehensive judgement of the strength of the ABV in its entirety and hence, the measurement quality of the GCE 1162 reading examination. It is important to recognize the following three caveats that affect the two final steps of evaluation. The first caveat relates to the notion of an ‘integrated evaluative judgement’ (Messick, 1992) which presupposes no single truth. As Newton (2017a: 63) acknowledges, ‘different evaluators might well reach different judgements, even on the basis of the same corpus of evidence and analysis’. While the validity evidence gathered in this study focuses on a priori aspects of the examination, the second caveat relates to the a posteriori parameters which need to be considered when building a complete ABV, including a social consequential dimension. The third caveat is that the impact of a high-stakes national examination is widespread, therefore, reaching a definitive verdict of sufficient validity is necessarily a collective public responsibility. The fact that in reality validation studies are normally undertaken by an individual or a small group of academics and specialists poses a dilemma. This study has attempted to mitigate the dilemma by taking into consideration the views and opinions of numerous stakeholders, however, the study needs to be complemented by validation studies of a larger scale.

Inference	Validation question	Claim	Assumptions	Provisional evaluation status
1. Specifications and administration	Are the intended purposes, constructs and administrative procedures of the examination clearly and sufficiently articulated?	The intended purposes, constructs and administrative procedures of the examination are clearly and sufficiently articulated.	a. The purposes of the examination are indicated.	Accepted with concerns
			b. It is possible to identify the primary purpose(s) when there is a multiplicity of purposes.	Plausible rejection
			c. Purposes attributed to the examination are achievable and non-conflicting.	Plausible rejection
			d. Purposes for which the results are unfit are indicated.	Rejected
			e. The constructs of the examination are indicated.	Accepted with concerns
			f. Detailed explanations of what the constructs entail are given.	Plausible rejection
			g. The constructs reflect a general consensus of the views of experts in relevant fields with specific consideration of Singapore's context.	Accepted with concerns

			h. The constructs align with the recommendations and learning outcomes of the broader curriculum.	Accepted with concerns
			i. Security procedures are in place to ensure confidentiality and fairness.	Accepted
			j. Feedback channels are available.	Plausible rejection
			k. Administrative procedures are documented and accessible for public scrutiny.	Rejected
			l. Intra and cross organizational collegiality and research are promoted.	Plausible rejection

Figure 9a: Summary of provisional evaluation status of the assumptions underpinning the specifications and administration inference relating to the GCE 1162 reading examination

Inference	Validation question	Claim	Assumptions	Provisional evaluation status
2. Test-taker characteristics	Are the characteristics and needs of Singaporean test-takers taken into consideration?	The characteristics and needs of Singaporean test-takers are taken into careful consideration.	a. The examination is supported by knowledge of adolescence and adolescent literacy.	Accepted with concerns
			b. The examination appeals to the reading interests of Singaporean adolescents.	Plausible rejection
			c. The examination is relevant and authentic to Singaporean adolescents, paralleling their real life needs.	Plausible rejection
			d. The examination takes into account new forms of reading literacy.	Plausible rejection

Figure 9b: Summary of provisional evaluation status of the assumptions underpinning the test-taker characteristics inference relating to the GCE 1162 reading examination

Inference	Validation question	Claim	Assumptions	Provisional evaluation status
3. Cognitive parameters	Are the cognitive requirements of the GCE 1162 reading examination appropriate and do the reading constructs sampled indicate broader competence beyond the examination?	The cognitive requirements of the examination are appropriate and the reading constructs sampled indicate broader competence beyond the examination.	a. The examination takes into account the different dimensions of reading assessment (e.g. text comprehension, knowledge and application of language and literature, multiple text reading for problem-solving, and reading volume and interest).	Plausible rejection
			b. There is adequate representation of lower-order thinking items (LOT).	Accepted with concerns
			c. There is adequate representation of higher-order thinking items (HOT).	Plausible rejection
			d. There is adequate representation of items at each specific cognitive level (remember, understand, apply, analyse, evaluate and create).	Plausible rejection
			e. The examination takes into account different reading levels (local and global).	Accepted with concerns

			f. The examination takes into account different reading types (expeditious and careful).	Accepted with concerns
			g. Statistical analyses are employed in field testing to refine items in the actual examination.	Accepted with concerns
			h. There is alignment between the measurement objectives of the examination and the learning objectives in the syllabus.	Accepted with concerns
			i. The examination assesses constructs that are relevant to real-life reading contexts beyond the syllabus.	Plausible rejection

Figure 9c: Summary of provisional evaluation status of the assumptions underpinning the cognitive parameters inference relating to the GCE 1162 reading examination

Inference	Validation question	Claim	Assumptions	Provisional evaluation status
4. Contextual parameters	Are the characteristics of the test items and passages appropriate and fair?	The characteristics of test items and passages in examination are appropriate and fair.	a. A variety of suitable item types is employed to assess reading constructs.	Accepted with concerns
			b. Mark schemes are well defined for differentiating the quality of answers.	Plausible rejection
			c. There is adequate representation of passages of different discourse modes and text purposes.	Accepted with concerns
			d. There is adequate representation of passages with different propositional content.	Plausible rejection
			e. Passages in general possess literary merit.	Plausible rejection
			f. Passages in general are of a suitable readability level.	Accepted with concerns
			g. The contextual parameters of the examination support the assessment of constructs that are relevant to real-life reading contexts beyond the syllabus.	Plausible rejection

Figure 9d: Summary of provisional evaluation status of the assumptions underpinning the contextual parameters inference relating to the GCE 1162 reading examination

Figures 9a to 9d present the provisional evaluation status of each assumption, judged using supporting evidence and rebuttals presented in Chapters 5 to 8. Building on these provisional evaluation statuses of the 32 assumptions, the first step now is to assign an evaluation status to each of the four inferences as a whole (Shaw & Crisp, 2012; Verheij, 2005). The evaluation status of an inference is justified when a claim and supporting evidence are accepted and rebuttals rejected or when the supporting evidence is stronger than the rebuttals. Conversely, the evaluation status is defeated when a claim and evidence are rejected and rebuttals accepted or when the rebuttals are more convincing than the supporting evidence. If the assumptions underlying the inference are not investigated, the evaluation status of the inference will be unevaluated. Evaluations of the four a priori inferences are shown below and summarized in Figure 9e.

The first of these is the specifications and administration inference. The associated sub research question asks, ‘Are the intended purposes, constructs and administrative procedures of the examination clearly and sufficiently articulated?’ The provision of detailed examination purposes, constructs and administrative procedures for the GCE 1162 reading examination, in documents such as the syllabus and test specifications, are essential in preventing the inappropriate use of test scores, construct underrepresentation and irrelevance, and miscommunication among stakeholders. Data from semi-structured interviews and document analysis suggested that improvements are needed to address the issues of ambiguous or undocumented examination purposes and constructs before the inference can be considered justified. A transparent administrative system, maintained by specifying, for example, the selection criteria for item setters and advisers, and procedures for determining the cut-scores, was also lacking as argued in the VA. The evaluation status assigned to the specifications and administration inference is, therefore, plausibly defeated.

The second is the test-taker characteristics inference and corresponding sub research question, ‘Are the characteristics and needs of Singaporean test-takers taken into consideration?’ Reading proficiency can only be interpreted meaningfully if the GCE 1162 reading examination demonstrates awareness of the relevant test-taker characteristics. Evidence gathered from semi-structured interview and document analysis supported the assumption that the GCE 1162 reading examination is

designed with adequate knowledge of adolescence and adolescent literacy. Several threats to the VA were, however, identified, including the relatively low appeal of the passages to test-takers, their relevance, authenticity and sensitivity to new forms of reading literacy. The test-taker characteristics inference is, hence, awarded an evaluation status of plausibly defeated.

The third a priori inference is the cognitive parameters inference. The accompanying sub research question of the inference asks, ‘Are the cognitive requirements of the GCE 1162 reading examination appropriate and do the reading constructs sampled indicate broader competence beyond the examination?’ Data from semi-structured interviews, document analysis and expert judgement relating to 22 sets of GCE 1162 reading examination papers revealed that the test design takes into account the different reading levels, reading types and learning objectives in the syllabus. In addition, statistical analyses, including both Classical Test Theory and Item Response Theory, are regularly carried out by SEAB during field tests to refine items in the actual examination. An evaluation status of justified with concerns is accorded to the cognitive parameters inference, bearing in mind that the inadequate representation of reading dimensions and higher-order thinking items at the analyse, evaluate and create levels are threats to the VA.

The fourth is the contextual parameters inference which addresses the sub research question, ‘Are the characteristics of the test items and passages appropriate and fair?’ Data amassed by semi-structured interview, document analysis and expert judgement underpinned strengths in the VA such as varied item types, passages of suitable readability level, and passages of different discourse modes and text purposes. These strengths were, unfortunately, offset by rebuttals that included a less than complete and accurate mark scheme, a general tendency for passages to focus heavily on only values and attitudes and passages that scored low on literary merit. The contextual parameters of the examination also offered insufficient support for the assessment of constructs relevant to real-life reading contexts beyond the syllabus. As such, the evaluation status given to the contextual parameters inference is plausibly defeated.

As summarized in Figure 9e, of the four inferences, one is justified with concerns and three are plausibly defeated. As the GCE 1162 examination is a high-stakes

national examination, the burden to provide strong and credible IAs and VAs is similarly high-stake. For this reason, ‘when assigning an evaluation status any error must be made in favour of the test-taker and test-user’ (Henning, 2014: 218), with the intent of improving the measurement quality of the GCE 1162 reading examination.

Inference	Evaluation status
1. Specifications and administration	Plausibly defeated
2. Test-taker characteristics	Plausibly defeated
3. Cognitive parameters	Justified with concerns
4. Contextual parameters	Plausibly defeated

Figure 9e: Summary of evaluation status of the four a priori inferences underpinning the GCE 1162 reading examination

The evaluation statuses of the four a priori inferences are drawn together in the second pivotal step to answer the main research question, ‘To what degree have the intended measurement objectives of the GCE 1162 reading examination been achieved?’ The main research question, as explained in Chapter 1, is essentially a question about validity. The present study adopts the definition of validity as ‘*fundamentally* a measurement concept, tantamount to measurement quality’ (Newton, 2017a: 11, original emphasis), although it recognizes that validity, as construed by Messick (1989b), comprises a social consequential facet. The focal point of this study is, therefore, upon constructing an ABV founded on inferences internal to the examination itself. In the ABV, the IAs and VAs underlying the four inferences are backed up by the supportive nature of some of the validity evidence collected. Rebuttal evidence, however, suggests considerable threats to validity that have to be addressed in order to strengthen the ABV which in turn gives credence to the measurement quality of the examination. The overall conclusion that the measurement quality of the GCE 1162 reading examination is at a moderately

unsatisfactory level is cautiously made.¹ In other words, while the ABV assumptions are partially substantiated, major weaknesses in the ABV are also detected.

9.3 Beyond the a priori inferences

An ABV premised on the a priori inferences of specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters was established in the study. The four a priori, or before-the-test, inferences are by no means exhaustive and in this subsection, a short discussion of three a posteriori, or after-the-test, inferences is offered. The three chosen a posteriori inferences, scoring, criterion-related, and washback and impact, form part of Weir's (2005) socio-cognitive validity framework as outlined in Chapter 1. The interactions between these a priori and a posteriori inferences, though not within the scope of this study, may well furnish further insights into the validity evaluation of the GCE 1162 reading examination.

9.3.1 Scoring

The first a posteriori inference to be sketched out is scoring. The scoring inference accounts for the extent to which test scores for a group of test-takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test-taker (Berkowitz, Wolkowitz, Fitch & Kopriva, 2000). Put differently, the scoring inference can be seen as a superordinate term for all aspects of reliability (Weir, Vidaković & Galaczi, 2013). Bachman (1990) contends that reliability is fundamentally about minimizing the effects of measurement error and attaining a test score as close as possible to the true score, a hypothetical score a test-taker would obtain if no error enters an examination.

¹ As Newton (2017a: 63) maintains, validity, or measurement quality in the present study, being 'a property that comes in degrees', can only be quantified 'impressionistically, using categories like "very low", or "low" or "moderate" or "high" (or other such terms)'. The decision to use any of these terms must, however, be backed by a strong ABV, which the study has sought to construct.

Further, scales used in project evaluation studies, such as the six-point scale (Highly satisfactory, satisfactory, moderately satisfactory, moderately unsatisfactory, unsatisfactory and highly unsatisfactory) used by the Operations Evaluation Department of the World Bank (Liebenthal, Michelitsch & Tarazona, 2005; The World Bank, 2005) could possibly be modified and used to rate the overall strength of an ABV.

The smaller the measurement error, the greater the reliability and the more substantiated the scoring inference.

There are three main sources of measurement error, namely, test-takers, markers and the examination itself. First, test-takers being human are not always consistent in their performance. Changes in motivation, concentration and health are likely to affect the quality of their responses. Test-takers may know the answer but fatigue, distractions, carelessness, omission and misinterpretation of instructions may prevent them from giving a correct response. On the other hand, test-takers may make random guesses and get some items right by chance. Second, markers are a source of potential error. In large-scale assessments like the GCE 1162 examination which involves a large group of markers, the problem of inter-rater reliability is made more acute. There must be a degree of agreement among markers, that is to say answers of the same quality will be given the same score by different markers. There is also the issue of intra-rater reliability or the degree of scoring consistency among examination scripts marked by a single marker. Third, there are test-specific sources of error. As illustrated in earlier chapters, an examination needs to be a good representation of the constructs, free of bias and administered fairly and ethically to yield reliable scores.

Internal consistency coefficients, such as the Cronbach Alpha, are commonly used to assess the reliability of test scores for a group of test-takers.² Generally accepted as ‘the industry standard’ (Khalifa & Weir, 2009: 148), the Cronbach Alpha is expressed as a function of the number of items in a test, the average covariance between pairs of items and the variance of the overall score. The Cronbach Alpha ranges from 0.00 to 1.00—the higher the value, the higher the estimated level of reliability. When interpreting the Cronbach Alpha, it must be borne in mind that higher levels of reliability can be expected for examinations with more items,

² Other than internal consistency coefficients, test-retest coefficients and alternative-form coefficients are often used too to evaluate reliability. The former are simply obtained by administering the same examination to a group of test-takers twice and correlating the scores. The latter are derived by administering a parallel and comparable form of the examination to the test-takers in the second seating. While these two types of coefficient represent intuitively appealing procedures to estimate reliability, they are not without serious limitations. For example, reactivity may compromise the reliability of the second examination as the test-takers can become sensitized to and familiarized with the items. Unlike the test-retest and alternative-form coefficients, internal consistency coefficients can be calculated with a single test administration.

relatively uniform item types and test tasks, and a wide range of test-taker ability. Most high-stakes examinations report Cronbach Alpha coefficients that exceed 0.80, although a very high Cronbach Alpha coefficient (exceeding 0.90) may be neither possible nor desirable for reading examinations like the GCE 1162 reading examination which employ a wider variety of passages and item types (Khalifa & Weir, 2009).

The standard error of measurement is an additional reliability statistic that is perhaps more functional than internal consistency coefficients for making confident and defensible decisions about test-takers with borderline scores. Conceptually, the standard error of measurement is used to determine a band around a test-taker's score within which that test-taker's true score would probably fall if the examination were administered to them repeatedly. It can be computed using the test score standard deviation and internal consistency coefficient. On the basis of the standard error of measurement, an examination board can estimate how far test-takers' scores would vary by chance alone if they were to take the examination over and over again. With all other factors held constant, the narrower the standard error of measurement is, the narrower the band of possible variation will be, or the more consistently raw scores will reflect the test-takers' actual proficiency.

The *Standards for Educational and Psychological Testing*, published by the American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME) (2014), describes at length the guiding principles for ensuring reliability and thereby upholding the scoring inference. The *Standards for Educational and Psychological Testing* advocates that within feasible limits, 'appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use' (AERA et al., 2014: 42), specifically 'Standard 2.3: For each total score, subscore or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported' (AERA et al., 2014: 43) and 'Standard 2.14: [...] Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score' (AERA et al., 2014: 46). It remains unclear whether acceptable levels of reliability are being met for the reading paper and other components of the GCE 1162

examination as neither the internal consistency coefficients nor the standard errors of measurement are publicly disclosed. No documented reasons seem to be available to explain this deficiency. The need to provide stakeholders and users of the GCE 1162 reading examination with sound and transparent evidence for the scoring inference is in accordance with the argument put forward throughout the study for a less opaque examination system.

9.3.2 Criterion-related

This subsection continues the thread of a posteriori validity evidence by presenting a concise discussion of the criterion-related inference. The criterion-related inference is validated when equivalence of different forms of the same examination can be demonstrated and a relationship drawn between the examination and some external criterion believed to be a measure of the same proficiency (Khalifa & Salamoura, 2011). In other words, comparability is central to this discussion about the criterion-related inference of the GCE1162 reading examination. There are five aspects to be considered.

The first of these aspects is comparability between parallel versions of the GCE 1162 reading examination. Parallel versions of the examination are designed in conformance to the same test specifications and syllabus, and administered under the same conditions. For example, between the years 2012 and 2015, there are eight parallel versions of the GCE 1162 reading examination, with two examinations administered per year (see Figure 4f). Even with parallel versions of an examination, the overall difficulty of each examination still varies, that is, the mean score obtained by a random sampling of test-takers would differ. To compensate for the differences in difficulty, statistical methods such as equating are used so that scores and grades may be used interchangeably across all parallel versions of the examination (Newton, Baird, Goldstein, Patrick & Tymms, 2007; Kolen & Brennan, 2004).

A second aspect is the less straightforward situation of comparability between non-parallel versions of the GCE 1162 reading examination. When, for example, changes were made to the GCE 1162 reading examination format in 2012 and 2016 or to the Chinese as a second language (CL2) syllabus in 2011, the examination papers before

and after the changes are non-parallel versions of the examination. Scores from these non-parallel versions of the GCE 1162 reading examination can no longer be equated as the reading constructs measured would expectedly have changed. Linking can be used instead to put the scores from two or more non-parallel examinations on the same scale (Kolen & Brennan, 2004).

Another aspect is the highly complex concept of comparability between the GCE 1162 examination in general, and its reading paper specifically, with other Singapore-Cambridge General Certificate of Education Ordinary-Level (GCE O-Level) examinations. For instance, how does the GCE 1162 reading examination compare with the GCE 1116 Higher Chinese and GCE 1153 Chinese Language 'B' reading examinations? Is grade A1 in the GCE 1162 examination of the same standard as grade A1 in other mother tongue examinations such as the GCE 1148 Malay as a second language examination and the GCE 1157 Tamil as a second language examination? In sum, SEAB is expected to ensure comparability of examination standards between and across all GCE O-Level examinations from one year to the next.

A fourth aspect to review is cross-test comparability. Taylor (2004b) posits that comparing an examination with other available examinations which claim to measure similar constructs offers important information to test users and stakeholders. Key cross-test comparability investigations include internal studies between the Cambridge English as a second language examinations and examinations offered by the Educational Testing Service (Bachman, Davidson, Ryan & Choi, 1995), and between the International English Language Testing System and Cambridge Main Suite Examination (Taylor, 2004a). More recently, Zhu (2015) examines the similarities and differences between reading examinations in three large-scale assessments, namely the Programme for International Student Assessment, the Progress in International Reading Literacy Study and the National Assessment of Educational Progress. Throughout the present study, comparisons were made, when appropriate, between the GCE 1162 reading examination and the Cambridge International General Certificate of Secondary Education (IGCSE) and International Baccalaureate (IB) suite of CL2 examinations, as well as the Hanyu Shuiping Kaoshi (HSK 汉语水平考试), specifically Levels 4 to 5. Much more extended research is

needed, of course, to map the GCE 1162, IGCSE, IB and HSK CL2 reading examinations onto a comparison scale. As pointed out in Chapter 6 on test-taker characteristics, the world is becoming ‘flat’ (Friedman, 2007), enabling Singaporean adolescents easier access to global study and work opportunities. Overseas admission officers and employers will need to know how the GCE 1162 examination measures up to that of other CL2 qualifications awarded to applicants, such as the IGCSE, IB and HSK.

The fifth aspect is comparison with external standards, such as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2018). Since its publication in 2001, the CEFR has become highly influential in Europe and beyond. As Khalifa & Salamoura (2011) observe, there is a growing interest for examination boards to link their examinations to the CEFR in the field of language assessment. In Singapore, the CEFR project was implemented in the early 2010s by MOE and SEAB in accordance with the Mother Tongue Languages Review Committee’s recommendation of ‘[providing] proficiency descriptors to more explicitly spell out the language skills and levels of attainment our students should achieve at various key stages of learning’ (MTLRC, 2011: 15). Some of the preliminary outcomes of the CEFR project were shared by MOE at the 2014 International Association for Educational Assessment conference held in Singapore (Wang, Lee, Lim & Lea, 2014). The official set of proficiency descriptors, findings, and technicalities behind linking national CL2 examinations such as the GCE 1162 examination to the CEFR —strong supporting evidence for the criterion-related inference— were, however, not publicly released.

In conclusion, to substantiate the criterion-related inference for the GCE 1162 reading examination, SEAB has to take into consideration all five aspects of comparability discussed above. Ensuring that the examination is equated or linked across parallel and non-parallel versions, as well as with other GCE O-Level examinations and external criteria is essential when results of the GCE 1162 reading examination have significant consequences for test-takers and are rarely reversible. This is the point in the narrative where it becomes necessary to look beyond the design, administration and technicalities of the examination into the wider world where national examinations and grades have far-reaching washback and impact.

9.3.3 Washback and impact

An account of Singapore's success story was provided in Chapter 3 on the Singaporean context. As described by the Organization for Economic Cooperation and Development (OECD, 2011: 160), Singapore has rapidly transformed from a third world nation to a 'gleaming global hub of trade, finance and transportation' in less than five decades, an economic miracle often associated with Singapore's strong conviction for a world-class education system. Singapore's highly regarded education system produces students consistently ranked top by international education indicators, for example, in 2015 students came first in reading, mathematics and science in the influential Programme for International Student Assessment rankings (OECD, 2016). While these stellar results paint an impressive picture of rapid development of and commitment to education, Singaporean students today are under immense pressure to perform well at national examinations that control access to elite schools, scholarships, future jobs and social status (Teo, 2017; Zaccheus, 2017; Teng, 2016). Recent findings from OECD suggest that Singaporean students experience significantly higher levels of anxiety over examinations and grades compared to their peers from 71 other countries and economies (Davie, 2017). The Singaporean examination system, with its streaming of students through the Primary School Leaving Examination and Singapore-Cambridge General Certificate of Education Ordinary-Level and Advanced-Level Examinations, has been infamously likened to 'a pressure cooker' (Teng & Yang, 2016).

In the context of Chinese language learning and testing in Singapore specifically, many academics (e.g. Curdt-Christiansen, 2014; Guo 2011; Chew, 2007) have illustrated the limitations of Chinese functioning as a second language in a predominantly English as a first language environment. Under such circumstances, Guo (2011) proffers that it is hardly surprising that not many Singaporean students acquire genuine communicative competence in CL2, let alone become lifelong learners and readers of the language. Singaporean students and their parents may expend considerable effort in doing well at CL2 national examinations such as the GCE 1162 examination, yet may end up frustrated over unsatisfactory grades. Recommendations for a more differentiated CL2 curriculum using authentic materials and communicative-approach pedagogy have been made (e.g. Chin 2016;

Tan 2016). To date, however, few studies have been conducted to evaluate the CL2 national examinations in Singapore and to provide any alternative solution to their possible negative washback and impact.

As early as the 1870s, Latham (1877) described examinations as having profound implications for students, teachers, parents and members of the public. Madaus (1988) and Pearson (1988) elaborated that examinations are instrumental in shaping the motivation and behaviours of students and teachers as well as educational goals and processes. In 1989, Messick (1989b) first introduced the concept of consequential validity to the validity argument as outlined in Chapter 1, maintaining that it is necessary to ascertain whether the intended and unintended social consequences of test use are consistent with test purposes and social values. Messick's concept of consequential validity is generally understood to encompass washback and impact (Khalifa & Weir, 2009). Washback can be defined as the influence tests have on teaching and learning, focusing specifically on the narrower contexts of the classroom and school (Hamp-Lyons, 2000; Wall, 1997). Impact on the other hand is concerned with the wider influences of tests on the community and society at large and can be seen as a superordinate which subsumes washback (Hawkey, 2006; McNamara, 2000).

As Cheng, Watanabe and Curtis (2004) encapsulate, test washback and impact have become an integral part of validation studies and will continue to be so. Washback and impact warrant even more attention, perhaps, in Confucian societies like Singapore where academic success is highly valued as a means of acquiring public office and for achieving moral perfection, often defining self-worth. Sim (2014) draws attention to the common perception that pragmatism remains the key philosophy driving teaching in Singapore, that teachers teach in the way they believe will help more students to pass their examinations. Sim's sentiments are echoed by many of the interviewees in this study who affirm that rather than the official curriculum determining what is taught and how it is taught, it is the national examinations that dictate what is learned and how it is learned. Additionally, washback and impact will be most intense where test-takers and stakeholders perceive a test as challenging and the results as important (Green, 2007), as observed in the case of the GCE 1162 reading examination.

To conclude, understanding of the GCE 1162 reading examination may greatly benefit from research into its washback and impact. A potential area of focus is how the examination can be re-designed and administered, taking into consideration the findings presented in this present study, to bring about beneficial changes for test-takers, stakeholders and Singaporean society at large. In essence, the challenge lies in producing a GCE 1162 reading examination worth teaching to, such that success in the examination translates into proficiency in the reading dimensions, skills and approaches which Singaporean society wishes to encourage. Attention must also be given to providing transparent and comprehensive information, and supporting stakeholders throughout the examination process (Saville, 2003). It is only when examinations are ‘properly conceived and implemented’ (Popham, 1987: 680) that positive social consequences become achievable. Other plausible areas of focus include increasing the assessment literacy of teachers (e.g. Zhang & Soh, 2016; Kunnan & Zhang, 2015), constructing a test agenda for achieving fairness, equity and social justice (e.g. Kunnan, 2008, 2004), and monitoring standards and ensuring accountability (e.g. Newton, 2008; Linn, 2000). A word of caution is necessary, however, that whilst much investigation can be done to identify and avoid some of the possible negative washback and impact, there is a limit to what is attainable or even perhaps desirable (Davies, 1997). In other words, it is not possible for policy makers and examination boards to take account of all possible social consequences.

9.4 Directions for future research

The three a posteriori inferences, scoring, criterion-related, and washback and impact, suggest areas for future validation studies of the GCE 1162 reading examination. In addition, further validity evidence can be obtained by using other research methods such as protocol analysis, survey and statistical analysis of scores from the examination.

Data are generated in protocol analysis through a participant’s verbal response to a task or a probe. The act of reading a text in an examination involves cognitive processes that are not for the most part observable. By means of asking test-takers to articulate whatever goes through their minds, however, researchers can catch glimpses into the hidden cognitive processes which, in turn, help to form even more

accurate judgements about the measurement quality of an examination (Anderson, Bachman, Perkins & Cohen, 1991).

Surveys can also be designed to assess the prevalence of attitudes and knowledge with regard to the GCE 1162 reading examination. By utilizing surveys, a much larger population of stakeholders and test-takers can be sampled than is possible through in-depth interviews as carried out in the present study. Further, although scores from the actual examination are unlikely to be available for research, the GCE 1162 reading examination can be replicated on a smaller scale and the scores analysed statistically. Answer scripts from the replicated examination can also be studied to gain insights into how test-takers respond to items and how scripts are marked and scored.

Validation, as characterized by Messick (1989b), is an ongoing process of collecting evidence for an assessment. New developments in the fields of validity and validation, and reading, or reforms in assessment and education policies in Singapore will therefore necessitate re-examination of the inferences, claims, assumptions, supporting evidence and rebuttals in the study. Given the limitations of the present study, I would argue strongly that a comprehensive and cogent account of the measurement quality of the GCE 1162 reading examination has been presented. The chapter now concludes by drawing attention to the key implications and impact of the study.

9.5 Key implications and impact

Wood (1993: 151-152) in the influential publication *Assessment and Testing* asserted that:

The examining boards have been lucky not to have been engaged in validity argument. Unlike reliability, validity does not lend itself to sensational reporting. [...] Validation work is unglamorous and needs to be painstaking but has to be done. As long as the examination boards make claims that they are assessing this or that ability or skill, they are vulnerable to challenge from disgruntled individuals.

The primacy of validity in testing and assessment has been consistently affirmed over the last thirty years, with the *Standards for Educational and Psychological Testing* (AERA et al., 2014: 11) referring to validity as ‘the most fundamental consideration in developing tests and evaluating tests’. Whilst validity as theory has been carefully developed in countries such as the United Kingdom and United States of America, validity as practice, or validation, has often lagged behind (Stobart, 2012), due to the ‘unglamorous’ and ‘painstaking’ nature of validation and its susceptibility ‘to challenge from disgruntled individuals’ as recounted above by Wood. This relative lack of interest in validation becomes even more pronounced in the Singaporean context, where a survey of existing literature reveals hardly any published works on the validation of Singapore’s national examinations. The present study is therefore, as I have stated in Chapter 1, an important landmark as it provides the most extensive validation study, if not the first detailed evaluation, of a national examination in Singapore.

The study, driven by theoretical underpinnings in the fields of validity and validation, and reading, was contextualized within the Singapore assessment and education landscape. Weir’s (2005) socio-cognitive validity framework and Kane’s (2009, 2006) ABV were used to frame the study around four a priori inferences specifically, specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters. Methodologically rooted in philosophical pragmatism and employing a mixed methods research design, the study has generated findings that have significant practical and theoretical implications that are summarized below.

9.5.1 Practical implications and impact

The primary aim of the study was to evaluate the degree to which the intended measurement objectives of the GCE 1162 reading examination have been achieved. In answering this main research question, several threats to validity have been identified which SEAB and MOE need to ameliorate or help resolve. Rectifying these threats to validity will strengthen the ABV and, hence, the measurement quality or validity of the examination.

First, SEAB and MOE need to set forth clearly the purposes of the GCE 1162 reading examination, presenting the rationale behind how scores from the examination are intended to be interpreted and consequently used. Users of the examination should also be cautioned against utilizing scores for unsupported purposes. In particular, the recently added purpose of improving instructional guidance and bringing about meaningful learning by the Curriculum Planning and Development Division of MOE (CPDD, 2014) has to be revisited. Achieving this ambitious purpose requires the GCE 1162 reading examination to be designed as an examination worth teaching to, resulting in positive washback on teaching and students' reading habits.

Second, extensive theoretical and empirical research on CL2 reading should be spearheaded by SEAB and MOE. Further understanding of the reading needs, interests and challenges of Singaporean adolescents in the age of the Internet and new literacies will aid refinement of the constructs of the GCE 1162 reading examination. The reading constructs should be specified in detail and referenced against external criteria such as the IGCSE and IB suite of CL2 examinations, HSK and CEFR.

Third, the GCE 1162 reading examination could encompass more varied dimensions of reading assessment such as multiple text reading for problem-solving, and reading volume and interest. I have argued in Chapter 2 that Singaporean adolescents are increasingly required in the age of the Internet to process multiple complex texts to develop a lucid intertextual representation. Reconfiguring the examination to include multiple text reading for problem-solving better aligns it to real-life reading contexts. Additionally, extensive voluntary reading is indicative of fluent reading and MOE envisions nurturing a generation of life-long CL2 readers and learners (CPDD, 2011). MOE and SEAB could, therefore, explore ways in which the assessing of reading volume and interest can be integrated into the examination, for example, as school-based assessment in the form of a reading portfolio or presentation. Expected reading volume can be recommended in the syllabus, and official supplementary readers can also be designed and students tested on their understanding and appreciation of these readers in the GCE 1162 reading examination.

Fourth, semi-structured interview, document analysis and expert judgement data strongly suggest that the proportion of scores allocated to higher-order thinking (HOT) and lower-order thinking (LOT) items has not changed significantly over the past decade, between May 2006 and May 2016. Only 8.79% of items assessed are HOT items, accounting for 16.36% of the total score, indicating inadequate representation. This is despite recommendations made by the Chinese Language Curriculum and Pedagogy Review Committee (2004) to include more HOT items and reduce the proportion of items that assesses LOT skills such as factual recall (remember) and literal comprehension (understand). It is, therefore, imperative that test designers increase the proportion and weightage of HOT items if scores from the GCE 1162 reading examination are to be representative of comprehension at the analyse, evaluate and create cognitive levels as the Syllabus 2011 claims.

Fifth, subject matter experts (SMEs) and interviewees advocate the potential benefits of repackaging the GCE 1162 reading examination so that it can be more authentic and relevant to Singaporean adolescents. At present, passages with relatively low literary merit that explicitly expound values and attitudes are over-represented in the examination. A revamped examination should ideally select texts of higher literary value and from a wider range of propositional content. Suggestions put forth by the SMEs and interviewees include extracts from canonical texts, lifestyle articles, expository texts on science, geography and history intended for general readers and texts on contemporary culture including film, art and literature. More varied item types such as information transfer, matching headings, and summarizing and comparing multiple texts could also be considered for inclusion in the examination.

Sixth, tangible steps should be taken by SEAB and MOE in ensuring that mark schemes are improved and publicly available, especially for the short answer and open-ended items, to minimize any threats to the valid interpretation of scores of the GCE 1162 reading examination. Steps include designing mark schemes with a clear understanding of the reading constructs and an accurate prediction of the Outcome Space. Mark schemes should also state explicitly the principles to which markers must abide in order that responses of varying quality are scored consistently. Furthermore, items that do not adhere to the general principles of test design, such as

those with flawed distractors and keys or item stems that unwittingly provide clues to the correct answer, should be avoided.

Seventh, SEAB and MOE should urgently seek to promulgate a culture of transparency. Given the high-stakes nature of the GCE 1162 reading examination, all examination procedures, for example the selection criteria for item setters, markers and advisers, and statistical information such as cut-scores, internal consistency coefficients and standard errors of measurement should be defensible and publicly accessible. A move toward transparency as *Standards for Educational and Psychological Testing* advocates will empower stakeholders and test users to engage in robust discussions and research, improving the measurement quality of the examination in the long term. The dissemination of information could take the form of public forums, workshops and continuing professional development training by assessment specialists at SEAB and academics at partner organizations such as the National Institute of Education and the Singapore Centre for Chinese Language. Whilst SEAB has been raising literacy in formative assessment among Singapore's teachers and encouraging good assessment practices in the classroom (SEAB 2017c), more could be done to promote the systematic review and evaluation of national examinations. SEAB (2017c: 3) has recently announced:

SEAB has captured a wealth of knowledge, experiences and information assets over the years. These should be captured and stored effectively for knowledge sharing, providing a platform to incubate ideas and engender innovation, facilitate higher quality decisions and position SEAB for its next phase of growth and development.

This Knowledge Management System (SEAB, 2017c) of an established and highly regarded local and regional examination board is much anticipated by SMEs and interviewees and will, I believe, facilitate more validation studies like the present study.

9.5.2 Theoretical implications and impact

The present study has not only uncovered the strengths and limitations of the GCE 1162 reading examination and offered recommendations for improvement in its measurement quality—the implications and impact of this study go beyond these practical aspects at the micro-level. The study has provided extensive illumination of perspectives on and understandings of validity and validation, the reading construct and the Singaporean context. It might not be feasible for a full-scale validation study such as this to be carried out routinely for all subjects examined by SEAB. As one of the first detailed investigations of a national examination in Singapore, however, the study offers a research foundation and viable frameworks from which smaller-scale and more routine validation studies could be developed.

In the course of my analysis of the GCE 1162 reading examination, I have observed an inextricable relationship at the meso-level between the validation process and the context in which it is carried out. Messick (1988: 43) foreshadowed these concerns when acknowledging:

The practical use of measurements for decision making and action is or ought to be applied science, recognizing that applied science always occurs in a political context. Indeed, social and political forces are sometimes so salient that we may need a new discipline to deal explicitly with the politics of applied science.

Since Messick (1989b), the washback and impact of examinations on test-takers, institutions and society have been a source of concern as briefly considered earlier in this chapter. There is, however, little discussion in research literature about how test-takers, institutions and society can influence and shape the process of validation itself, thereby facilitating or hindering it. The relationship is bi-directional, as hypothesized in Chapter 1, rather than linear. The specific social, political, cultural and educational environments in which validation occurs will inevitably determine its feasibility and meaningfulness.

Validating a national examination in Singapore is no easy feat as control has been enforced by SEAB and MOE to circumscribe what evidence can be collected and published. Official statements about national examinations are often ambiguous and unofficial narratives are laborious to obtain. Singapore might risk becoming, as Scott, Posner, Martin and Guzman (2015: 131) describe, ‘an extraordinarily sensitive society’ where ‘the government [...] is assiduous in sanctioning only those projects, schemes, programmes of work and enterprises that it favours.’ Statism might have served the young nation well in the past, but as Ho (2016) and Mahbubani (2015) caution, red tape, or excessive bureaucracy and reluctance to share information, will be a significant impediment to research and development. By contextualizing the validation of the GCE 1162 reading examination, this study has manifested that validation processes are socially situated and that validation responsibilities need to be publicly documented and co-ordinated.

At the macro-level, the present study has established the adequacy of Weir’s (2005) socio-cognitive validity framework and Kane’s (2009, 2006) ABV for amassing validity evidence in ways which are feasible. In this study, a unitary view of validity was assumed as perceived by Messick (1996, 1994, 1989b). A unified approach to validity has, nevertheless, long been criticized for offering little by way of usable advice to those working on test evaluation and construction in the field of assessment (Lissitz, 2009). Weir’s socio-cognitive validity framework was therefore adapted in order to conceptualize the validation process within a temporal frame thereby identifying four a priori inferences for which evidence was to be collected. A separate chapter was then dedicated to each a priori inference namely, specifications and administration, test-taker characteristics, cognitive parameters and contextual parameters. In line with the most recent progress in thinking about validation, supporting evidence and rebuttals were subsequently organized into a persuasive measurement argument using Kane’s ABV comprising IAs and VAs with inferences, claims and assumptions. Kane’s ABV in particular serves to remind that any argument is only as strong as its weakest link, and it is thus reasonable for a validation study to target the likely weaknesses in the argument as is the case in this study. As knowledge of validity and validation, the reading construct and the Singaporean context advances, new weaknesses undermining the argument will be diagnosed as demanding attention. In sum, at the macro-level, Weir’s socio-cognitive

validity framework and Kane's ABV have been shown able to accommodate and strengthen the validation research of a CL2 reading examination in the Singaporean context.

I have become increasingly aware in the writing of this research of the importance and complexity of validation studies. Student interviewee Omega candidly commented during the interview that '*Cher*,³ Singaporean students were born to be tested!', and there is much truth, I believe, in this remark. National examinations are, indeed, indispensable measurement tools in Singapore, where they remain inseparable from teaching and learning. Drawing on Foucault's analytics of power, national examinations can be seen as possessing strong potential as mechanisms for social control. In Chapter 3 on the Singaporean context, I provided an account of the Foucauldian perception of the invisibility and pervasiveness of power in modern society. Foucault (1980: 39) argued that disciplinary power in modern society 'reaches into the very grain of individuals, touches their bodies and inserts itself into their actions and attitudes, their discourses, learning processes and everyday lives'. National examinations, as integral components of the Singaporean education system, are avenues through which 'a synaptic regime of power, a regime of its exercise *within* the social body rather than *from above* it' (Foucault, 1980: 39, original emphases) is established. Combining the techniques of hierarchical surveillance and normalizing judgement, national examinations enable society to construct individuals in particular ways, rendering them easily supervisable, efficient and productive (Foucault, 2012).

The first technique of disciplinary power is hierarchical surveillance. The national examinations monitor students, identify students who are not performing adequately, regulate students' behaviour and enable comparisons to be made. Such knowledge is then used to document these individuals, exerting powerful effects on their lives by judging and controlling what they are, and are not, eligible to do, subsequently streaming them into relatively permanent places in society. As Foucault (2012: 189) postulated:

³ *Cher* means *teacher* in vernacular Singaporean English.

The examination [...] introduces individuality into the field of documentation. [...] The examination that places individuals in a field of surveillance also situates them in a network of writing; it engages them in a whole mass of documents that capture and fix them. The procedures of examination were accompanied at the same time by a system of intense registration and of documentary accumulation.

The disciplinary power of national examinations is also enacted through the technique of normalizing judgement. Normalizing judgement occurs through comparison, such that individuals are compared to the norm which at once creates a field of differentiation (Foucault, 2012). In national examinations, normalizing judgement 'is most evident and familiar as a distribution of ability and as a concomitant typology of rank positions' (Ball, 2013: 51). National examinations, as articulated by many of the interviewees in this study, dictate not only the criteria against which students are measured, but also teachers and schools. Normalizing judgement imposes on all stakeholders a notion of objectivity that acts to bind them to a truth about national examinations. This truth in turn shapes the stakeholders' perception of what constitutes legitimate and useful knowledge, and the identities and relative worth of students, teachers and schools.

In summing up Foucault's description of ways through which society imposes discipline, Shohamy (2001) concludes that few devices are as powerful, or as capable of dictating as many decisions, as high-stakes national examinations. The authority of national examinations is even more pronounced in highly-centralized educational systems like Singapore that place a considerable premium on academic excellence. A single national examination score can independently trigger an automatic admission, promotion, placement, graduation and even future employment decision. Given the power of national examinations, especially their impact on the lives of individual adolescents and their capacity to participate fully in society, solutions for controlling their often unchallenged power must be sought (Shohamy, 2001).

At the point of writing, the Minister for Education, Ong Ye Kung (2018), has, at the Schools Work Plan Seminar 2018, announced major changes to keep the power of

examinations in check. This proposal stands to ameliorate the pressure many students, parents and teachers are facing. The first key change is the removal of all assessments and examinations for Primary One and Two students from 2019. The second change will affect Secondary One students who will no longer have a mid-year examination from 2019. From 2020 to 2021, there will be no mid-year examination for Primary Three, Primary Five and Secondary Three students. The third change focuses on report books in schools which will no longer reveal a student's position in relation to their class or cohort. In addition, MOE will set guidelines for schools such that there will be only one class test per subject per term that can be counted toward the year-end score. The fifth change is with the removal of assessments and examinations at Primary One and Two, MOE will adjust the academic criteria for its awards to recognize students' attitudes to learning such as diligence and collaboration. Minister Ong (2018) explained that 'students will benefit when some of their time and energy devoted to drilling and preparing for the examinations is instead allocated to preparing them for what matters to their future'. These recent changes are in tandem with the abolition of secondary school ranking in 2012⁴ and the changing of the Primary Six Leaving Examination (PSLE) scoring system from T-scores to wider grade bands in 2013.⁵ Together, these significant reforms pave the way for the next phase in Singapore's education—the 'learn for life' phase where there is less emphasis on examinations, drilling and competition, and where students derive more joy while learning, and learn for life (Ong, 2018).

Following Minister Ong's announcement on 28 September 2018, newspapers and social media in Singapore keenly debated the necessity of examinations. In the short span of a month, more than 40 commentaries and letters have been published in the two major local broadsheet newspapers, *The Straits Times* and *Lianhe Zaobao* (《联

⁴ The former ranking of Secondary schools in Singapore was based on student performance in the GCE O-Level examinations in the preceding year. This school ranking initiative, which was instituted in 1992, was replaced by school banding in 2004. Banding was in turn abolished in 2012 and replaced by a system designed to emphasize holistic education (MOE, 2012).

⁵ Prime Minister Lee Hsien Loong (2013) announced at the 2013 National Day Rally that the PSLE T-score will eventually be removed in 2021 and replaced with wider grade bands that reflect the student's individual performance and not their performance relative to their peers. In doing so, Prime Minister Lee reasoned that excessive stress and competition among students and parents will be reduced.

合早报》), for example Davie (2018), Long (2018) and The Lianhe Zaobao Editorial (2018). Several television and radio programmes about the reform in assessment have also been aired, including a panel discussion with Minister Ong (Huang, 2018). Public opinion is divided. There are parents and educators who opine that ‘standards will fall, academic rigour will be compromised and [...] [Singapore] risks cultivating a generation of weak-willed and under-motivated students, unable to cope with the pressures of competition and thus unprepared for the world’ (Kuah, 2018). Some critics even fear that as MOE’s focus shifts from examinations to niche talents, students from disadvantaged backgrounds may be left behind, thereby widening the gap between the lower and upper classes (Tan, 2018). On the other hand, there are others who applaud MOE’s bold attempt to redress the rigidity that has crept into the Singaporean education system, allowing students more space to be creative and imaginative by breaking away from grade obsession (The Lianhe Zaobao Editorial, 2018; The Straits Times Editorial, 2018). Kaur (2018) goes further by exploring the possibility of removing the national examinations. The key problem, Kaur (2018) maintains, ‘is not that there are too many exams [...] ultimately, it still comes down to the two big ones—the PSLE and GCE O-Level’, parents and educators ‘will continue to push [the students], in some cases, beyond breaking point’ if high-stakes national examinations were to stay.

What is the future for national examinations in Singapore? Minister Ong (2018) acknowledges that the Singaporean education and assessment system is undergoing a ‘quiet revolution’ despite being internationally lauded for its quality and robustness. It is certainly a time of change, as Singaporean academic Ng (2017: 41-42) incisively sums up:

[Singapore] has to abandon its obsession with learning for examinations. It is now focusing on learning for life, embracing holistic education, and developing its young people to think critically and creatively. [...] It is important to recognize the philosophy here. Singapore changes when it is still successful. Timely change occurs in anticipation of the future. It is change launched from a position of strength rather than one of desperation. But it takes courage to change when one is successful.

Indeed, it takes courage to change when the Singapore education and assessment system is widely seen to be successful. It also takes wisdom to retain effective measures and to anchor changes in timeless values. Minister Ong cautioned that as with all other major policy decisions, abolishing the national examinations involves tradeoffs which advocates ‘must be able to name and justify’ (Ng, 2018). Critics of the national examinations need to realize the potential of examinations and how they can lead to improvement in learning, teaching and policy planning, and uphold meritocracy, the cornerstone of Singaporean society. As Minister Ong pointed out, an inadequate understanding of tradeoffs is hampering deep discussion on national examination issues in Singapore (Ng, 2018). The way forward, therefore, is to promote a culture of transparency, knowledge transfer, and shared authority and responsibility. Such a culture would encourage stakeholders to develop a critical view of national examinations as well as to act on it by evaluating the quality of the examinations and questioning their purposes and consequences. Stakeholders would also be in a better position to engage in purposeful reviews of how Singapore’s national examinations could be designed and administered as examinations worth teaching to, and how they could be complemented by other forms of assessment. While there is still much work to be done toward constructing a quality discourse on national examinations in Singapore, I am confident that the validation study presented here is a meaningful step in this direction.

References

- Ahmed, A. & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice*, 18(3), 259-278.
- Alagumalai, S., Curtis, D. D. & Hungi, N. (2005). *Applied Rasch measurement: A book of exemplars*. Dordrecht: Springer.
- Alderson, J. C. (1978). *A study of the cloze procedure with native and non-native speakers of English*. Retrieved March 7, 2018 from https://www.era.lib.ed.ac.uk/bitstream/1842/6711/1/D076597_1.pdf
- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds), *Reading in a foreign language* (pp. 1-24). London: Longman.
- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Alderson, J. C. & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, 35, 79-113.
- Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L. & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York: Routledge.
- Alderson, J. C. & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, 5(2), 253-270.
- Alexander, P. A. & Fox, E. (2004). A historical perspective on reading research and practice. In R. B. Ruddell & N. J. Unrau (Eds), *Theoretical models and practices of reading* (pp. 33-68). Newark, DE: International Reading Association.
- All-Party Committee of the Singapore Legislative Assembly. (1956). *Report of the All-Party Committee of the Singapore Legislative Assembly on Chinese education*. Singapore: Government Printer.
- Allen, A. (2011). Michael Young's 'The rise of the meritocracy': A philosophical critique. *British Journal of Educational Studies*, 59(4), 367-382.
- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1974). *Standards for educational and psychological tests*. Washington, DC: APA.
- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: APA.

- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- American Psychological Association (APA). (1952). Technical recommendations for psychological tests and diagnostic techniques: Preliminary proposal. *American Psychologist*, 7(8), 461-475.
- American Psychological Association (APA), American Educational Research Association (AERA) & National Council on Measurement in Education (NCME). (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: APA.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Anderson, L. W. & Krathwohl, D. R. (Eds). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Anderson, N. J., Bachman, L. F., Perkins, K. & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, 8, 41-66.
- Anderson, R. C. & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research: A research annual* (pp. 231-256). Greenwich, CT: JAI Press.
- Ang, B. C. (2003). The teaching of the Chinese language in Singapore. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 335-352). Singapore: Eastern Universities Press.
- Angoff, W. H. (1971). Scale, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.
- Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp. 19-32). Hillsdale: Lawrence Erlbaum Associates.
- Argyris, C. (1999). *On organizational learning*. Oxford: Blackwell.

- Artelt, C., Schiefele, U. & Schneider, W. (2001). Predictor of reading literacy. *European Journal of Psychology of Education, 16*(3), 363-383.
- Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds), *The psychology of learning and motivation volume II* (pp. 89-195). New York: Academic Press.
- August, D. & Shanahan, T. (Eds). (2006). *Developing literacy in second language learners*. Mahwah: Lawrence Erlbaum Associates.
- Aw, G. P. (2015). Extensive reading (Part 1): Theory and teaching design [泛读教学实证研究（一）—理论与教学设计]. In G. P. Aw (Ed.), *A collection of empirical research studies in Chinese teaching and learning* [华语文教学实证研究—新加坡中小学经验] (pp. 271-292). Taipei: Wanjuanlou Publishing.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*, 671-704.
- Bachman, L. F., Davidson, F., Ryan, K. & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press.
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Baddeley A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423.
- Baddeley A. D. & Hitch G. J. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47-89). New York: Academic Press.
- Ball, S. (1994). *Education reform: A critical and post-structural approach*. Buckingham: Open University Press.
- Ball, S. (2013). *Foucault, power and education*. London: Routledge.
- Ball, S. (2016). *Michel Foucault and education policy analysis*. London: Routledge.
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J.: Prentice Hall.
- Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive theory*. Englewood Cliffs, N.J.: Prentice Hall.

- Barrett, M. (1992). Words and things: Materialism and method. In M. Barrett and A. Phillips (Eds), *Destabilizing theory: Contemporary feminist debates* (pp. 201-219). Cambridge: Polity Press.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.
- Beardsmore, H. B. (2003). Language shift and cultural implications in Singapore. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 85-98). Singapore: Eastern Universities Press.
- Bedlington, S. S. (1978). *Malaysia and Singapore: The building of new states*. London: Cornell University Press.
- Bell, S. M. & McCallum, R. S. (2008). *Handbook of reading assessment: A one-stop resource for prospective and practicing educators*. Boston: Pearson.
- Bellack, A. & Hersen, M. (1984). *Research methods in clinical psychology*. New York: Pergamon.
- Berkowitz, D., Wolkowitz, B., Fitch, R. & Kopriva, R. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy makers*. Washington: US Department of Education.
- Bernhardt, E. B. (2005). Progress and procrastination in second language reading. *Annual review of applied linguistics*, 25, 133-150.
- Bernhardt, E. B. (2011). *Understanding advanced second language reading*. New York: Routledge.
- Bernstein, B. (1971). *Class, codes and control: Theoretical studies towards a sociology of language*. London: Routledge & Kegan.
- Bingham, W. V. D. (1937). *Aptitudes and aptitude testing*. New York: Harper & Brothers Publishers.
- Birch, B. (2007). *English L2 reading: Getting to the bottom*. Mahwah: Lawrence Erlbaum Associates.
- Black, P. (2003). *Testing: Friend or foe? Theory and practice of assessment and testing*. London: Routledge Falmer.
- Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W.H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: David McKay Company.
- Bokhorst-Heng, W. (1998a). *Language and imagining the nation in Singapore*. (Unpublished doctoral thesis). University of Toronto, Toronto.

- Bokhorst-Heng, W. (1998b). Language planning and management in Singapore. In J. A. Foley (Ed.), *English in new cultural contexts: Reflections from Singapore* (pp. 287-309). Singapore: Oxford University Press.
- Borsboom, D., Cramer, A. O. J., Keivit, R. A., Scholten, A. Z. & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp.135-170). USA: Information Age Publishing.
- Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.
- Brandom, R. (2011). *Perspectives on pragmatism: Classical, recent, and contemporary*. Cambridge, MA: Harvard University Press.
- Bray, G. B. & Barron, S. (2004). Assessing reading comprehension: The effects of text-based interest, gender, and ability. *Educational Assessment*, *9*(3, 4), 107-128.
- Breen, M. (1985). Authenticity in the language classroom. *Applied Linguistics*, *6*, 60-70.
- Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R.L. Brennan (Ed.), *Educational measurement* (pp.3-16). Washington, DC: American Council on Education/Praeger.
- British Educational Research Association. (2011). *Ethical Guidelines for Educational Research*. Retrieved March 7, 2018 from <http://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-Ethical-Guidelines-2011.pdf>
- Britt, M. A. & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology*, *25*, 313-339.
- Bruner, J. S. (1966). *Toward a theory of instruction*. New York: Norton.
- Bruner, J. S. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.
- Bruning, R. H., Schraw, G. J. & Norby, M. M. (2011). *Cognitive psychology and instruction*. Boston: Pearson.
- Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R. & Curtis, S. A. (1921). Report of the Standardization Committee. *Journal of Educational Research*, *4* (1), 78-80.
- Burton, D. & Bartlett, S. (2009). *Key issues for education researchers*. London: Sage Publications.
- Cambridge Assessment. (n.d.). *About us*. Retrieved March 7, 2018 from <http://www.cambridgeassessment.org.uk/about-us/>

- Carroll, B. J. (1980). *Testing communicative performance*. Oxford: Pergamon Press.
- Carroll, J. B. (1961). Fundamental considerations in testing for English proficiency of foreign students. In J. E. Alatis (Ed.), *Testing the English proficiency of foreign students* (pp. 30-40). Washington, D.C.: Centre for Applied Linguistics.
- Carruthers, K. (2012). *The teaching and learning of Chinese in schools: Developing a research agenda to support growth*. Retrieved March 7, 2018 from https://ciforschools.files.wordpress.com/2015/01/2012-conference-paper_the-teaching-and-learning-of-chinese-in-schools_kcarruthers.pdf
- Carver, R. P. (1990). *Reading rate: A review of research and theory*. San Diego: Academic Press.
- Carver, R. P. (1997). Reading for one second, one minute, or a year from the perspective of reading theory. *Scientific Studies on Reading*, 1, 3-43.
- Chall, J. S. (1996). *Stages of reading development*. Fort Worth, TX: Harcourt Brace.
- Chall, J. S. & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.
- Chang, L. P. (2012). The study of the vocabulary size at the CEFR levels for CFL/CSL learners [对应于欧洲共同架构的华语词汇量]. *Journal of Chinese Language Teaching* [华语文教学研究], 9(2), 77-96.
- Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple... *Language Testing*, 29(1), 19-27.
- Chapelle, C. A., Enright, M. K. & Jamieson, J. (Eds). (2008). *Building a validity argument for the Test of English as a Foreign Language (TOEFL)*. London: Routledge.
- Chapelle, C. A., Enright, M. K. & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3-13.
- Cheah, C. M. (2003). *Teaching and assessment*. Singapore: Singapore Chinese Teachers' Union.
- Chen, D. Y. (2013, September 9). English has gradually become Singaporeans' mother tongue [英语已逐渐成为新加坡人的母语]. *Lianhe Zaobao* [联合早报], p. 16.
- Chen, X., Dronjic, V. & Helms-Park, R. (Eds). (2015). *Reading in a second language: Cognitive and psycholinguistic issues*. New York: Routledge Publishers.
- Cheng, C. C. (1998). Quantification for understanding language cognition. In B. Tsou, B. Y. Lai, W. G. Chan & S. Y. Wang (Eds), *Quantitative and*

- Computational Studies on the Chinese Language* (pp. 15-30). Hong Kong: City University of Hong Kong.
- Cheng, C. C. (2005). Computing the degree of difficulty in lexical semantics and sentence reading [词汇语意与句子阅读难易度计量]. In *The 6th Chinese Lexical Semantics Workshop* (pp. 261-265). Fujian: Xiamen University.
- Cheng, L. Y., Watanabe, Y. & Curtis, A. (Eds). (2004). *Washback in language testing: Research contexts and methods*. NJ: Lawrence Erlbaum Associates.
- Cheong, Y. Y. (2017). Assessment in Singapore: Perspectives for classroom practice. *Assessment in Education: Principles, Policy & Practice*, DOI:10.1080/0969594X.2017.1309354.
- Chew, C. H. (2007). *Singapore's Chinese language education: A global perspective* [全球化环境下的华语文与华语文教育]. Singapore: Youth Book Company.
- Chew, E. C. T. & Lee, E. (Eds). (1991). *A history of Singapore*. Singapore: Oxford University Press.
- Chia, Y. T. (2015). *Education, culture and the Singapore developmental state: "World-soul" lost and regained?* New York: Palgrave Macmillan.
- Chin, C. K. (2011). *Chinese language curriculum and pedagogies of Singapore*. Nanjing: Nanjing University Press.
- Chin, C. K. (2016). The future: New directions of Singapore Chinese language teaching. In K. C. Soh (Ed.), *Teaching Chinese language in Singapore: Retrospect and challenges* (pp. 27-42). Singapore: Springer.
- Chinese Language Curriculum and Pedagogy Review Committee (CLCPRC). (2004). *Report of the Chinese Language Curriculum and Pedagogy Review Committee*. Singapore: Ministry of Education.
- Chomsky, N. (2002). *Syntactic structures*. Berlin: Mouton de Gruyter.
- Chua, B. H. (2003). Multiculturalism in Singapore: An instrument of social control. *Race and Class*, 44(3), 58-77.
- Chua, S. C. (1964). *Report on the census of population 1957*. Singapore: State of Singapore Department of Statistics.
- Clariana, R. B., Wolfe, M. B. & Kim, K. (2014). The influence of narrative and expository lesson text structures on knowledge structures: Alternate measures of knowledge structure. *Educational Technology Research and Development*, 62(4), 601-616.
- Clarke, M. A. (1980). The short circuit hypothesis of ESL reading—or when language competence interferes with reading performance. *The Modern Language Journal*, 64(2), 203-209.

- Clymer, T. (1968). What is reading?: Some current concepts. In H. M. Robinson (Ed.), *Innovation and change in reading instruction* (pp.7-29). Chicago: University of Chicago Press.
- Cohen, L., Manion, L. & Morrison, K. (2011). *Research methods in education*. London: Routledge.
- Coiro, J. (2003). Exploring literacy on the internet: Reading comprehension on the internet: Expanding our understanding of reading comprehension to encompass new literacies. *The Reading Teacher*, 56, 458-464.
- Coiro, J., Knobel, M., Lankshear, C. & Leu, D. J. (Eds). (2008). *Handbook of research on new literacies*. Mahwah, NJ: Lawrence Erlbaum.
- Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. Retrieved March 7, 2018 from http://english.hanban.org/node_8002.htm
- Covington, M. A., He, C., Brown, C., Naci, L. & Brown, J. (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale: CASPR research report 2006-01*. Athens, GA: The University of Georgia Artificial Intelligence Centre.
- Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage Publications.
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage Publications.
- Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications.
- Creswell, J. W., Plano Clark, V. L., Gutmann, M. & Hanson, W. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds), *Handbook of mixed methods in social and behavioural research* (pp. 209-240). Thousand Oaks, CA: Sage Publications.
- Crisp, V. & Shaw, S. (2012). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies*, 38(2), 209-222.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1980a). Selection theory for a political world. *Public Personnel Management*, 9(1), 37-50.
- Cronbach, L. J. (1980b). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade. Proceedings of the 1979 Educational Testing Service Invitational Conference* (pp. 99-108). San Francisco, CA: Jossey-Bass.

- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp. 3-17). Hillsdale: Lawrence Erlbaum Associates.
- Cronbach, L. J. (1989). Construct validation after thirty years. In Linn, R. (Ed.), *Intelligence: measurement, theory, and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Crooks, T. J., Kane, M. T. & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, 3(3), 265-285.
- Crossley, S. A., Allen, D. B. & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23, 84-102.
- Crossley, S. A., Greenfield, J. & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42, 475-493.
- Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. London: Sage Publications.
- Crystal, D. (2012). *English as a global language*. New York: Cambridge University Press.
- Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience*. New York: Harper Perennial.
- Cumming, J. & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in education: Principles, policy & practice*, 6(2), 177-194.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of educational research*, 49(2), 222-251.
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon: Multilingual Matters.
- Curdt-Christiansen, X. L. (2014). Planning for development or decline? Education policy for Chinese language in Singapore. *Critical Inquiry in Language Studies*, 11(1), 1-26.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.
- Curriculum Planning and Development Division (CPDD). (2002). *The Secondary Chinese Language Syllabus 2002* [中学华文课程标准]. Singapore: Ministry of Education.

- Curriculum Planning and Development Division (CPDD). (2010). *The English Language Syllabus 2010*. Singapore: Ministry of Education.
- Curriculum Planning and Development Division (CPDD). (2011). *The Secondary Chinese Language Syllabus 2011* [中学华文课程标准]. Singapore: Ministry of Education.
- Curriculum Planning and Development Division (CPDD). (2014). *MOE Secondary Chinese Assessment Guide for Educators* [中学华文评价指引]. Singapore: Ministry of Education.
- Dale, E. & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 17(2), 37-54.
- Dale, E. & Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1), 19-26.
- Davie, S. (2017, August 20). Singapore students suffer from high levels of anxiety: Study. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/singapore/education/spore-students-suffer-from-high-levels-of-anxiety-study>
- Davie, S. (2018, October 11). Fewer exams, more time for joy in learning? Please don't stop reducing exams. *The Straits Times*. Retrieved October 27, 2018 from <https://www.straitstimes.com/opinion/fewer-exams-more-time-for-joy-in-learning>
- Davies, A. (1997). Demands of being professional in language testing. *Language testing*, 14(3), 328-339.
- Davies, A. & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language learning* (pp. 795 -814). Mahwah: Lawrence Erlbaum Associates.
- Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 3, 499-545.
- De Vaus, D. A. (2001). *Research design in social research*. London: Sage Publications.
- Dechant, E. (1991). *Understanding and teaching reading: An interactive model*. Hillsdale: Lawrence Erlbaum Associates.
- Dewey, J. (1925). *Experience and nature*. Whitefish, MT: Kessinger Publishing.
- Diamantidaki, F., Pan, L. & Carruthers, K. (Eds). (2008). *Mandarin Chinese teacher Education: Issues and solutions*. London: UCL IOE Press.
- Diggins, J. P. (1994). *The promise of pragmatism: Modernism and the crisis of knowledge and authority*. Chicago: Chicago University Press.

- Dixon, L. Q. (2005). Bilingual education policy in Singapore: An analysis of its sociohistorical roots and current academic outcomes. *International Journal of Bilingual Education and Bilingualism*, 8(1), 25-47.
- Douglas, D. (1997). Language for specific purposes testing. In C. Clapham & D. Carson (Eds), *Encyclopaedia of language in education: Language testing and assessment* (pp. 112-120). Dordrecht: Kluwer Academic.
- Dowling, P. & Brown, A. (2010). *Doing research/ reading research: Re-interrogating education*. New York: Routledge.
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837.
- DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.
- Duke, N. K. (2005). Comprehension of what for what: Comprehension as a non-unitary construct. In S. G. Paris & S. A. Stahl (Eds), *Children's reading comprehension and assessment* (pp. 93-104). Mahwah, NJ: Erlbaum.
- Dunnette, M. D. (1992). It was nice to be there: Construct validity then and now. *Human Performance*, 5(1), 157-169.
- Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, 16(10), 640-647.
- Educational Testing Service. (2014). *The Educational Testing Service standards for quality and fairness*. Retrieved March 7, 2018 from <https://www.ets.org/s/about/pdf/standards.pdf>
- Engeström, Y. (2008). *From teams to knots: Activity-theoretical studies of collaboration and learning at work*. Cambridge: Cambridge University Press.
- English Language Institute of Singapore. (2014). Frameworks for disciplinary literacy. *ELIS Research Digest*, 1(6), 72-86. Retrieved March 7, 2018 from <http://fliphtml5.com/vscu/onxv>
- Fabos, B. (2008). The price of information: Critical literacy, education and today's Internet. In D. J. Leu, J. Coiro, M. Knobel & C. Lankshear (Eds), *Handbook of research on new literacies* (pp. 839- 870). Mahwah, NJ: Erlbaum.
- Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research*, 4(1), 6-16.
- Field, J. (2004). *Psycholinguistics: The key concepts*. London: Routledge.
- Fielding, N. & Fielding, J. (1986). *Linking data*. Beverly Hills, CA: Sage Publications.

- Fielding, N. & Thomas, H. (2008). Qualitative interviewing. In N. Gilbert (Ed.), *Researching social life* (pp. 123-144). London: Sage Publications.
- Flesch, R. (1943). *Marks of readable style: A study in adult education*. New York: Teachers College, Columbia University.
- Flesch, R. (1951). *The art of clear thinking*. New York: Collier.
- Flesch, R. (1955). *Why Johnny can't read?* New York: Harper and Row.
- Fletcher, C. R. (1994). Levels of representation in memory for discourse. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 589-607). San Diego: Academic Press.
- Foucault, M. (1980). *Power/knowledge: Selected interviews and other writings 1972-1977*. New York: Pantheon Books.
- Foucault, M. (1991). *The Foucault reader*. New York: Penguin.
- Foucault, M. (2012). *Discipline and punish: The birth of the prison*. New York: Vintage Books.
- Fraser, C. A. (2007). Reading rate in L1 Mandarin Chinese and L2 English across five reading tasks. *The Modern Language Journal*, 91(3), 372-394.
- Freebody, P. (2003). *Qualitative research in education: Interaction and practice*. London: Sage Publications.
- Friedman, T. L. (2007). *The world is flat: A brief history of the twenty-first century*. New York: Picador.
- Fry, E. (1977). Fry's readability graph: Clarifications, validity and extension to level 17. *Journal of reading*, 21(3), 242-252.
- Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
- Fung, A., Graham, M. & Weil, D. (2007). *Full disclosure: The promise and perils of transparency*. Cambridge: Cambridge University Press.
- Gathercole, V. C. M. (Ed.). (2013). *Issues in the assessment of bilinguals*. Bristol: Multilingual Matters.
- Geranpayeh, A. & Taylor, L. (2013). *Examining listening: Research and practice in assessing second language listening*. Cambridge: Cambridge University Press.
- Goh, C. B. & Gopinathan, S. (2008). The development of education in Singapore since 1965. In S. K. Lee, C. B. Goh, B. Fredriksen & J. P. Tan (Eds), *Toward a better future: Education and training for economic development in Singapore since 1965* (pp. 12-38). Washington: The World Bank.

- Goh, C. T. (1999, May 5). Speech by Prime Minister Goh Chok Tong on Singapore 21 debate in parliament. *National Archives of Singapore Online Archives*. Retrieved March 7, 2018 from <http://www.nas.gov.sg/archivesonline/data/pdfdoc/1999050503/gct19990505d.pdf>
- Goh, H. H., Lin, J. Z. & Zhao, C. S. (2013). *The frequency dictionary of daily Chinese words encountered by Singapore students*. Singapore: NTU-SCCL Press.
- Goh, K. S. (1979). *Report on the Ministry of Education 1978*. Singapore: Ministry of Education.
- Goh, Y. S. (2010). *The globalization of Chinese: A Singapore perspective*. Beijing: The Commercial Press.
- Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly*, 5, 9-30.
- Goodman, K. S. (1985). Unity in reading. In H. Singer & R. Ruddell (Eds), *Theoretical models and process of reading* (pp. 813-840). Newark: International Reading Association.
- Gopinathan, S. (1974). *Towards a national system of education in Singapore 1945-1973*. Singapore: Oxford University Press.
- Gopinathan, S. (1997). Education and development in Singapore. In J. Tan, S. Gopinathan & W. K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 33-53). Singapore: Prentice Hall.
- Gopinathan, S. (2003). Language policy changes 1979-1997: Politics and pedagogy. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 19-44). Singapore: Eastern Universities Press.
- Gough, P. B. (1972). One second of reading. In J. F. Kavanagh & I. G. Mattingly (Eds), *Language by ear and by eye: The relationships between speech and reading* (pp. 331-358). Cambridge: MIT Press.
- Gough, P. B. (1984). Word recognition. In P. D. Pearson, R. Barr, M. L. Kamil & P. Mosenthal (Eds), *Handbook of reading research volume I* (pp. 225-254). New York: Longman.
- Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly*, 25(3), 375-406.
- Grabe, W. (2000). Reading research and its implications for reading assessment. In A. Kunnan (Ed.), *Fairness and validation in language assessment*, (pp. 226-262). Cambridge: Cambridge University Press.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

- Grabe, W. & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow: Longman.
- Green, A. B. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge: Cambridge University Press.
- Greene, J. C. & Caracelli, V. (1997). Defining and describing the paradigm issue in mixed method evaluation. *New Directions for Evaluation*, 74, 5-17.
- Grellet, F. (1987). *Developing reading skills*. Cambridge: Cambridge University Press.
- Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, 31, 111-133.
- Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal*, 29(2), 75-91.
- Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Thousand Oaks, CA: Sage Publications.
- Guest, G., Namey, E. E. & Mitchell, M. L. (2013). *Collecting qualitative data: A field manual for applied research*. Los Angeles: Sage Publications.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427-439.
- Guion, R. M. (1980). On trinitarian doctrines on validity. *Professional Psychology*, 11, 385-398.
- Guion, R. M. (1998). *Assessment, measurement and prediction for personnel decisions*. Hillsdale: Lawrence Erlbaum Associates.
- Gulikers, J. T., Bastiaens, T. J. & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, 52(3), 67-86.
- Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5(10), 511-517.
- Guo, X. (2011). Chinese language teaching in Singapore: Variations and objectives [华文教学在新加坡]. *Journal of Chinese Language Education* [华文学刊], 9(1), 1-16.
- Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives*, 11(1-2), 1-18.
- Haladyna, T. M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53(4), 999-1010.

- Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.
- Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Hambleton, R. K. & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of educational measurement*, 14(2), 75-96.
- Hamers, J. F. & Blanc, M. H. A. (2004). *Bilinguality and bilingualism*. Cambridge: Cambridge University Press.
- Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System*, 28(4), 579-591.
- Hargis, G., Carey, M., Hernandez, A. K., Hughes, P., Longo, D., Rouiller, S. & Wilde, E. (2004). *Developing quality technical information: A handbook for writers and editors*. Upper Saddle River, N.J.: Prentice Hall Professional Technical Reference.
- Hargreaves, E. (2007). The validity of collaborative assessment for learning. *Assessment in Education: Principles, Policy & Practice*, 14 (2), 185-199.
- Hargreaves, E. (2013). Assessment for Learning and Teacher Learning Communities: UK teachers' experiences. *Teaching Education*, 24 (3), 327-344.
- Harris, T. L. & Hodges, R. E. (Eds). (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark: International Reading Association.
- Hawkey, R. (2006). *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.
- Heaton, J. B. (1988). *Writing English language tests*. New York: Longman.
- Hedgcock, J. S. & Ferris, D. R. (2009). *Teaching readers of English: Students, texts, and contexts*. New York: Routledge.
- Hegel, G. W. F. (1985). *Introduction to the lectures on the history of philosophy*. Oxford : Clarendon Press.
- Heidegger, M. (2004). *What is called thinking?* New York: Harper and Row.
- Heng, S. K. (2011, February 8). Opening address by Mr Heng Swee Keat, Minister for Education, at the Ministry of Education Work Plan Seminar 2011. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/news/speeches/opening-address-by-mr-heng-swee-keat-minister-for-education--at-the-ministry-of-education-moe-work-plan-seminar--on-thursday--22-september-2011-at-1000-am-at-ngee-ann-polytechnic-convention-centre>

- Heng, S. K. (2012, February 8). Prepared remarks for Mr Heng Swee Keat, Minister for Education, on 'education for competitiveness and growth' at the Singapore Conference in Washington D.C., USA. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/speech-by-mr-heng-swee-keat-at-the-singapore-conference-washington-dc-usa.pdf>
- Heng, S. K. (2014, April 9). Opening address by Mr Heng Swee Keat, Minister for Education, at the International Conference of Teaching and Learning with Technology at the Suntec International Convention and Exhibition Centre. *MOE Speeches*. Retrieved March 7, 2018 from <http://www.aps.sg/files/in-the-news/opening-address-by-mr-heng-swee-keat-at-the-international-conference-of-teaching-and-learning-with-technology.pdf>
- Heng, S. K. (2015, September 22). Keynote address by Mr Heng Swee Keat, Minister for Education, at the Ministry of Education Work Plan Seminar 2015. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/news/speeches/keynote-address-by-mr-heng-swee-keat-minister-for-education--at-the-ministry-of-education-work-plan-seminar-2015--on-tuesday--22-september-2015-at-9-15am-at-ngee-ann-polytechnic-convention-centre>
- Henning, A. S. (2014). *An argument-based validation of the Teacher Performance Assessment in Washington state*. (Unpublished doctoral thesis). Durham University, Durham.
- Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newbury House.
- Herber, H. (1978). *Teaching reading in content areas*. Englewood Cliffs: Prentice-Hall.
- Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. New York: Guilford Press.
- Ho, C. L. & Alsagoff, L. (1998). English as the common language in multicultural Singapore. In J. A. Foley (Ed.), *English in new cultural contexts: Reflections from Singapore* (pp. 201-217). Singapore: Oxford University Press.
- Ho, K. L. (2014, May 26). PSLE minefields to avoid. *The Straits Times*. Retrieved March 7, 2018 from <http://www.asiaone.com/news/edvantage/psle-minefields-avoid>
- Ho, K. P. (2016). *The ocean in a drop: Singapore: The next fifty years*. Singapore: World Scientific Publishing.
- Holstein, J. A. & Gubrium, J. F. (1995). *The active interview*. London: Sage Publications.

- Hong Kong Examinations and Assessment Authority. (2013). *School-based assessment: Overview* [校本评核—简介]. Retrieved March 7, 2018 from <http://www.hkeaa.edu.hk/tc/sba/introduction/>
- House, E. R. (1977). *The logic of evaluative argument*. Los Angeles: Center for the Study of Evaluation, University of California Los Angeles.
- Huang, H. L. (Producer). (2018, September 30). *On air with Minister: Episode 1 Mr Ong Ye Kung* [空中访民情第1集：教育部长王乙康] [Television broadcast]. Singapore: Mediacorp.
- Hudson, T. (1998). Theoretical perspectives on reading. *Annual Review of Applied Linguistics*, 18, 43-60.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- International Reading Association. (2002). *Integrating literacy and technology in the curriculum: A position statement*. Newark, DE: International Reading Association.
- Internet World Stats. (2016). *Usage and population statistics*. Retrieved March 7, 2018 from <http://www.internetworldstats.com/stats.htm>
- Irwin, J. W. (1986). *Teaching reading comprehension process*. Englewood Cliffs: Prentice-Hall.
- Isaacs, T., Zara, C. & Herbert, G. (2013). *Key concepts in educational assessment*. London: Sage Publications.
- James, J. (2003). Linguistic realities and pedagogical practices in Singapore: Another perspective. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 99-116). Singapore: Eastern Universities Press.
- James, W. (1909). *The meaning of truth: A sequel to 'pragmatism'*. New York: Longmans, Green and Company.
- Jing, X. X. (1995). Assessing the readability of Chinese language instructional materials in China: Formulating a Chinese readability index [中文国文教材的适读性研究—适读年级值的推估]. *Educational Research and Information* [教育研究资讯], 3(3), 113-127.
- Johnson, R. B. & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14-26.

- Johnson, R. B., Onwuegbuzie, A. J., de Waal, C., Stefurak, T. & Hildebrand, D. (2016). Unpacking pragmatism for mixed methods research. In D. Wyse, N. Selwyn, E. Smith & L. E. Suter (Eds), *The BERA/SAGE handbook of educational research* (pp. 259-279). London: Sage Publications.
- Johnston, P. H. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly*, 19(2), 219-239.
- Josey, A. (2012). *Lee Kuan Yew: The crucial years*. Singapore: Marshall Cavendish.
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds), *English in the world: Teaching and learning the language and literature* (pp. 11-30). Cambridge: Cambridge University Press.
- Kachru, B. B. (1992). *The other tongue: English across cultures*. Urbana: University of Illinois Press.
- Kamil, M. L., Pearson, P. D., Moje, E. & Afflerbach, P. (Eds). (2011). *Handbook of reading research volume IV*. London: Routledge.
- Kane, M. T. (1990). *An argument-based approach to validation*. Iowa: ACT Research Report Series.
- Kane, M. T. (1992a). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. T. (1992b). The assessment of professional competence. *Evaluation and the Health Professions*, 15(2), 163-182.
- Kane, M. T. (1992c). *Viewpoints: The validity of assessments of professional competence*. Retrieved March 7, 2018 from <http://files.eric.ed.gov/fulltext/ED343958.pdf>
- Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions*, 17(2), 133-159.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135-170.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Washington, DC: American Council on Education/Praeger.
- Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 39-64). Charlotte, NC: Information Age Publishing.

- Karantonis, A. & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4-12.
- Kaur, K. (2018, September 30). Want children to discover joy of learning? Overhaul PSLE. *The Straits Times*. Retrieved October 27, 2018 from <https://www.straitstimes.com/singapore/education/want-children-to-discover-joy-of-learning-overhaul-psle>
- Keenan, J. M., Betjemann, R. S. & Olson, R. K. (2008). Reading comprehension tests vary in the skills they access: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, 12(3), 281-300.
- Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book Company.
- Kelly, A. V. (2009). *The curriculum: Theory and practice*. London: Sage Publications.
- Khalifa, H. & Salamoura, A. (2011). Criterion-related validity. In L. Taylor (Ed.), *Examining speaking* (pp. 259-292). Cambridge: Cambridge University Press.
- Khalifa, H. & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
- Khamid, H. M. A. (2015, August 3). Public debate to advance with civil society's participation: DPM Tharman. *Channel NewsAsia*. Retrieved March 7, 2018 from <http://www.channelnewsasia.com/news/singapore/public-debate-to-advance/2025098.html>
- Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale: Lawrence Erlbaum Associates.
- Kintsch, W. (1998). *Comprehension: A framework for cognition*. New York: Cambridge University Press.
- Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. Ruddell & N. Unrau (Eds), *Theoretical models and processes of reading* (pp. 1270-1328). Newark: International Reading Association.
- Kintsch, W. & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, 74, 828- 834.
- Kirkpatrick, A. (2010). *English as a lingua franca in ASEAN*. Hongkong: Hongkong University Press.
- Klein, S. B. (2015). *Learning: Principles and applications*. Los Angeles: Sage Publications.

- Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, 57, 1-44.
- Koh, G. (2011, April 28). Singapore General Election 2011: What moves the voters? *Institute of Policy Speeches*. Retrieved March 7, 2018 from http://lkyspp2.nus.edu.sg/ips/wp-content/uploads/sites/2/2013/08/GK_CAB-GE2011_280411.pdf
- Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer Verlag.
- Kong, D. (2017, September 27). Boost Chinese language standards in schools. *The Straits Times*. Retrieved March 7, 2018 from <https://www.straitstimes.com/forum/letters-in-print/boost-chinese-language-standards-in-schools>
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge: Harvard University Press.
- Kramsch, C. (1993). *Context and culture in language teaching*. Oxford: Oxford University Press.
- Krashen, S. (1983). Newmark's "Ignorance Hypothesis" and current second language acquisition theory. In S. Gass & L. Selinker (Eds), *Language transfer in language learning*. Rowley: Newbury House.
- Krashen, S. (2004). *The power of reading*. Portsmouth: Heinemann.
- Kuah, A. W. J. (2018, October 1). Why move to reduce examinations and emphasis on grades is disconcerting, but necessary. *Today*. Retrieved October 27, 2018 from <https://www.todayonline.com/commentary/why-move-to-reduce-examinations-is-disconcerting-necessary>
- Kucer, S. B. (2001). *Dimensions of literacy: A conceptual base of teaching reading and writing in school settings*. Mahwah: Lawrence Erlbaum Associates.
- Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural equation modelling approach*. Cambridge: Cambridge University Press.
- Kunnan, A. J. (1998). Approach to validation in language assessment. In A.J. Kunnan (Ed.), *Validation in language assessment* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds), *European language testing in a global context* (pp. 27-48). Cambridge: Cambridge University Press.
- Kunnan, A. J. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. J. Weir (Eds), *Multilingualism and*

- assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229-251). Cambridge: Cambridge University Press.
- Kunnan, A. J. & Zhang, L. M. (2015). Responsibility in language assessment. In H. Yang (Ed.), *The sociology of language testing* (pp. 211-231). Shanghai: Shanghai Foreign Language Press.
- Kuo, C. Y. & Jernudd, B. H. (1994). Balancing macro and micro-sociolinguistic perspectives in language management: The case of Singapore. In T. Kandiah & J. Kwan-Terry (Eds), *English and language planning: A south-east asian contribution* (pp. 70-89). Singapore: Times Academic Press.
- Kvale, S. & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing*. Thousand Oaks: Sage Publications.
- Kwok, K. W. (2001). Chinese-educated intellectuals in Singapore: Marginality, memory and modernity. *Asian Journal of Social Science*, 29(3), 495-519.
- LaBerge, D. & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book*. London: Longman.
- Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Deighton, Bell and Company.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds), *Vocabulary and applied linguistics* (pp. 126-132). London: MacMillan.
- Lee, E. (2008). *Singapore: The unexpected nation*. Singapore: Institute of Southeast Asian Studies.
- Lee, H. L. (2013, August 18). Prime Minister Lee Hsien Loong's National Day Rally 2013. *Prime Minister's Office*. Retrieved October 27, 2018 from <https://www.pmo.gov.sg/newsroom/prime-minister-lee-hsien-loongs-national-day-rally-2013-english>
- Lee, K. Y. (1965, December 14). Summary of the speech by the Prime Minister Mr Lee Kuan Yew in Parliament when he moved the motion of thanks to the Yang di-Pertuan Negara for his address. *National Archives of Singapore Online Archives*. Retrieved March 7, 2018 from <http://www.nas.gov.sg/archivesonline/data/pdfdoc/lky19651214a.pdf>
- Lee, K. Y. (1978, March 4). Two speeches (combined & edited) by the Prime Minister, Mr Lee Kuan Yew, at Istana Chap Goh Mei Reception and Tanjong Pagar Community Centre Scholarships Presentation. *National Archives of Singapore Online Archives*. Retrieved f March 7, 2018 from <http://www.nas.gov.sg/archivesonline/data/pdfdoc/lky19780304.pdf>

- Lee, K. Y. (2000). *From third world to first: The Singapore story (1965-2000)*. Singapore: Times Media.
- Lee, K. Y. (2004, November 25). Speech by Minister Mentor Lee Kuan Yew at the parliamentary debate on the report of the Chinese Language Curriculum and Pedagogy Review Committee. *Singapore Government Press Release*. Retrieved March 7, 2018 from <http://www.nas.gov.sg/archivesonline/speeches/view.html?filename=2004112501.htm>
- Lee, K. Y. (2011, September 6). Speech by Mr Lee Kuan Yew, former Minister Mentor and current Senior Advisor to Government of Singapore Investment Corporation at the launch of the English Language Institute of Singapore (ELIS) at the Marina Bay Sands Expo and Convention Centre. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/speech-by-mr-lee-kuan-yew-at-elis-launch.pdf>
- Lee, Y. J. (2010). Nothing but the truth?: Dilemmas and conflicts during research reporting on educational change. In A. D. Henshall & B. C. Fontanez (Eds), *Educational change* (pp.173-188). Hauppauge, NY: Nova Science Publishers.
- Lee, Y. J. (2014). Science education in a straightjacket: The interplay of people, policies, and place in an East-Asian developmental state. In A. L. Tan, C. L. Poon & S. S. L. Lim (Eds), *Inquiry into the Singapore science classroom: Research and practices* (pp.151-171). Dordrecht: Springer.
- Lee, Y. J., Hung, W. L. D. & Cheah, H. M. (2009). IT and educational policy in the Pacific-Asian region. In Voogt, J. & Knezek, G. (Eds), *International handbook of information technology in education* (pp. 1119-1132). Dordrecht: Springer.
- Leong, W. K. (1999). *Language teaching and assessment* [语文教学与测试]. Singapore: SNP Publishing.
- Leong, W. S. Cheng Y. & Tan, K. (Eds). (2014). *Assessment and learning in schools*. Singapore: Pearson.
- Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J. & Henry, L. A. (2013). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In R. B. Ruddell & D. Alvermann (Eds), *Theoretical models and processes of reading* (pp. 1150-1181). Newark, DE: International Reading Association.
- Leu, D. J., McVerry, J. G., O'Byrne, W. I., Zawilinski, L., Castek, J. & Hartman, D. K. (2009). The new literacies of online reading comprehension and the irony of no child left behind: Students who require our assistance the most, actually receive it the least. In L. M. Morrow, R. Rueda & D. Lapp (Eds), *Handbook of research on literacy and diversity* (pp. 173-194). New York: Guilford.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K. & Patz, R. J. (1998, April). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the National Council for Measurement in Education annual meeting, San Diego, California.

- Lewis, D. M., Mitzel, H. C., Green, D. R. & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.
- Lewis-Beck, M. S., Bryman, A. E. & Liao, T. F. (Eds). (2004). *The Sage encyclopaedia of social science research methods*. Thousand Oaks, CA: Sage Publications.
- Liebethal, A., Michelitsch, R. & Tarazona, E. (2005). *Extractive industries and sustainable development: An evaluation of World Bank Group experience*. Washington, DC: World Bank Publications.
- Lim, K. S. (1965, November 19). Speech by the Minister for Finance, Mr Lim Kim San, at the Singapore Manufacturer's Association luncheon at Imperial room. *National Archives of Singapore Online Archives*. Retrieved March 7, 2018 from <http://www.nas.gov.sg/archivesonline/data/pdfdoc/PressR19651119.pdf>
- Lim, L. (2013). Meritocracy, elitism and egalitarianism: A preliminary and provisional assessment of Singapore's primary education review. *Asia Pacific Journal of Education*, 33(1), 1-14.
- Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage Publications.
- Linderholm, T. & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, 94, 778-784.
- Linderholm, T., Virtue, S., Tzeng, Y. & van den Broek, P. (2004). Fluctuations in the availability of information during reading: Capturing cognitive processes using the landscape model. *Discourse Processes*, 37, 165-186.
- Linn, R. L. (2000). Assessments and accountability. *Educational researcher*, 29(2), 4-16.
- Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. USA: Information Age Publishing.
- List, A. & Alexander, P. A. (2017). Analysing and integrating models of multiple text comprehension. *Educational Psychologist*, 52(3), 143-147.
- Liu, Y. L. & Song, S. Z. (1992). Calculating and ranking of Chinese characters and words [论汉语教学字词的统计与分级]. In the Office of Chinese Language Council (Ed.), *The guidelines of HSK vocabulary and characters* [汉语水平考试词汇与汉字等级大纲] (pp. 1-22). Beijing: Beijing Language College Press.
- Lively, B. A. & Pressey, S. L. (1923). A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9, 389-398.

- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635-694.
- Loke, K. K. (1994). Policy intentions and policy outcomes: A comparative perspective on the Singapore bilingual education system. *Compare: A Journal of Comparative and International Education*, 24(1), 53-65.
- Long, P. Y. (2018, October 26). Examinations not the best way to assess a person [考试非评估个人最好办法]. *Lianhe Zaobao* [联合早报], p. 25.
- Lorch, J., Kluzewitz, M. & Lorch, E. (1995). Distinctions among reading situations. In R. Lorch & E. O'Brien (Eds), *Sources of coherence in reading* (pp. 375-398). Hillsdale, NJ: L. Erlbaum.
- Lunzer, E. A. & Gardner, K. (1979). *The effective use of reading*. London: Heinemann.
- Lu, S. X. (1987). *Perspectives on language education* [吕叔湘论语文教学]. Jinan: Shandong Education Publishing House.
- Mackey, W. F. (1987). Bilingualism and multilingualism. In U. Ammon, N. Dittmar & K. Mattheier (Eds), *Sociolinguistics: An international handbook of the science of language and society*, (pp. 699-713). Berlin: Walter de Gruyter.
- MacLeod, R. (1982). *Days of judgement: Science, examinations and the organization of knowledge in late Victorian England*. Driffield: Nafferton Books.
- Magliano, J. P., Millis, K., Ozuru, Y. & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), *Reading comprehension strategies* (pp. 107-136). New York: Erlbaum.
- Mahbubani, K. (2015, June 13). Trust the people, share government data. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/opinion/trust-the-people-share-government-data>
- Marshall, B. (2015). Learning, pedagogy and assessment. In D. Scott & E. Hargreaves (Eds), *The Sage handbook of learning* (pp. 254-262). London: Sage Publications.
- Mauzy, D. K. & Milne, R. S. (2002). *Singapore politics under the People's Action Party*. London: Routledge.
- Maxwell, J. A. (1996). *Qualitative research design: An interactive approach*. Thousand Oaks, CA: Sage Publications.
- McCarthy, T. & Ellis, E. (1999, July 19). Singapore lightens up. *Time*. Retrieved March 7, 2018 from <http://content.time.com/time/world/article/0,8599,2054247,00.html>

- McCormick, T. (1988). *Theories of reading in dialogue: An interdisciplinary study*. New York: University Press of America.
- McNamara, D. S., Ozuru, Y., Graesser, A. C. & Louwrese, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 573-578). Mahwah, NJ: Erlbaum.
- McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.
- McNamara, T. F. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3(1), 31-51.
- McNamara, T. P., Miller, D. L. & Bransford, J. D. (1991). Mental models and reading comprehension. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds), *Handbook of reading research volume II* (pp. 490-511). New York: Longman.
- Meara, P. & Milton, J. (2003). *X_Lex, the Swansea Levels Test*. Newbury: Express.
- Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Thousand Oaks, CA: Sage Publications.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp. 33-45). Hillsdale: Lawrence Erlbaum Associates.
- Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.
- Messick, S. (1989b). Validity. In R. Linn (Ed.), *Educational measurement* (pp.13-103). New York: Macmillan.
- Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (pp. 1487-1495). New York: Macmillan Publishing Company.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.
- Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 241-256.
- Michael, E. J. (2006). *Public policy: The competitive framework*. Melbourne: Oxford University Press.
- Miller, W. L. & Crabtree, B. J. (1999). The dance of interpretation. In B. J. Crabtree & W. L. Miller (Eds), *Doing qualitative research* (pp. 127-143). London: Sage Publications.
- Ministry of Education (MOE). (1971). *Examinations division annual report 1971*. Singapore: MOE.
- Ministry of Education (MOE). (2004a, July 23). Chinese language 'B' syllabus for students with exceptional difficulties in learning Chinese language, bonus points scheme for students strong in mother tongue languages for admission to selected courses in NUS and NTU, extension of bonus points scheme for students eligible to apply to SAP schools. *MOE Press Releases*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/chinese-language-39-b-39-syllabus-for-students-press-release.pdf>
- Ministry of Education (MOE). (2004b, March 31). Formation of Singapore Examinations and Assessment Board. *MOE Press Releases*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/formation-of-singapore-examinations-amp-assessment-board.pdf>
- Ministry of Education (MOE). (2010a). *Building a national education system for the 21st century: The Singapore experience*. Retrieved March 7, 2018 from http://www.edu.gov.on.ca/bb4e/Singapore_CaseStudy2010.pdf
- Ministry of Education (MOE). (2010b, March 9). MOE to enhance learning of 21st century competencies and strengthen art, music and physical education. *MOE Press Releases*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/moe-to-enhance-learning-of-21s.pdf>
- Ministry of Education (MOE). (2011). *The Mother Tongue Languages Review Committee report: Nurturing active learners and proficient users*. Singapore: Ministry of Education.
- Ministry of Education (MOE). (2012, September 12). MOE removes secondary school banding and revamps school awards. *MOE Press Releases*. Retrieved October 27, 2018 from <https://www.moe.gov.sg/news/press-releases/moe-removes-secondary-school-banding-and-revamps-school-awards>
- Ministry of Education (MOE). (2014a). *Education statistics digest 2014*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/publications/education-statistics-digest/esd-2014.pdf>

- Ministry of Education (MOE). (2014b). Student drop-out rate for primary, secondary and ITE levels. *MOE Parliamentary Replies*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/news/parliamentary-replies/student-drop-out-rate-for-primary-secondary-and-ite-levels>
- Ministry of Education (MOE). (2015a). *About us*. Retrieved March 7, 2018 from <http://www.moe.gov.sg/about/>
- Ministry of Education (MOE). (2015b). *Language programmes*. Retrieved March 7, 2018 from <http://www.moe.gov.sg/education/secondary/language-programmes/>
- Ministry of Education of the People's Republic of China. (2011). *Language syllabus* [语文课程标准]. Beijing: Beijing Normal University Publishing Group.
- Modern Language Association. (2015). *New MLA survey report shows advanced language study grows for several languages despite lower overall language enrolments in US colleges and universities*. Retrieved March 7, 2018 from http://www.mla.org/pdf/2013_enrollment_survey_pr.pdf
- Moore, A. (2015). Knowledge, curriculum and learning: 'What did you learn in school?' In D. Scott & E. Hargreaves (Eds), *The Sage handbook of learning* (pp. 144-154). London: Sage Publications.
- Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research*, 1(1), 48-76.
- Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, 14(2), 5-13.
- Mother Tongue Languages Review Committee (MTLRC). (2011). *The Mother Tongue Languages Review Committee report: Nurturing active learners and proficient users*. Singapore: Ministry of Education.
- Mullis, I. V. S. & Martin, M. O. (Eds). (2015). *PIRLS 2016 framework*. Retrieved March 7, 2018 from http://timss.bc.edu/pirls2016/downloads/P16_Framework_2ndEd.pdf
- Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge University Press: Cambridge.
- National Institute of Child Health and Human Development. (2001). *Put reading first: The research building blocks for teaching children to read*. Washington, DC: U.S. Government Printing Office.
- National Library Board. (2017). *2016 national reading habits study: Findings on teenagers*. Singapore: National Library Board.

- Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125-173.
- Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, 14(2), 149-170.
- Newton, P. E. (2008). *Monitoring national attainment standards: OFQUAL report*. Retrieved March 7, 2018 from http://dera.ioe.ac.uk/8639/1/083916_monitoring_national_attainment_standards.pdf
- Newton, P. E. (2012a). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10(1-2), 1-29.
- Newton, P. E. (2012b). Questioning the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10(1-2), 110-122.
- Newton, P. E. (2017a). *An approach to understanding validation arguments: OFQUAL report*. Retrieved March 7, 2018 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/653070/An_approach_to_understanding_validation_arguments.pdf
- Newton, P. E. (2017b). Assessment dilemmas. *Research Intelligence*, 133, 18-20.
- Newton, P. E. (2017c). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, 36(2), 5-15.
- Newton, P. E. & Baird, J. A. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 173-177.
- Newton, P. E., Baird, J. A., Goldstein, H., Patrick, H. & Tymms, P. (Eds). (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Newton, P. E. & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18(3), 301-319.
- Newton, P. E. & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage Publications.
- Ng, C. W. (2018, May 27). Singapore education system is not a free lunch. *Channel NewsAsia*. Retrieved October 27, 2018 from <https://www.channelnewsasia.com/news/commentary/scrap-psle-singapore-education-system-not-a-free-lunch--10277802>
- Ng, P. C. L. (2011). Language planning in action: Singapore's multilingual and bilingual policy. *Ritsumeikan Asia Pacific Journal*, 30, 1-12.
- Ng, P. C. L. (2014). Mother tongue education in Singapore: Concerns, issues and controversies. *Current Issues in Language Planning*, 15(4), 1-15.

- Ng, P. T. (2017). *Learning from Singapore: The power of paradoxes*. Singapore: Taylor and Francis.
- Ng, T. C. (2016). The past: An review of five reviews. In K. C. Soh (Ed.), *Teaching Chinese language in Singapore: Retrospect and challenges* (pp. 3-10). Singapore: Springer.
- Nuttall, C. (1996). *Teaching reading skills in a foreign language*. London: Heinemann.
- Odlin, T. (1989). *Language transfer*. New York: Cambridge University Press.
- Ong, Y. K. (2016, April 21). What SkillsFuture is about. *The Straits Times*. Retrieved March 7, 2018 from <https://www.straitstimes.com/opinion/what-skillsfuture-is-about>
- Ong, Y. K. (2018, September 28). Opening address by Mr Ong Ye Kung, Minister for Education, at the Schools Work Plan Seminar 2018. *MOE Speeches*. Retrieved October 27, 2018 from <https://www.moe.gov.sg/news/speeches/opening-address-by-mr-ong-ye-kung--minister-for-education--at-the-schools-work-plan-seminar>
- Organization for Economic Cooperation and Development (OECD). (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Retrieved March 7, 2018 from <http://www.oecd.org/pisa/46623978.pdf>
- Organization for Economic Cooperation and Development (OECD). (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*. Retrieved March 7, 2018 from <http://www.mecd.gob.es/dctm/inee/internacional/pisa-2015-frameworks.pdf?documentId=0901e72b820fee48>
- Pakir, A. (1997). Education and invisible language planning: The case of the English language in Singapore. In J. Tan, S. Gopinathan & W.K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 57-74). Singapore: Prentice Hall.
- Pan, X. H. (2010, April 27). Chinese language education in Singapore: Holding the last line of defence [华文，不能让底线失守]. *Lianhe Zaobao* [联合早报], p. 20.
- Pang, E. F. & Lim, L. (1997). The school system and social structure in Singapore. In J. Tan, S. Gopinathan & W. K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 363-368). Singapore: Prentice Hall.
- Papert, S. (1980). *Mindstorms*. New York: Basic Books.
- Parsons, T. (1962). *The structure of social action*. New York: Free Press.

- Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice*. Thousand Oaks, CA: Sage Publications.
- Pavlov, I. P. (1897). *The work of the digestive glands*. London: Griffin.
- Pearson, I. (1988). Tests as levers of change. In D. Chamberlain & R. J. Baumgardner (Eds), *ESP in the classroom: Practice and evaluation* (pp. 98-107). London: Modern English.
- Pearson, P. D. (2009). The roots of reading comprehension instruction. In S. E. Israel & G. G. Duffy (Eds), *Handbook of research on reading comprehension* (pp. 3-31). New York: Routledge.
- Pearson, P. D. & Johnson, D. D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart & Winston.
- Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, 12, 286-302.
- Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds), *The neurocognition of language* (pp. 167-208). Oxford: Oxford University Press.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357-383.
- Perfetti, C. A. & Curtis, M. E. (1986). Reading. In R. F. Dillon & R. J. Sternberg (Eds), *Cognition and instruction* (pp. 13-57). San Diego: Academic Press.
- Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher*, 24(7), 5-12.
- Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.
- Piaget, J. (1964). Development and learning. In R. Ripple & V. Rockcastle (Eds), *Piaget rediscovered* (pp. 78-119). Washington, DC: U.S. Office of Education, National Science Foundation.
- Piaget, J. (1969). *Science of education and the psychology of the child*. New York: Viking.
- Piaget, J. (1970). Piaget's theory. In P. Mussen (Ed.), *Carmichael's manual of child psychology* (pp. 703-732). New York: Wiley.
- Pollitt, A., Ahmed, A., Baird, J. A., Tognolini, J. & Davidson, M. (2008). Improving the quality of GCSE Assessment. *The Qualifications and Curriculum Authority*. Retrieved March 7, 2018 from <http://www.camexam.co.uk>
- Popham, W. J. (1987). The merits of measurement driven instruction. *Phi Delta Kappa*, 68, 679-682.

- Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13.
- Popper, K. (2002). *The logic of scientific discovery*. London: Routledge.
- Prime Minister's Office, Singapore Department of Statistics, Ministry of Home Affairs, Immigration and Checkpoints Authority & Ministry of Manpower. (2017). *Population in brief 2017*. Retrieved March 7, 2018 from <https://www.strategygroup.gov.sg/docs/default-source/default-document-library/population-in-brief-2017.pdf>
- Prior, L. (2003). *Using documents in social research*. London: Sage Publications.
- Qi, H. Y. (2012). A discussion on the importance of teaching, assessment and testing from the perspective of the Singapore primary level Chinese language syllabus. *Journal of Chinese Language Education*, 19, 1-14.
- Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, 56, 283-307.
- Rahman, T. & Mislevy, R. J. (2017). *Educational Testing Service report: Integrating cognitive views into psychometric models for reading comprehension assessment*. Retrieved March 7, 2018 from <https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12163>
- Ratnam-Lim, C. T .L. & Tan, K. H. K. (2015). Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assessment in Education: Principles, Policy & Practice*, 22(1), 61-78.
- Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In B. Laufer & P. Bogaards (Eds), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209-227). Amsterdam: John Benjamins.
- Reid, T. (2011). *Essays on the intellectual powers of man*. Cambridge: Cambridge University Press.
- Riches, C. & Genesee, F. (2006). Crosslanguage and crossmodal influences. In F. Genesee, K. Lindholm-Leary, W. Saunders & D. Christian (Eds), *Educating English language learners: A synthesis of research evidence* (pp. 64-108). New York: Cambridge University Press.
- Roach, J. (1971). *Public examinations in England 1850-1900*. Cambridge: Cambridge University Press.
- Roberts, T. A., Christo, C. & Shefelbine, J. A. (2010). Word recognition. In M. L. Kamil, P. D. Pearson, E. B. Moje & P. Mosenthal (Eds), *Handbook of reading research volume IV* (pp. 229-258). New York: Longman.
- Robson, C. (2016). *Real world research: A resource for users of social research methods in applied settings*. Chichester, NJ: John Wiley & Sons.

- Romaine, S. (1995). *Bilingualism*. Malden: Blackwell Publishers.
- Rorty, R. (1999). *Philosophy and social hope*. London: Penguin Books.
- Rose, C. (Producer). (2009, October 23). *Charlie Rose* [Television broadcast]. Boston: Public Broadcasting Service.
- Rosenblatt, L. R. (1978). *The reader, the text, the poem: The transactional theory of the literary work*. Carbondale: Southern Illinois University Press.
- Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance* (pp. 573-603). New York: Academic Press.
- Russell, B. (1997). *Problems of philosophy*. Oxford: Oxford University Press.
- Ryan, R. M. & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Education Psychology*, 25, 54-67.
- Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp.113-129). Hillsdale: Lawrence Erlbaum Associates.
- Sahlgren, G. H. (Ed.). (2014). *Tests worth teaching to: Incentivising quality in qualifications and accountability*. London: The Centre for Market Reform of Education.
- Saville, N. (2003). The process of test development and revision within UCLES EFL. In C.J. Weir & M. Milanovic (Eds), *Continuity and innovation: Revising the Cambridge proficiency in English examination 1913-2002* (pp. 57-120). Cambridge: Cambridge University Press.
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64, 913-951.
- Schunk, D. H. (2012). *Learning theories: An educational perspective*. New Jersey: Prentice Hall.
- Scott, D. (1997). Qualitative approaches to data collection and analysis: Examinations and schools. In G. McKenzie, J. Powell & R. Usher (Eds), *Understanding social research: Perspectives on methodology and practice* (pp. 155-172). London: The Falmer Press.
- Scott, D. (2000). *Reading educational research and policy*. London: RoutledgeFalmer.
- Scott, D. (2010). *Education, epistemology and critical realism*. London: Routledge.

- Scott, D. (2011). Assessment reforms: High stakes testing and knowing the contents of other minds. In R. Berry & B. Adamson (Eds), *Assessment reform in education: Policy and practice* (pp. 155-164). Amsterdam: Springer.
- Scott, D. (2015). *New perspectives on curriculum, learning and assessment*. Dordrecht: Springer.
- Scott, D. (2016). *Education systems and learners: Knowledge and knowers*. London: Macmillan Palgrave.
- Scott, D. & Hargreaves, E. (2015). An introduction and a theory of learning. In D. Scott & E. Hargreaves (Eds), *The Sage handbook of learning* (pp. 1-15). London: Sage Publications.
- Scott, D., Husbands, C., Slee, R., Wilkins, R. & Terano, M. (2015). *Policy transfer and educational change*. London: Sage Publications.
- Scott, D., Posner, C. M., Martin, C. & Guzman, E. (2015). *Interventions in education systems: Reform and development*. London: Bloomsbury Publishing.
- Sharpe, L. & Gopinathan, S. (1997). Effective island, effective schools: Repair and restructuring in the Singapore school system. In J. Tan, S. Gopinathan & W. K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 369-384). Singapore: Prentice Hall.
- Shaw, S. D. & Crisp, V. (2011). Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges. *Research Matters, 11*, 14-19.
- Shaw, S. D. & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters Special Issue, 3*, 3-44.
- Shaw, S. D. & Crisp, V. (2015). Reflections on a framework for validation: Five years on. *Research Matters, 19*, 31-37.
- Shaw, S. D., Crisp, V. & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Policy, Principles & Practice, 19*(2), 159-176.
- Shaw, S. D. & Newton, P. E. (2012, April). Cracks in construct validity theory. *Paper presented at the National Council on Measurement in Education Annual Meeting*, British Columbia: Vancouver.
- Shaw, S. D. & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
- Shen, H. (2005). Linguistic complexity and beginning-level L2 Chinese reading. *Journal of the Chinese language Teachers Association, 40*(3), 1-28.

- Shen, H. & Ke, C. (2007). Radical awareness and word acquisition among non-native learners of Chinese. *The Modern Language Journal*, 91, 97-111.
- Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information*, 22, 63-75.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman.
- Shohamy, E. & Hornberger, N. H. (Eds). (2008). *Encyclopedia of language and education: Language testing and assessment*. New York: Springer.
- Shu, H. & Anderson, R. C. (1999). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. Inhoff & H. C. Chen (Eds), *Reading Chinese script: A cognitive analysis* (pp.1-19). Mahwah: Lawrence Erlbaum Associates.
- Silver, R. E. (2005). The discourse of linguistic capital: Language and economic policy planning in Singapore. *Language Policy*, 4(1), 47-66.
- Sim, J. J. (2014). *The washback effects of the O level English language examination on Singaporean teachers*. (Unpublished master's dissertation). National Institute of Education, Singapore.
- Simpson, J. H. & Wigglesworth, G. (2008). *Children's language and multilingualism: Indigenous language use at home and school*. London: Continuum.
- Simpson, T. L. (2002). Dare I oppose constructivist theory? *The Educational Forum*, 66, 347-354.
- Sin, D. (1998). *Saints, sinners and Singaporeans: A collection of poems*. Singapore: Angsana Books.
- Sin, Y. & Ng, W. M. (2018, February 14). The Chinese Singaporean identity: A complex, ever changing relationship. *The Straits Times*. Retrieved March 7, 2018 from <https://www.straitstimes.com/opinion/a-complex-ever-changing-relationship>
- Singapore Examinations and Assessment Board (SEAB). (2013). *SEAB annual report 2012/2013*. Retrieved March 7, 2018 from https://www.seab.gov.sg/pages/media/Publications/annualReport/annualReport_12_13/index.html
- Singapore Examinations and Assessment Board (SEAB). (2014a). *GCE 1162 examination information booklet* [华文1162 试卷说明]. Singapore: Ministry of Education.
- Singapore Examinations and Assessment Board (SEAB). (2014b). Robust processes in setting examination questions and marking papers. *SEAB Media Replies*.

- Retrieved March 7, 2018 from
<http://www.seab.gov.sg/publicCommunications/mediaReplies/SEABForumLetterReply-RobustProcessesinSettingExaminationQuestionsandMarkingPapers.pdf>
- Singapore Examinations and Assessment Board (SEAB). (2014c). *SEAB annual report 2013/2014*. Retrieved March 7, 2018 from
https://www.seab.gov.sg/pages/media/Publications/annualReport/annualReport_13_14/pdf/AnnualReport2014.pdf
- Singapore Examinations and Assessment Board (SEAB). (2015a). *National examinations: General information*. Retrieved March 7, 2018 from
<https://www.seab.gov.sg/pages/nationalExaminations/GOL/general.asp>
- Singapore Examinations and Assessment Board (SEAB). (2015b). *National examinations: Syllabuses*. Retrieved March 7, 2015 from
https://www.seab.gov.sg/pages/nationalExaminations/GOL/School_Candidates/2015_GCE_O.asp
- Singapore Examinations and Assessment Board (SEAB). (2015c). *SEAB annual report 2014/2015*. Retrieved March 7, 2018 from
https://www.seab.gov.sg/pages/media/Publications/annualReport/annualReport_14_15/pdf/AnnualReport2015.pdf
- Singapore Examinations and Assessment Board (SEAB). (2017a). *Our vision, mission and values*. Retrieved March 7, 2018 from
<http://www.seab.gov.sg/pages/about/vision.asp>
- Singapore Examinations and Assessment Board (SEAB). (2017b). *Publications*. Retrieved March 7, 2018 from <https://www.seab.gov.sg/pages/media/publications>
- Singapore Examinations and Assessment Board (SEAB). (2017c). *SEAB annual report 2016/2017*. Retrieved March 7, 2018 from
https://www.seab.gov.sg/pages/media/Publications/annualReport/annualReport_16_17/pdf/AnnualReport2017.pdf
- Singh, B. (2015). *Quest for political power: Communist subversion and militancy in Singapore*. Singapore: Marshall Cavendish.
- Siok, W. T. & Fletcher, P. (2001). The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. *Developmental Psychology*, 37, 886-899.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.
- Sismondo, S. (1993). Some social constructions. *Social Studies of Science*, 23, 515-553.

- Skinner, B. F. (1974). *About behaviorism*. New York: Knopf.
- Slife, B. D. & Williams, R. N. (1995). *What's behind the research: Discovering hidden assumptions in the behavioural sciences*. Thousand Oaks, CA: Sage Publications.
- Smith, C. S. (2003). *Modes of discourse: The local structure of texts*. Cambridge: Cambridge University Press.
- Smith, F. (2004). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Mahwah: Lawrence Erlbaum Associates.
- Smith, J. K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. *Educational Researcher*, 12(3): 6-13.
- Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Retrieved March 7, 2018 from http://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1465.pdf
- Spolsky, B. (2000). Language testing in *The Modern Language Journal*. *The Modern Language Journal*, 84, 536-552.
- Spolsky, B. (2004). *Language policy*. Cambridge: Cambridge University Press.
- Srivastava, P. & Hopwood, N. (2009). A practical iterative framework for qualitative data analysis. *International journal of qualitative methods*, 8(1), 76-84.
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16, 32-71.
- Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds), *Handbook of reading research volume II* (pp. 418-452). New York: Longman.
- Statista (2017). *The most spoken languages worldwide*. Retrieved March 7, 2018 from <https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/>
- Stewart, V. & Wang, S. H. (2005). *Expanding Chinese language capacity in the United States*. New York: Asia Society.
- Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.
- Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161-179.
- Stobart, G. (2012). Foreword. *Research Matters Special Issue*, 3, 1.

- Stromso, H. I. & Braten, I. (2002). Norwegian law students' use of multiple sources while reading expository texts. *Reading Research Quarterly*, 37, 208-227.
- Sun, H. Y. (1992). *Chinese readability formulas* [中文易懂性公式]. (Unpublished master's dissertation). Beijing normal university, Beijing.
- Sung, Y. T., Lin, W. C., Dyson, S. B, Chang, K. E. & Chen, Y. C. (2015). Levelling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2), 371-391.
- Sweet, A. & Snow, C. E. (Eds). (2003). *Rethinking reading comprehension*. New York: The Guilford Press.
- Tan, C. B. (Ed.). (2013). *Routledge handbook of the Chinese diaspora*. New York: Routledge.
- Tan, C. L. (Ed.). (2011). *From practice to practical: Teaching and learning of Chinese as a second language*. Nanjing: Nanjing University Press.
- Tan, C. L. (2016). The present: An overview of teaching Chinese language in Singapore. In K. C. Soh (Ed.), *Teaching Chinese language in Singapore: Retrospect and challenges* (pp. 11-26). Singapore: Springer.
- Tan, D. W. L. (2018, October 3). Removing exams not the way to go. *The Straits Times*. Retrieved October 27, 2018 from <https://www.straitstimes.com/forum/letters-in-print/removing-exams-not-the-way-to-go>
- Tan, G. P. A. (2004). *A sociolinguistic analysis of the 'bilingual approach to the teaching of Chinese language' in Singapore*. Retrieved March 7, 2018 from <https://scholarbank.nus.edu.sg/bitstream/10635/14560/1/TanGPA.pdf>
- Tan, J. (Ed.). (2012). *Education in Singapore: Taking stock, looking forward*. Singapore: Pearson.
- Tan, K. H. K. & Deneen, C. C. (2015). Aligning and sustaining meritocracy, curriculum and assessment validity in Singapore. *Assessment Matters*, 7(1), 31-52.
- Tan, S. H. (2003). Theoretical ideals and ideologized reality in language planning. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 45-64). Singapore: Eastern Universities Press.
- Tan, T. K. Y. (1986, March 21). Main and development estimates of Singapore for the financial year 1st April, 1986 to 31st March, 1987. *Singapore Parliament Reports*. Retrieved March 7, 2018 from http://sprs.parl.gov.sg/search/topic.jsp?currentTopicID=00059938-ZZ¤tPubID=00069511-ZZ&topicKey=00069511-ZZ.00059938-ZZ_1%2Bid012_19860321_S0002_T00021-budget%2B

- Tan, Y. K., Chow, H. K. & Goh, C. (2008). *Examinations in Singapore: Change and continuity (1891-2007)*. Singapore: World Scientific.
- Tashakkori, A. & Teddlie, C. (Eds). (2003). *Handbook of mixed methods in social and behavioural research*. Thousand Oaks, CA: Sage Publications.
- Taylor, L. (2004a). IELTS, Cambridge ESOL examinations and the Common European Framework. *Research Notes*, 18, 2-3.
- Taylor, L. (2004b). Issues of test comparability. *Research Notes*, 15, 2-5.
- Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.
- Teng, A. (2016, October 30). Exam stress among the young: When grades define worth. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/singapore/when-grades-define-worth>
- Teng, A. & Yang, C. (2016, April 17). Going beyond grades: Evolving the Singapore education system. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/singapore/education/going-beyond-grades-evolving-the-singapore-education-system>
- Teo, J. (2017, March 14). More children and teens are stressed out. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/singapore/health/more-children-and-teens-are-stressed-out>
- Teo, K. S., Soh, Y. A., Wong, C. C. & Chua, L. K. (2014, May). *An exploratory study on the use of two standard setting methods in the validation of mother tongue language descriptors: The Singapore experience*. Paper presented at the 40th International Association for Educational Assessment Conference, Singapore.
- Tharman, S. (2004, November 8). Speech by Mr Tharman Shanmugaratnam, Minister for Education, at the Academy Of Principals' Global Education Conference 2004. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/speech-by-mr-tharman-shanmugaratnam-minister-for-education-at-academy-of-principals-global-education-conference.pdf>
- Tharman, S. (2005, September 22). Speech by Mr Tharman Shanmugaratnam, Minister for Education, at the Moe Work Plan Seminar 2005. *MOE Speeches*. Retrieved March 7, 2018 from <https://www.moe.gov.sg/docs/default-source/document/media/archive/speech-by-mr-tharman-shanmugaratnam-minister-for-education-at-the-moe-work-plan-seminar-2005.pdf>
- The Guardian Editorial. (2013, December 5). David Cameron urges British students to ditch French and learn Mandarin. *The Guardian*. Retrieved March 7, 2018 from <http://www.theguardian.com/politics/2013/dec/05/david-cameron-ditch-french-learn-mandarin-china>

- The Lianhe Zaobao Editorial. (2014, July 7). Elevate the social status of Chinese and Mandarin [提高华语文的社会地位]. *Lianhe Zaobao* [联合早报], p. 21.
- The Lianhe Zaobao Editorial. (2018, October 2). Patience in breaking free from grade obsession [耐心摆脱分数主义]. *Lianhe Zaobao* [联合早报], p. 20.
- The Office of Chinese Language Council International. (2014). *Hanyu Shuiping Kaoshi (HSK)*. Retrieved March 7, 2018 from http://english.hanban.org/node_8002.htm
- The Office of Chinese Language Council International. (2016). *New reflections on Hanyu Shuiping Kaoshi (HSK)* [对新汉语水平考试的新思考]. Retrieved March 7, 2018 from <http://www.chinesetest.cn/gonewcontent.do?id=5589526>
- The Straits Times Editorial. (1990, March 17). Bilingual education: The three issues Singapore must tackle. *The Straits Times*, 1, 23.
- The Straits Times Editorial. (1997, January 1). Language, stability and the future. *The Straits Times*. Retrieved March 7, 2018 from <http://eresources.nlb.gov.sg/newspapers/digitised/issue/straitstimes19970101-1>
- The Straits Times Editorial. (2018, October 4). Giving students more space to learn. *The Straits Times*. Retrieved October 27, 2018 from <https://www.straitstimes.com/opinion/st-editorial/giving-students-more-space-to-learn>
- The World Bank. (2005). *Implementation completion and results report: Guidelines*. Washington, DC: World Bank Publications.
- The World Bank. (2015). *Overview of China*. Retrieved March 7, 2018 from <http://www.worldbank.org/en/country/china/overview>
- Thomson, S., Hillman, K. & De Bortoli, L. (2013). *A teacher's guide to PISA reading literacy*. Retrieved March 7, 2018 from https://www.acer.edu.au/files/PISA_Thematic_Report_-_Reading_-_web.pdf
- Thorndike, E. L. (1921). *The teacher's word book*. New York: Teachers College, Columbia University.
- Toh, C. C. (1964, May 16). Speech by the Deputy Prime Minister, Dr Toh Chin Chye, at the meeting of the Singapore Advisory Committee of the University of Cambridge Local Examinations Syndicate. *Singapore Government Press Release*. Retrieved March 7, 2018 from <http://www.nas.gov.sg/archivesonline/speeches/record-details/7831251b-115d-11e3-83d5-0050568939ad>
- Toh, E. & Ong, A. (2011, October 8). English in school, mandarin at home: Mr Lee urges parents to ensure the young do not lose grasp of language. *The Straits Times*. Retrieved March 7, 2018 from <http://eresources.nlb.gov.sg/newspapers/Digitised/Issue/straitstimes20111008-1>

- Toh, P. G. & Leong, S. C. (Eds). (2014). *Assessment in Singapore: Perspectives for classroom practice*. Singapore: Toppan Leefung Private Limited.
- Toh, P. G. & Leong, S. C. (Eds). (2016). *Assessment in Singapore Volume 2: Strategies and methods for classroom practice*. Singapore: Toppan Leefung Private Limited.
- Toulmin, S. (2003). *The uses of argument: Updated edition*. Cambridge: Cambridge University Press.
- Tremewan, C. (1996). *The political economy of social control in Singapore*. Basingstoke: Palgrave Macmillan.
- Trocki, C. A. (2005). *Singapore: Wealth, power and the culture of control*. London: Routledge.
- Tsui, A. B. M. & Tollefson, J. W. (2007). Language policy, culture, and identity in Asian contexts. New York: Lawrence Erlbaum Associates.
- Tuinman, J. & Brady, M. (1974). How does vocabulary account for variance on reading comprehension tests? A preliminary instructional analysis. In P. Nacke (Ed.), *Interaction: Research and practice for college-adult reading. Twenty-third yearbook of the National Reading Conference* (pp. 176-184). Clemson, SC: National Reading Conference.
- Unaldi, A. (2010). *Investigating reading for academic purposes: Sentence, text and multiple texts*. Retrieved March 7, 2018 from <http://uobrep.openrepository.com/uobrep/handle/10547/279255>
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (1953). *The use of vernacular languages in education*. Paris: UNESCO.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2001). *International conference on education 46th session: Final report*. Paris: UNESCO.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2014). *New UN resolution calls for UNESCO to continue its catalysing role in the fight against illiteracy*. Retrieved March 7, 2018 from <https://en.unesco.org/news/new-resolution-calls-unesco-continue-its-catalysing-role-fight-against-illiteracy>
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2016). *Education: Statistics on literacy*. Retrieved March 7, 2018 from <http://www.unesco.org/new/en/education/themes/education-building-blocks/literacy/resources/statistics>

- University of Connecticut. (2015). *Online Reading Comprehension Assessment*. Retrieved March 7, 2018 from <http://www.orca.uconn.edu/>
- Urquhart, A. H. & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York: Longman.
- Valencia, S. W. & Pearson, P. D. (1987). Reading assessment: Time for a change. *The Reading Teacher*, 40(8), 726-732.
- Verheij, B. (2005). Evaluating arguments based on Toulmin's scheme. *Argumentation*, 19, 347-371.
- Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N. & North, B. (2009). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
- Vygotsky, L. S. (1987). *The collected works of L.S. Vygotsky: Problems of general psychology*. New York: Plenum.
- Wall, D. (1997). Test impact and washback. In C. Clapham & D. Corson (Eds), *Encyclopedia of language education*, (pp. 291-302). Dordrecht: Kluwer Academic Publisher.
- Wang, C. M., Lee, W. H., Lim, H. L. & Lea, S. H. (2014). *Development of the proficiency descriptors framework for the teaching, learning and assessment of mother tongue languages in Singapore*. Retrieved March 7, 2018 from http://www.iaea.info/documents/paper_371f2788e.pdf
- Wang, L. (2008). Some concepts of readability formula and relevant research paradigm as well as the research tasks of formula in TCFL [可读性公式的内涵及研究范式—兼议对外汉语可读性公式的研究任务]. *Language teaching and linguistics study* [语言教学与研究], 6, 46-53.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20, 158-178.
- Wee, K. W. (1989, January 9). President's address. *Singapore Parliament Reports*. Retrieved March 7, 2018 from <http://sprs.parl.gov.sg/search/topic.jsp?currentTopicID=00061244-ZZ>
- Wei, X.N. (2012). The interpretation and enlightenment of "catchwords" in reading instruction [语文阅读教学“流行语”的解读与启示]. *Curriculum, Teaching Material and Method* [课程. 教材. 教法], 3(9), 44-49.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, 9(1), 27-55.

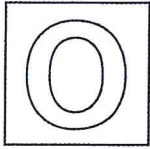
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.
- Weir, C. J., Hawkey, R., Green, A. & Devi, S. (2012). The cognitive processes underlying the academic reading construct as measured by IELTS. In L. Taylor & C. J. Weir (Eds), *Research in reading and listening assessment* (pp. 212-269). Cambridge: Cambridge University Press.
- Weir, C. J., Hawkey, R. Green, A., Devi, S., Unaldi, A. (2009). The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university. In L. Taylor & P. Thompson (Eds), *IELTS research report volume 9* (pp. 97-156). London: British Council/IDP Australia.
- Weir C. J. & Porter, D. (1994). The multi-divisible or unitary nature of reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language*, 10(2), 1-19.
- Weir, C. J., Vidaković, I. & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English language examinations 1913-2012*. Cambridge: Cambridge University Press.
- Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, MA: Harvard University Press.
- Wenger, E. (2009). A social theory of learning. In K. Illeris (Ed.), *Contemporary theories of learning: Learning theorists...in their own words* (pp. 209-218). London: Routledge.
- Widdowson, H. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.
- Wigfield, A. (2000). Facilitating young children's motivation to read. In L. Baker, M. J. Dreher & J. T. Guthrie (Eds), *Engaging young readers* (pp. 140-158). New York: Guilford.
- Williams, E. & Moran, C. (1989). Reading in a foreign language at intermediate and advanced levels with particular reference to English. *Language Teaching*, 22, 217-228.
- Williams, E. & Standish, P. (2015). Learning and philosophy. In D. Scott & E. Hargreaves (Eds), *The Sage handbook of learning* (pp. 51- 61). London: Sage Publications.
- Wilson, H. E. (1978). *Social engineering in Singapore: Educational policies and social change 1819-1972*. Singapore: Singapore University Press.
- Wolf, M. (2007). *Proust and the squid: The story and science of the reading brain*. New York: Harper.

- Wong, F. & Gwee, Y. H. (1980). *Official reports on education: Straits settlements and the federated Malay states (1870-1939)*. Singapore: Pan Pacific Book Distributors.
- Wood, R. (1993). *Assessment and testing: A survey of research*. Cambridge: Cambridge University Press.
- Woon, W. (1992, September 13). Good values can be passed on in any language. *The Straits Times*. Retrieved March 7, 2018 from <http://eresources.nlb.gov.sg/newspapers/Digitised/Issue/straitstimes19920913-1>
- World Health Organization. (2011). *Maternal, new-born, child and adolescent health: Adolescent development*. Retrieved March 7, 2018 from http://www.who.int/maternal_child_adolescent/topics/adolescence/dev/en/
- Xia, Z. J. (2001). Understanding and supporting reading development in the elementary and middle grades [试论中小學生語文閱讀能力的層級結構及其培養]. *Curriculum, Teaching Material and Method* [課程. 教材. 教法], 2, 8-13.
- Xu, D. M., Chew, C. H. & Chen, S. C. (2003). Language use and language attitudes in the Singapore Chinese community. In S. Gopinathan, A. Pakir, W. K. Ho & V. Saravanan (Eds), *Language, society and education in Singapore: Issues and trends* (pp. 133-154). Singapore: Eastern Universities Press.
- Yang, C. X. & Yang, Z. M. (2001). Supporting reading development in the middle grades [试论中学語文閱讀能力的培養]. *Education Exploration* [教育探索], 2, 38-39.
- Yang, S. J. (1971). *A readability formula for Chinese language*. (Unpublished doctoral thesis). University of Wisconsin, Wisconsin.
- Yip, J. S. K. (1997). Reflections and renewal in education. In J. Tan, S. Gopinathan & W. K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 385-400). Singapore: Prentice Hall.
- Yip, J. S. K., Eng, S. P. & Yap, J. Y. C. (1997). 25 years of educational reform. In J. Tan, S. Gopinathan & W. K. Ho (Eds), *Education in Singapore: A book of readings* (pp. 3-32). Singapore: Prentice Hall.
- Young, M. D. (1958). *The rise of the meritocracy*. London: Thames and Hudson.
- Zaccheus, M. (2017, August 27). Study-linked stress a growing concern. *The Straits Times*. Retrieved March 7, 2018 from <http://www.straitstimes.com/singapore/education/study-linked-stress-a-growing-concern>
- Zeng, J. & Wan, M. H. (2012). Reading pedagogy: Perspectives from cognitive psychology [心理學視域下的語文閱讀教學內容重建]. *Education Research Monthly* [教育學術月刊], 2, 73-76.

- Zhang, L. M. & Soh, K. C. (2016). Assessment literacy of Singapore Chinese language teachers in primary and secondary schools. In K. C. Soh (Ed.), *Teaching Chinese language in Singapore: Retrospect and challenges* (pp. 85-103). Singapore: Springer.
- Zhang, N. Z. (2000). A quantitative analysis of the readability of Chinese language teaching materials [汉语教材语料难度的定量分析]. *Chinese Teaching in the World* [世界汉语教学], 3, 83-88.
- Zhao, S. & Liu, Y. (2008). Home language shift and its implications for language planning in Singapore: From the perspective of prestige planning. *The Asia Pacific-Education Researcher*, 16(2), 111-126.
- Zhao, S. & Liu, Y. (2010). Chinese education in Singapore: Constraints of bilingual policy from the perspectives of status and prestige planning. *Language Problems and Language Planning*, 34(3), 236-258.
- Zhou, M. (2003). Design teaching with “purpose as guidance and activity as centre” and develop reading ability [“以目标为导向、活动为中心”设计教学发展语文阅读能力]. *Theory and Practice of Education* [教育理论与实践], 11, 53-54.
- Zhu, X. H. (2014). *Assessment for learning: Principles and strategies* [促进学习的语文评估：基本理念与策略]. Beijing: People’s Education Press.
- Zhu, X. H. (2015). *Assessment for learning: Reading* [促进学习的阅读评估]. Beijing: People’s Education Press.
- Zieky, M. & Perie, M. (2006). *A primer on setting cut scores on tests of educational achievement*. Retrieved March 7, 2018 from https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf
- Zwaan, R. A. & Rapp, D. N. (2006). Discourse comprehension. In M. Traxler & M. A. Gernsbacher (Eds), *Handbook of psycholinguistics* (pp. 725-764). San Diego: Elsevier.

Appendix A

GCE 1162 reading examination latest examination format (May 2016 onwards) sample paper



SINGAPORE – CAMBRIDGE
General Certificate of Education Ordinary Level

华文
试卷二

1160/2

CHINESE
Paper 2

SPECIMEN PAPER
With effect from 2016 examination

1 hour 30 minutes

Additional Materials: Answer Book

考生须知

1. 本试卷有三项试题：综合填空、阅读理解一和阅读理解二。
2. 细读每一项的说明后才作答。
3. 答案必须写在所提供的作答簿内。

本试卷的试题印在第3页至第13页。



Singapore Examinations and Assessment Board

©MOE

一 综合填空(每题1分,共占10分)

根据短文的内容和上下文的意思,选出括号中最适当的词语,然后把代表它的数字写在作答簿的格子内。

在日常生活中,你是否有过心情 Q1 (1 恶劣 2 沉静 3 安逸 4 懈怠) 得什么都不想做经验呢? 其实,这两者之间有着因果关系:当你开始懒散时,人就很容易变得 Q2 (1 悲凉 2 焦虑 3 沮丧 4 急躁); 当你振作起来时,那些让你陷入心理 Q3 (1 处境 2 困境 3 境界 4 境地) 的事,反而都忘光了。

与其无助地坐在那儿,不如把时间拿来做些别的事,例如完成作业,或者去运动,这总比 Q4 (1 咎由自取 2 怨声载道 3 不务正业 4 无所事事) 来得好。

体力劳动可以让脑袋清醒,反应也较 Q5 (1 灵敏 2 警惕 3 利落 4 积极)。你看许多企业家,工作已经非常忙碌了,他们仍然会抽空去运动,让身心保持最好的 Q6 (1 常态 2 状态 3 姿态 4 形态)。这样,才能在瞬息万变的商场中,迅速地做出正确的 Q7 (1 抉择 2 选项 3 展望 4 企盼)。

再说,当我们身体活动时, Q8 (1 精神 2 四肢 3 心灵 4 身手) 得以舒展,心情也为之开朗,一切烦恼就烟消云散了。如果能这么做,又怎么会因为一时想不开而发生 Q9 (1 意外 2 灾难 3 祸害 4 悲剧) 呢?

一个顽强的选手不抵达终点是不会放弃的,他们也不会被生活上小小的困难给 Q10 (1 践踏 2 侵占 3 打垮 4 干扰)。也许我们不是运动员,但是却可以学习运动员的精神。

(接下页)

二 阅读理解一（每题2分，共占20分）

根据短文的内容，选出各题最适当的答案，然后把代表它的数字写在作答簿的格子内。

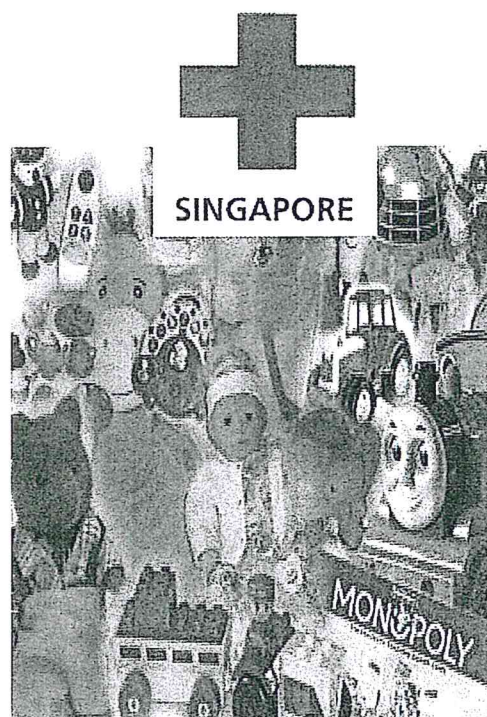
A组（Q11）

分享愉悦，捐赠玩具

为了纪念印度洋大海啸一周年，新加坡红十字会将于本月26日至30日举办“分享愉悦，捐赠玩具”的活动。

红十字会将通过民间组织和志愿福利团体，把玩具运送到灾区，希望为儿童带来欢乐。在把玩具运往灾区前，义工将为玩具进行消毒并包装。

有意捐赠者，可在早上11时至傍晚7时之间，将玩具交到淡滨尼地铁站旁的“海啸摄影展”会场。



Q11 这篇短文的用意是什么？

- (1) 鼓励公众踊跃出席“海啸摄影展”。
- (2) 赞扬一些民间组织和志愿福利团体。
- (3) 向公众保证义工会为所有玩具消毒并包装。
- (4) 呼吁公众支持“分享愉悦，捐赠玩具”的活动。

B组(Q12—Q14)

美铃满足于自己拥有的特别财富——歌唱。

美铃与歌唱结下不解之缘，是从中学开始的。她拜师学艺，学唱艺术歌曲，参加了多次的音乐考试。毕业后，她当了音乐老师，并且负责指导学校的合唱团。

许多朋友和她见面时都不叫她的名字，直呼“那个很爱唱歌的人”。对于这个称呼，她引以为荣。现在，美铃已经退休了，但依然到音乐学校学习中国民歌，也学习流行歌曲的演唱技巧，歌艺也日益精进。

在家里，美铃的女儿能弹钢琴，母女偶尔会一弹一唱，自得其乐。不论是什么音乐，只要听见了，她体内的歌唱细胞就开始活跃起来。她觉得唱歌不只能消磨时间，也为她带来了心灵的喜悦。

Q12 为什么美铃满足于自己的歌唱？

- (1) 歌唱让她消磨时间。
- (2) 歌唱让她得到许多赞美。
- (3) 歌唱让她得到家人的支持。
- (4) 歌唱让她的生活十分充实。

Q13 美铃的歌艺越来越好的主要原因是什么？

- (1) 热爱歌唱，不断学习。
- (2) 朋友的支持，家人的鼓励。
- (3) 和音乐结下良缘，获得名师的指导。
- (4) 天生有才华，从中学开始就接触音乐。

Q14 以下哪一点是文中没有提到的？

- (1) 美铃学过流行音乐。
- (2) 美铃善于表演乐器。
- (3) 美铃能唱中国民歌。
- (4) 美铃参加音乐考试。

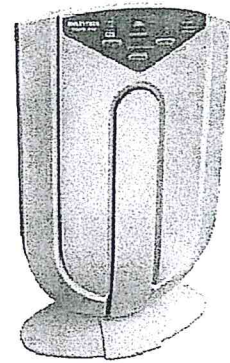
(接下页)

C组 (Q15—Q16)

新一代空气净化器登场

在人口密集的大城市，绿色空间有限，空气污染的情况相对严重，导致人们容易生病，影响体质。因此，环新公司推出了一种有别于使用过滤网的新型空气净化器。

- 特点：
- 1.采用最新的离子科技，操作时能不断释放有益健康的负离子。
 - 2.体形轻巧，操作时不会发出声响。
 - 3.耗电量少过6瓦特，可以一整天使用。



这种新型空气净化器的用途广泛，有除臭、去毒、净化空气的功能，适合安装在住家、酒店、娱乐场所等地方。

环新公司正在明达坊举行三天产品展销会，提供特别优惠。凡在这三天（2月5日至2月7日）前来购买的顾客，可以获得八折优惠。

赶快！别错失良机！

Q15 空气净化器能除臭、去毒和净化空气，靠的是_____

- (1) 不断释放的负离子。
- (2) 能消除杂质的过滤网。
- (3) 轻巧，可随身携带的体形。
- (4) 省电，可全天空使用的功能。

Q16 广告“别错失良机”中的“良机”指的是什么？

- (1) 净化器的价格优惠，十分难得。
- (2) 在展销期间购买产品，享受八折优惠。
- (3) 产品展销会只举行三天，没购买很可惜。
- (4) 购买用途广泛的空气净化器，保持健康。

(接下页)

D组(Q17—Q20)

我小时候是一个公认的非常淘气的坏男孩。九岁那年，母亲去世了，父亲娶了继母。我们第一次见面时，父亲告诉继母，我是整个邻里中最坏的男孩。这让我伤心透了，更想表现得坏一些来气气父亲。

但出乎我意料的是，继母没有显露出嫌恶的表情，反而面带微笑地向我走来，抚摸着我的头，对父亲说：“你错了。他不是最坏的男孩，而是最聪明的男孩，只是他还没找到发挥才能的地方。”

这句话改变了我，改变了我的一生，至今想起来，我仍然觉得心里热乎乎的。

如果你也曾经像我那样被认为是“最坏的孩子”，请不要气馁，你要对自己说“我是行的，只是还没找到自己的优点”。相信自己，然后努力向好的方面转变，你会发现你的确是行的。

自信是前进的动力。只要心存自信，外力是不容易摧毁你的。相反的，如果连自己都不相信自己，别人还会认为你行吗？有了自信，就有争取成功的勇气，敢于朝着目标踏上第一步。

当然，我还必须进一步说明，自信是指适度的自信，也就是既不自卑，也不自大。只要认识自己的优点，了解自己的局限，你就不会被恶意的抨击打败，也不至于被一时的成就冲昏脑袋。

Q17 每当想起第一次和继母见面的情景，“我”心中为什么热乎乎的？

- (1) 继母是一位善良热情的长辈。
- (2) 继母和一般继母的形象不同。
- (3) 继母的笑容令“我”感到很意外。
- (4) 继母给了“我”从没得过的鼓励。

Q18 “我”对坏孩子的劝告是什么？

- (1) 忘掉过去所做的事，不要气馁。
- (2) 下决心做一个好人，争取成功。
- (3) 努力表现自己的特长，并改过自新。
- (4) 设法掌握一生的命运，且实事求是。

Q19 自信是前进的动力，主要在于它能_____

- (1) 避开外来的障碍。
- (2) 激发奋斗的勇气。
- (3) 使人认清前进的目标。
- (4) 令别人对自己有信心。

Q20 “适度的自信”指的是什么？

- (1) 很清楚自己的优点多于缺点。
- (2) 不因他人的抨击而心灰意冷。
- (3) 不因一时的胜利而得意忘形。
- (4) 在自卑和自大之间取得平衡。

(接下页)

三 阅读理解二（10题 40分）

根据短文的内容，完成下列各题，并把答案写在作答簿的横线上。

A组（Q21—Q25）

从前，我不论在家里，还是在工作场所，都喜欢“帮助对方”。当孩子问我作文怎么写时，我就详细地告诉她可以这样写、那样写。当护士长时，我跟在护士背后，收这捡那的。有时护士打完针，针筒就留在宝宝旁边，我常会紧张地帮她们收拾。

有一天，护理部的主管告诉我：“与其送鱼给他，不如教他捕鱼的方法。”我听了心中立即有所领悟，彻底改变了一贯的做法。

当晚刚好女儿赶着交作文，又来问我怎么写了。这次，我只告诉她大纲，让她自己思考，把想法表达出来。起初，她愁眉苦脸，似乎不知如何下笔。但不久之后，就渐入顺境。结果，写出来的作文，水平竟大大提高了，不只出现了许多独特的见解，所用的比喻也很有创意。

此后，我对人对事的态度就从“帮助对方”改成“让他自助”。

护士小姐没做完的事，我会提醒她，由她自行收拾。最后她们都知道工作该怎么做才完善。

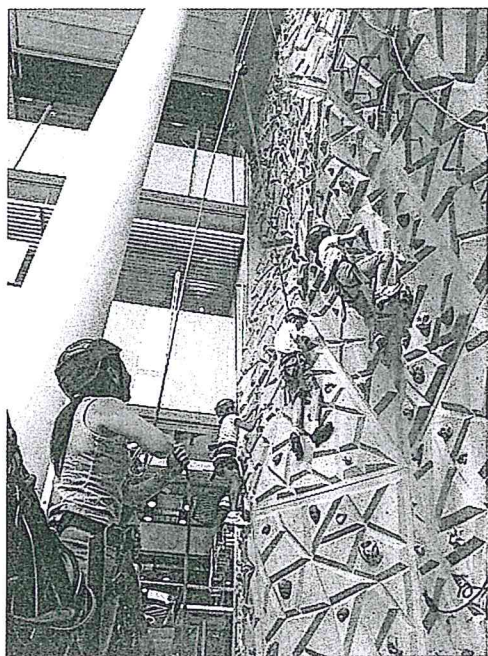
早上读报，获知我国政府基于经济表现良好，决定给全民派发“增长配套”。这等于是在金钱上惠及百姓，而且是收入越低，所得越多。这是政府关心民生的善举，令人欣慰。对一些穷人来说，可能还是一场及时雨。但更值得一提的是，报章的另一则新闻，报道政府将鼓励及帮助低薪工友继续提升工作技能，这可是国家级的“让他自助”了。

- Q21 试以“我”当护士长的工作为例，说明“帮助对方”和“让他自助”所造成的不同结果。（4分）
- Q22 “让他自助”使“我”的孩子的作文出现了怎样的变化？为什么会有这样的变化？（5分）
- Q23 “我”是否反对政府在金钱上帮助收入低的人民？从哪里看出来？（3分）
- Q24 试解释下面两个短语在文中的意思：（4分）
(a) 渐入顺境（第三段）
(b) 一场及时雨（第六段）
- Q25 “与其送鱼给他，不如教他捕鱼的方法”，你赞成这个说法吗？为什么？试举例说明。（4分）

(接下页)

B组(Q26-Q30)

今天，越来越多年轻人喜欢攀墙运动，像电影中的蜘蛛侠一样，在墙壁上轻松自如地走动。为什么年轻人喜欢攀墙运动呢？



“这不奇怪，人类的老祖宗——猴子，不是一天到晚爬上爬下的吗？”攀墙教练再诺半开玩笑地说，“喜欢攀墙的人可能‘原始的本性’还没有消失吧。”

同样是迎向最高点，登山运动考验的是脚力，要你一步一步地走向目的地。“新加坡没有高山，也没有险峻的石壁，年轻人喜欢富有挑战的活动，所以就来攀墙。”再诺解释，“攀墙运动靠的是四肢的力量，一个动作接着一个动作，攀向墙头。”

刚参加训练课程的周伟强和刘小明，就亲身体会了征服一面墙之后的满足感。这两名男孩抱着姑且一试的心理，结果攀上约三层楼高的墙，这是他们从来没有想过的。开始的时候，他们不觉得困难，但越往上攀就越辛苦。由于可供支持的点和抓的地方距离越来越远，人往往悬在半空中，恐惧一直压抑着信心。能够突破这种心理障碍，那种满足感是难以形容的。

攀墙运动对意志力的训练远远超过体力。进行这种运动的人都必须凭着坚定的意志力，利用墙壁上的凹凸处，一步一步地从地面爬到墙的顶端。就算你体力再好，但只要少了一点恒心，也未必能攀上墙头。

“攀墙运动提供的训练，对攀墙者有多方面的益处。”再诺肯定地说。

- Q26 文中“原始的本性”指的是什么？（2分）
- Q27 登山和攀墙这两项运动，有什么相同和不同的地方？（4分）
- Q28 为什么说攀上墙头后，那种满足感是难以形容的？（5分）
- Q29 攀墙运动对意志力的训练远远超过体力，你同意吗？为什么？（4分）
- Q30 本文对你的学习有什么启发？试写出你的看法。（5分）

完

Appendix B

GCE 1162 reading examination new examination format

(May 2012-November 2015) sample paper



SINGAPORE – CAMBRIDGE
General Certificate of Education Ordinary Level

华文
试卷二

1162/2

CHINESE
Paper 2

October/November 2013

1 hour 30 minutes

Additional Materials: Answer Book

考生须知

1. 本试卷有三项试题：综合填空、阅读理解一和阅读理解二。
2. 细读每一项的说明后才作答。
3. 答案必须写在所提供的作答簿内。

本试卷的试题印在第3页至第13页。



Singapore Examinations and Assessment Board

©MOE 2013

Oct/Nov 2013 Paper 2 (I)

一 综合填空(每题1分,共占10分)

根据短文的内容和上下文的意思,选出括号中最适当的词语,然后把代表它的数字写在作答簿的格子内。

天地万物,除了人类以外,没有其他的动物能使用语言。人类利用语言来 Q1 (1传递 2沟通 3交往 4释放) 信息,交流感情。

说话是一门学问。一句话要意思完整又能被人接受,需要事先 Q2 (1安排 2研究 3斟酌 4探讨),而说话的时间、地点与对象,也必须加以注意。一个善用语言的人 Q3 (1固然 2简直 3明明 4刚好) 就是技艺精湛的厨师。他对各种食材的调配,火候的拿捏,都要做到 Q4 (1熟能生巧 2无懈可击 3天衣无缝 4出类拔萃)。

雨露能 Q5 (1灌注 2灌输 3滋润 4滋生) 花草,贴切的话语也能让人听了心情舒畅。然而,不恰当的话语,会制造种种麻烦和冲突。因此,当我们用语言表情达意时,必须十分 Q6 (1谨慎 2细密 3精致 4稳重)。

语言是一把双刃剑,怎么使用要看掌握在谁的手中。自古以来,小人就善于利用 Q7 (1只言片语 2流言蜚语 3三言两语 4花言巧语) 骗取别人的信任,以达到不可告人的目的。反过来说,善良的人则往往会以真诚的话语来 Q8 (1化解 2解除 3松解 4解脱) 人与人之间的纠纷和怨恨。

我认为说话应该像画画儿一样,用笔要恰到好处,增一笔嫌多,减一笔则不足,这样才算 Q9 (1出众 2高明 3准确 4优秀)。说话时意思必须完整,不要因过繁而令人 Q10 (1劳累 2困倦 3厌烦 4恍惚),也不要因过简而使人费解。这样,就可以充分发挥语言的力量了。

(接下页)

二 阅读理解一（每题2分，共占20分）

根据短文的内容，选出各题最适当的答案，然后把代表它的数字写在作答簿的格子内。

A组（Q11）

视障者日后只要下载一个简单的智能手机应用程序，就能靠手机发出的指示，独自到达目的地。



这个采用全球卫星定位系统的应用程序，不但能让视障者确定自己的所在地，也能一步一步引导他们到要去的地方。比如，当他们搭巴士时，这个应用程序会通过声音告诉他们所在的巴士站和即将到站的巴士号码，让他们出门更方便。此外，用户还可以通过它了解天气情况。

来自理工学院的三名学生在老师的指导下，研发了这个导盲系统，这也是他们的毕业作品。

Q11 这篇短文的主要内容是什么？

- (1) 学生为了帮助视障者而研发出一款智能手机。
- (2) 学生采用全球卫星定位系统完成了毕业作品。
- (3) 新研发的应用程序能够让用户了解天气情况。
- (4) 新研发的导盲系统通过智能手机帮助视障者。

1162/2/2013

Oct/Nov 2013 Paper 2 (3)

B组(Q12 – Q13)

我家附近的一条路上，每逢周末都有跳蚤市场。跳蚤市场由居民委员会主办，居民只须花上10元，就可以租一个摊位来摆卖物品。

摊主所摆卖的东西可说是琳琅满目。有时发现摊位上别人不想要的东西，竟然是自己很想要的，你就会有意外的惊喜。

我们一般都以为跳蚤市场的摊主会漫天开价，其实不然，很多人常常能以低价买到自己所需要的东西。由于跳蚤市场就在地铁站附近，来来往往的人很多。一些外国女佣和客工收入有限，也很喜欢来这里买东西。

经营这样的摊子有苦也有乐。有时碰到骤雨，摊主就会手忙脚乱。不过，乐的是可以打发时间，赚些零用钱，同时也能为环保尽一份力。一举多得，何乐而不为？

Q12 外国女佣和客工为什么喜欢光顾跳蚤市场？

- (1) 跳蚤市场都很靠近地铁站。
- (2) 在跳蚤市场上能买到便宜货。
- (3) 跳蚤市场是由居民委员会主办的。
- (4) 在跳蚤市场上摆卖的东西种类繁多。

Q13 作者通过这篇短文表达了什么看法？

- (1) 跳蚤市场有它存在的价值。
- (2) 跳蚤市场处处都充满惊喜。
- (3) 跳蚤市场的摊主有苦也有乐。
- (4) 跳蚤市场满足了所有居民的需要。

(接下页)

电子阅读器 阅读乐趣多

电子阅读器的外形美观，携带方便，价格便宜，深受年轻人的欢迎。

随着科技的进步，电子阅读器的功能越来越多。为了满足用户的需求，它可以显示各种格式的电子书和播放多种格式的声频。用户不仅可以看书，还可以听书。



电子阅读器非常轻巧，机身薄，最轻的重量仅191克，机身最厚的部分仅有8.4毫米。用户可以将它放进口袋里，随时随地拿出来阅读。

电子阅读器也具有不反光、低辐射等特点，即使长时间阅读，眼睛也不会疲劳。哪怕是在阳光直射下，它也能保持清晰的显示效果，让用户尽情享受阅读乐趣。

Q14 从哪里看出电子阅读器的功能越来越多？

- (1) 它让用户感到物有所值。
- (2) 它长时间保持清晰的影像。
- (3) 用户一直阅读而不感到疲劳。
- (4) 用户用它来看和听各种电子书。

Q15 电子阅读器如何让用户随时随地轻松阅读？

- (1) 它的辐射低，适合长时间注视。
- (2) 它的价格便宜，人人都负担得起。
- (3) 它的体形轻巧，具有不反光的特点。
- (4) 它的功能多，能引起用户看书的兴趣。

(接下页)

D 组 (Q16 — Q20)

新加坡虽然是个小国，不过根据调查，平均每人排放的废气竟然排名全球第七。我国享有花园城市的美名，保护环境的工作却差强人意，这的确令人失望。

有人认为这样单纯计算人均废气排放量对我们是不公平的。新加坡缺乏天然资源，无法靠风力和水力来发电，只能依靠会排放废气的燃料来发电。虽然我国的废气排放只占全球的0.2%，可是我们也不能因此而推卸责任。

国人的生活方式迅速现代化，很多时候造成了浪费。专家指出，如果世界上每一个人都按照新加坡人的方式来生活，那么全世界就需要三个地球的资源才能满足所有人的需求。

我国政府已经征收了很高的汽车税和路税，可是汽车数量还是逐年增加。即使如此，政府抑制汽车数量的决心仍然没有动摇。另一方面，我国也向世界各国承诺将继续与全球气候暖化作战，改用天然气来发电以及研发清洁能源技术。

爱护地球，人人有责。由民间发起的“地球一小时”运动，目的在于提醒人们：在日常生活中只要减少能源的浪费，就能为保护地球尽一份力。

Q16 为什么作者会对我国人均废气排放量的排名感到失望？

- (1) 他认为新加坡的环保工作不够理想。
- (2) 他认为新加坡应该排在第七名之前。
- (3) 排名的计算方式对新加坡不太公平。
- (4) 排名会破坏新加坡花园城市的美名。

Q17 作者说：“如果世界上每一个人都按照新加坡人的方式来生活，那么全世界就需要三个地球的资源才能满足所有人的需求。”这句话说明了什么？

- (1) 新加坡是一个现代化的城市。
- (2) 新加坡是一个缺乏资源的国家。
- (3) 新加坡人往往不重视资源的节约。
- (4) 新加坡人不能推卸破坏环境的责任。

Q18 我国所采取的哪一项措施有助于减少废气的排放量？

- (1) 靠风力和水力发电。
- (2) 改用天然气来发电。
- (3) 采用清洁能源技术。
- (4) 抑制汽车数量增加。

Q19 “与全球气候暖化作战”这句话是什么意思？

- (1) 抵制破坏环境的国家。
- (2) 与全世界合作改善气候变化。
- (3) 通过行动来防止全球气候恶化。
- (4) 号召全世界使用天然气和清洁能源。

Q20 “地球一小时”运动发挥了什么作用？

- (1) 提醒我们地球是人类生存的唯一家园。
- (2) 提醒我们在日常生活中节约能源的重要性。
- (3) 提醒新加坡人要与他人分享地球上的资源。
- (4) 提醒新加坡人奢华的生活方式将会破坏地球。

(接下页)

三 阅读理解二 (10 题 40 分)

根据短文的内容, 完成下列各题, 并把答案写在作答簿的横线上。

A 组 (Q21 - Q25)

女舞蹈员穿着紫色长裙, 随着悠扬的乐曲, 像花间的蝴蝶, 翩翩起舞。男舞蹈员纯熟地带领女舞蹈员前进、后退、前俯、后仰……他们的舞姿令观众赏心悦目。

国标舞在本地还不是很普遍, 然而在国外却相当盛行。在外国, 不少年轻人觉得通过国标舞可以扩大社交圈子, 甚至结识终身伴侣。不过, 本地一名热爱国标舞的舞蹈员却认为: “学国标舞主要还是为了兴趣, 如果真的与舞伴发展成为情侣, 也算是一种额外收获吧!”

记者走访了三所国标舞学院, 受访的教练都认为与十多年前相比, 本地学国标舞的年轻人增加了许多。一名院长说: “人们已逐渐改变对国标舞的看法, 父母也愿意让孩子从小就学习这种舞蹈。”



另一所舞蹈学院的创办人认为: “年轻人开始意识到出席宴会时, 跳高雅的国标舞, 可以突显个人在社交礼仪上的修养, 因此对学跳国标舞产生兴趣。”

近年来, 国标舞经过多次改变, 以新名称“体育舞蹈”替代。年轻人逐渐把它当作一种极富挑战性的运动。由于它兼具文化娱乐和体育竞技的特点, 有很高的观赏性和技艺性, 因此许多国家和地区已经将它列为体育比赛的项目。

Q21 从哪里看出那对舞蹈员的舞艺精湛？(3分)

Q22 试从文中找出年轻人学习国标舞的三个原因。(4分)

Q23 记者走访国标舞学院后，得到哪些信息？(4分)

Q24 为什么国标舞能成为体育比赛的项目？(4分)

Q25 外国和本地的年轻人对学习国标舞有不同的看法。
你认同哪一种？为什么？(5分)

(接下页)

1162/2/2013

Oct/Nov 2013 Paper 2 (10)

B组(Q26—Q30)

朋友德力，只上了三次高山滑雪课程，便运槌如飞。他把这归功于滑雪教练，因为教练曾说过一句充满哲理的话：“别怕跌倒，雪那么软，怎么跌都跌不死。”

学滑雪，如果担心跌倒，还没起步，便先被心里的阴影绊倒了。

德力娓娓道来：“许多刚学滑雪的人从雪山俯冲下来时，看到那一望无际的白，脚便软了，想来个紧急刹车。结果呢，那原来强劲的冲势，反让他们跌得四脚朝天，有的甚至伤到椎骨呢！”

不怕跌，代表了“不怕失败”的精神，是一种斗志和毅力的表现。这种性格是自小由家庭培育出来的。

的确，在人生的旅途中，一句话往往足以影响我们的成败。

台湾一家身心灵整合中心的负责人就指出：在孩子成长过程中，长辈的言谈举止常常会影响孩子对事物的看法。比如，小孩一跌倒，长辈就心疼。祖母会大力拍打地板，责备地板害她的宝贝跌倒，这等于告诉孩子：“自己跌倒别人的错。”长此以往，孩子可能会在犯错后，不分青红皂白地归咎于人。

至于严厉的父亲，可能会“寓爱于骂”地斥责：“你是怎么搞的？好好的路竟然摔倒！”这会带来相反的效果：孩子以后遇到挫折，就认定是自己笨。

中心的负责人又指出，长辈应该让孩子认识到：跌倒没什么大不了，最重要的是跌倒之后，不论伤势如何，都应该冷静地面对，应该吸取教训，不要重蹈覆辙。

Q26 德力在学滑雪时，和其他初学者的表现有何不同？
(3分)

Q27 试分别解释下面两句话在文中的意思：(4分)

(a) 还没起步，便先被心里的阴影绊倒了(第二段)

(b) 脚便软了，想来个紧急刹车(第三段)

Q28 长辈的言行往往会影晌孩子的未来。作者是如何证明这一点的？(4分)

Q29 作者认为不怕跌，代表了“不怕失败”的精神，是一种斗志和毅力的表现。你同意他的看法吗？为什么？试举生活中的一个例子说明。(4分)

Q30 最后一段，“中心的负责人”所说的那句话给了你什么启示？试加以说明。(5分)

完

Appendix C

GCE 1162 reading examination old examination format

(May 2006-November 2011) sample paper

SINGAPORE-CAMBRIDGE
GENERAL CERTIFICATE OF EDUCATION ORDINARY LEVEL

MID-YEAR EXAMINATION

0
1162/2

华文
试卷二

May/June 2006

CHINESE
Paper 2

1 hour 30 minutes

Additional Material/s : Answer Sheet

考生须知

1. 本试卷有三项试题：综合填空、阅读理解一和阅读理解二。
2. 细读每一项的说明后才作答。
3. 答案必须写在所提供的作答卷上。

本试卷的试题印在第3页至第11页。



Singapore Examinations and Assessment Board

©MOE

May/June 2006 Paper 2(I)

一 综合填空 (每题1分, 共占10分)

根据短文的内容和上下文的意思, 选出括号中最适当的词语, 然后把代表它的数字写在作答卷的格子内。

外国一项调查发现: 自从有了手机, 人们较常迟到。因为大家都一“机”在手, 可以随时 Q1 (1联系 2说明 3交流 4嘱咐), 轻易更改见面的时间。

情况确实如此。朋友聚会, 十次有九次没准时碰面。大家都 Q2 (1兴高采烈 2人云亦云 3礼尚往来 4此起彼落)地传送“对不起, 我会迟到”的简讯, 结果聚会的时间总是延迟了。

因此, 有时候还真怀念手机出现以前的日子。说好了见面的时间, 大家都不敢 Q3 (1改口 2怠慢 3推辞 4食言)。两个人的约会, 不敢迟到, 怕对方独自一人等得太闷; 三个人的约会, 也不敢迟到, 怕其他人都准时, 自己迟到太 Q4 (1突兀 2滑稽 3狂妄 4尴尬); 四个或更多人的约会, 更不敢迟到, 怕大家走了, 把自己 Q5 (1擢 2甩 3揍 4漏)了。因为有这种心理 Q6 (1负担 2状况 3作用 4概念), 大家的时间观念都还很强。如今, 有了手机, 约会可以随时提前或延迟, 时间观念反而渐渐 Q7 (1流失 2微弱 3淡薄 4软化)了。

当然, 话说回来, 没有人会 Q8 (1忽略 2疑惑 3鄙视 4否定)手机对世人的重大贡献。例如, 万一出现突发状况, 不能准时赴约, 手机就为我们提供了解决问题的 Q9 (1良机 2秘诀 3途径 4门路), 让我们可以立即通知对方。

可见事情有正反两面, 重要的是: 我们应该在享受科技所带来的便利的同时, 又不忘保持大家 Q10 (1一贯 2一致 3一味 4一再)注重的守时习惯。

二 阅读理解一(每题2分,共占20分)

根据短文的内容,选出各题最适当的答案,然后把代表它的数字写在作答卷的格子内。

A组(Q11-Q16)

做家庭访问时,最令人憋一肚子闷气的,莫过于遇到明明是有人在家,却又偏偏不开门的情形。任凭你好言相劝,口水都快说干了,喉咙都快喊哑了,手指头都叩得快起泡了,甚至最后一记绝招也使出来了:“病人情况危急,须与家属联络”,结果还是不管用。

若遇上老人单独在家,这趟访问十之八九是要吃闭门羹的。由于最近老人遇上“假菩萨”而痛失棺材本,甚至被谋财害命的新闻满天飞,使他们闻风丧胆,一听到陌生人敲门,便吓得全身发抖,躲进房间,说什么都不肯开门。其实这也难怪他们,警方不是经常提醒老人,为了自身安全,不要随便开门吗?

碰上一些中年夫妇,情况更糟。他们应门后,虽然打开了木门,可是铁门仍然深锁。而打开木门也只是一刹那,犹如闪电划过黑夜,在第一句话还未画上句号时,就“碰”的一声又关上了。连对方长得什么模样,都没看清楚。这碗闭门羹真不是味道!唯一值得安慰的是,报告可以一切从简,三言两语便交代了。

Q11 做家庭访问时,以下哪种情形最令作者难受?

- (1) 叫门叫得口水都快干了。
- (2) 叩门叩得手指都快起泡了。
- (3) 分明有人在家却偏不开门。
- (4) 说了好话还是不见有人开门。

Q12 作者认为老人不肯开门的真正原因是_____。

- (1) 老人要响应警方的呼吁
- (2) 老人听到很多负面的新闻
- (3) 这是对付陌生人的可靠方法
- (4) 这是避免钱财被骗的有效做法

Q13 文中的“闭门羹”指的是什么？

- (1) 做访问的人喉咙喊哑。
- (2) 做访问的人白忙一阵。
- (3) 做访问的人被拒于门外。
- (4) 做访问的人憋一肚子气。

Q14 文中“打开木门也只是一刹那，犹如闪电划过黑夜”，是形容木门_____。

- (1) 很快地被打开
- (2) 没有很好地锁上
- (3) 留下了一条缝隙
- (4) 打开后便立即关上

Q15 文中“还未画上句号”指的是_____。

- (1) 还没开口说话
- (2) 还没把话说完
- (3) 还来不及把话记下来
- (4) 还来不及把话说清楚

Q16 家庭访问虽然没做成，但作者可以自我安慰的是，报告就可以_____。

- (1) 简单了事
- (2) 草草完成
- (3) 只写三句话
- (4) 写得很简练

B组(Q17-Q19)

有一年,在国外的公车上,看到一位母亲谢绝车上乘客的好意,让幼小的孩子拉着扶手站着。她说:“他已经能站了,让他自己站吧!”明知孩子站在拥挤的车厢里,难免会磕碰跌倒,却仍然让孩子自己站着。她相信经过磨练,孩子会越站越稳。

反观我国有不少家长,却不明白这个简单的道理。他们过分宠爱孩子,生活上什么都不放心,深怕孩子受伤害,所以样样包办,事事代劳。试想,这对孩子的成长有帮助吗?

Q17 这篇短文主要说明什么?

- (1) 家长要怎样栽培孩子。
- (2) 家长不懂得教育孩子。
- (3) 家长不应过度保护孩子。
- (4) 家长教育孩子的两种态度。

Q18 母亲为什么拒绝车上乘客的好意?

- (1) 她希望孩子学习自立。
- (2) 她希望孩子越来越好。
- (3) 她认为孩子可以拉着扶手。
- (4) 她不愿孩子接受别人的帮助。

Q19 作者对我国一些家长教育孩子的方法,有什么意见?

- (1) 对他们的做法不置可否。
- (2) 对他们的做法不以为然。
- (3) 觉得他们的做法无可厚非。
- (4) 觉得他们的做法无微不至。

C组 (Q20)

为了纪念印度洋大海啸一周年,新加坡红十字会将于本月26日至30日举办“分享愉悦,捐赠玩具”的活动。

红十字会将通过民间组织和志愿福利团体,把玩具运送到灾区,希望为儿童带来欢乐。在把玩具运往灾区前,义工将为玩具进行消毒并包装。

有意捐赠者,可在早上11时至傍晚7时之间,将玩具交到淡滨尼地铁站旁的“海啸摄影展”会场。

Q20 这篇短文的用意是什么?

- (1) 鼓励公众踊跃出席“海啸摄影展”。
- (2) 赞扬一些民间组织和志愿福利团体。
- (3) 向公众保证义工会为玩具消毒并包装。
- (4) 呼吁公众支持“分享愉悦,捐赠玩具”的活动。

三 阅读理解二(10题40分)

根据短文的内容,完成下列各题,并把答案写在作答卷的横线上。

A组(Q21-Q25)

李显龙总理在推广华语运动开幕仪式上的发言,可谓语重心长,也切中要害。他强调,华人学习华语,有经济上和文化上的理由。掌握华语,使我们保持认同感和文化的根,也让我们开启有数千年悠久历史的中华文明宝库。

如果我们只使用英语,而任由母语没落,有朝一日,将会失去固有的价值观和文化传统。因此,学习华语就不只是因为中国迅速崛起的经济因素而已。李总理的话,对一些因为认识上有盲点而轻视华语的华族同胞而言,肯定能起着发人深省的作用。

总理也提出了如何学好华语的建议:把华语当作活的语言,在日常生活中使用它;懂得讲华语的家长,应该以身作则,为年幼的孩子营造一个有利于学习华语的家庭环境;学校也应该积极地帮助学生有效地学习华语。倘若社会、家庭和学校三方面能紧密配合,那么,推广华语运动必定能继续取得成功。

总理也间接地向国人传达了另一个重要的信息:华语是华人的母语。虽然英语作为国家的行政语言和各族的共同语,是我们必须掌握的,但我们却不可因此而主宾错位,甚至抛掉母语。

这样的基本认识不仅适用于华族,也适用于其他族群。我们绝不能放弃自己的文化根本,我们是亚洲人,必须在放眼世界的同时,植根亚洲。

- Q21 文中指出学习华语有哪方面的理由? 试加以说明。(4分)
- Q22 李总理提出了哪些学好华语的建议? 怎样才能确保这些建议取得成效? (4分)
- Q23 文中为什么把母语和英语的关系看成是主宾关系? (4分)
-
- Q24 试解释下面两个短语在文中的意思。(4分)
- (a) 语重心长, 也切中要害(第一段)
- (b) 认识上有盲点(第二段)
- Q25 作者认为“我们是亚洲人, 必须在放眼世界的同时, 植根亚洲”。你同意吗? 为什么? (4分)

B组 (Q26 - Q30)

农历新年前夕，我收到一位多年不见的童年好友的贺卡。卡片上除了春节祝词，还写着：

“其实，曾经有好长一段时间，我很羡慕你有一个很好的家境，可以让你学琴……”

这位朋友热爱音乐。童年时彼此住家相距不远，放学她特别喜欢来我家听我弹琴。当时我主动表示很乐意教导她，还请她天天到我家练琴。如果她能持之以恒，要练就一身琴艺，并不是太困难的事。

但她来练琴，总是断断续续，没法持久。我一直不明白其中原因，我们相处得这么融洽，她又这么喜欢弹琴，为什么不能持之以恒呢？一直到踏入社会多年后，这迟来的告白才让我知道，原来，彼此家境的差距，曾让她产生这样的心理障碍，而这也就是她在学琴路上始终不能成功的主因。

一向来粗线条的我，那时只会羡慕她那温柔优雅的举止，而根本没有觉察到她的这种微妙心理。

“羡慕情结”，大概每个人都有，只是程度深浅各有不同。其实，这种情结，并非完全负面。处理得当的话，它也可以引导我们走上心灵所向往的路，借此发掘自己的潜能。生活不是一帆风顺的，如果我们能够化“阻力”为“助力”，这样的人生会更精彩。

- Q26 “我”的童年好友为什么会开始学琴? 结果又怎样? (4分)
- Q27 “我”的好友的“羡慕情结”是一种怎样的心理? 试加以说明。(4分)
- Q28 “我”觉得“羡慕情结”是正面还是负面的? 为什么? (4分)
- Q29 试解释下面两个短语在文中的意思。(4分)
- (a) 迟来的告白(第四段)
- (b) 粗线条(第五段)
- Q30 “我”觉得能够化“阻力”为“助力”的人生会更精彩。你同意吗? 为什么? (4分)

(完)

Appendix D

GCE 1162 examination test specifications

GCE O-LEVEL

CHINESE

华文

For Year of Examination from 2016



Singapore Examinations and Assessment Board

Issued in 2014
© MOE

目录

1. 考试目标	2
2. 试卷格式	3
3. 试卷蓝图	5
4. 试卷样本	7
5. 附录	
A 试卷一评分指引	
B 试卷三(口试)评分指引	

本考试纲要从 2016 年开始采用。

1. 考试目标

1.1 本科试卷供修读中学快捷课程的学生选考；普通学术课程的学生，若获得校方的批准，也可以选考。

1.2 本科试卷是遵照《中学华文课程标准》的相关教学目标和教学内容而编制的。中学华文（快捷课程）教学目标旨在加强学生的听说读写的能力，着重读写能力的培养。完成课程后，学生最终达到以下目标：

- 能听懂适合程度的记叙性、说明性、议论性和实用性语料
- 能针对较复杂的话题表达看法与感受，并与人进行有效的交流
- 能阅读适合程度的记叙性、说明性、议论性和实用性语料，并能进行文学欣赏
- 能写适合程度的记叙文、说明文、议论文和实用文，并能初步进行简单的文学创作

1.3 本科试卷主要考查学生下列语文能力：

- 聆听
- 会话
- 词语的认识和语言的应用
- 阅读理解
- 写作电子邮件或不同文体的文章

此外，本科试卷也考查学生综合运用语言技能的能力。

1.4 本科试卷按《试卷蓝图》编制，但可根据需要，酌情作调整。

2.3 试卷三：口试

10 -15 分钟

50 分/25%

这份试卷包括朗读短文和会话。在考试前，考生有 10 分钟的时间默读短文和观看录像短片。考生在限定的时间内，可以多次默读短文和观看录像短片。

第一部分：朗读短文

考生必须朗读一个短文。

第二部分：会话

考生针对所提供的录像短片，以及主考员的提问，跟主考员进行一段对话。

2.4 试卷三：听力理解

约 30 分钟

20 分/10%

这份试卷包括三个简短对话或语段，以及三个理解篇章，共有 10 道选择题。考生先听录音，然后回答问题。考查的内容包括日常会话、广告、说明、故事和新闻报道等。

3. 试卷蓝图

3.1 试卷一：写作（60分）

序数	考查项目	方式	范围	题数	分数比重	备注
一	实用文	开放式	电子邮件	2选1	20/10%	字数在150以上
	作文		记叙文、议论文和说明文	3选1	40/20%	字数在300以上 试卷一的考试格式与2012年的相同
共				2	60/30%	

3.2 试卷二：语文理解与应用（70分）

序数	考查项目	方式	范围	题数	分数比重	备注
一	综合填空	多项选择	1个短文	10	10/5%	试卷二的考试格式与2012年的相同
二	阅读理解一	多项选择	3至4个语料，如广告、传单、新闻报道等	10	20/10%	
三	阅读理解二	自由作答	2至3个短文	10	40/20%	
共				30	70/35%	

3.3 试卷三：口试与听力理解（70分）

序数	考查项目	方式	范围	题数	分数比重	备注
一	口试					
	1) 朗读短文	朗读	日常生活篇章、评论和新闻等	1	10/5%	从2016年起，朗读短文将采用电子版
	2) 会话	看录像短片，然后跟主考员进行对话	日常生活话题	1	40/20%	口试将以录像短片（55至60秒）取代图片
二	听力理解	听录音，然后回答多项选择式的题目	日常会话、广告、说明、故事、新闻报道等	10	20/10%	考试格式与2012年的相同
共				12	70/35%	

Appendix E

Letter of invitation to adult interviewees

Ms Cheong Yun Yee
c/o University College London Institute of Education
Gower Street,
London
WC1E 6BT
United Kingdom

31 July 2015

Dear Sir/Madam,

**Examining Second Language Reading:
A Critical Review of the Singapore-Cambridge General Certificate of Education
Ordinary-Level Chinese Language Examination**

I am an Education Officer previously with the Curriculum Planning Office of the Ministry of Education, Singapore (MOE). I am currently pursuing a PhD in Educational Testing under a MOE Postgraduate Scholarship at University College London Institute of Education.

As part of my research, I am conducting a study to find out what stakeholders know, think and feel about the GCE O-Level Chinese language reading examinations. In this connection, I am seeking your kind consent to be involved in this study. Your involvement will take one or both of the following forms:

- Interview (1 session of 2 hours to be conducted at a place of your convenience).
- Expert Panel (Participants will be grouped into pairs and each pair will be given 12 to 13 texts with the accompanying test items to peruse. The review process can be completed at home over the course of a month. The total time required is estimated to be 10 hours. Training will be provided).

Please be assured that standard rules regarding research ethics will be adhered to in order to protect the identities of participants. Names of participants will not be disclosed and all raw data collected will only be accessed by myself as the researcher and my supervisors. The requisite approval to conduct the research has already been obtained from MOE.

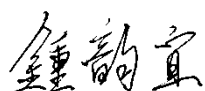
Should you require clarification regarding the research study, please do not hesitate to contact:

Ms Cheong Yun Yee Student Researcher Email: ycheong@ioe.ac.uk
HP: 90408122
Professor David Scott Main Supervisor Email: d.scott@ioe.ac.uk

Please indicate your consent/non-consent on the attached form.

Thank you and I look forward to your support and favourable response.

Yours faithfully



Cheong Yun Yee (Ms)

CONSENT FORM

To: Ms Cheong Yun Yee

Please check box

1. I confirm that I have read and understood the information sheet regarding the above study and have had the opportunity to ask questions.
2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving reason.
3. I agree to take part in the above study.
4. I agree to the interview being audio recorded.
5. I agree to the use of my responses in the form of anonymized quotations in the researcher's doctoral thesis, publications and presentations.

Name: _____

Date: _____

Signature: _____

Appendix F

Letter of invitation to parent/guardian of student interviewees

Ms Cheong Yun Yee
c/o University College London Institute of Education
Gower Street,
London
WC1E 6BT
United Kingdom

5 November 2015

Dear Parent/Guardian,

Request for Consent to Participate in Research Project

I am an Education Officer previously with the Curriculum Planning Office of the Ministry of Education. I am currently pursuing a PhD in Educational Testing through a MOE Postgraduate Scholarship at University College London Institute of Education.

As part of my research, I am conducting a study to find out what students know, think and feel about the GCE O-Level Chinese language reading examinations. In this connection, I seek your kind consent to allow your child/ward to be involved in this study. His/her involvement will take the following form:

- Interview (1 session of approximately 60 minutes) after school which will be audio recorded.

Please be assured that standard rules regarding research ethics will be adhered to in order to protect the identities of participating students. Participation in this study is optional and the activities will not be graded. Student names will not be disclosed and the audio recordings will only be accessed by myself as the researcher and my supervisors.

The requisite approval to conduct the research project has already been obtained from the Ministry of Education and the school. All research activities will be conducted between November 2015 and January 2016.

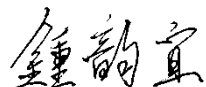
Should you require clarification regarding the research study, please do not hesitate to contact:

Ms Cheong Yun Yee	Student Researcher	Email: ycheong@ioe.ac.uk HP: 90408122
Professor David Scott	Main Supervisor	Email: d.scott@ioe.ac.uk

Please indicate your consent/non-consent on the attached form, which I will collect when I meet the students.

Thank you and I look forward to your support and favourable response.

Yours faithfully



Cheong Yun Yee (Ms)

CONSENT FORM

To: Ms Cheong Yun Yee

I am the parent of _____ of Class _____.

I have read the attached letter dated _____ and understood its contents.

I consent/do not consent* to my child/ward* taking part in the audio recorded research session.

Signature: _____

Date: _____

Name of parent/guardian: _____

* delete as necessary

Appendix G

Interview schedule

Interview Schedule [Final version]

I. Opening

Segment A: Establishing rapport and permission to record

您好！我是钟韵宜。[shake hands] 感谢您抽空接受我的采访。请问我可否录下这段采访的内容的内容？

Hello, I am Yunyee. [shake hands] Thank you for accepting my invitation to participate in this research interview. May I record this interview please?

Segment B: Self-introduction and purpose

我目前在伦敦大学学院 University College London 附属的教育学院修读评估与测试的博士课程，主要研究的是新加坡剑桥 GCE O 水准华文 1162 试卷评量阅读能力的方式。

这次采访您的目的是想了解您对这套试题以及华文作为第二语文阅读的看法。

I am currently pursuing a doctoral degree at the University College London Institute of Education. My research interests lie primarily in the field of summative assessment, particularly in the validation of large-scale examinations. My dissertation sets out to describe and review how the Singapore-Cambridge GCE O-Level Chinese Language Examination (GCE 1162) assesses second language reading ability.

I would like to ask you some questions in order to learn more about your opinions on Chinese as a second language CL2 reading, and more specifically, about the GCE 1162 reading examination.

Segment C: Motivation

我相信通过对您和其他教育工作者的采访，我们对 GCE 1162 设计的优缺点会有更透彻的分析。您所提出的意见、问题和建议将有助于我们知悉华文阅读与测试在新加坡未来的发展方向。

The data gathered through interviews with you and other stakeholders will be used to evaluate the GCE 1162 reading examination. Your views are extremely valuable in helping me understand what is important in reading assessment in the Singapore context.

Segment D: Confidentiality issues and time line

这次的访谈仅为学术研究用途。您的身份和谈话内容将会被严格保密，文章中如果引用任何实际例子，一定匿名处理。访谈预计一小时到一个半小时左右。

All of your information and responses will be kept confidential and used only for academic purposes. If any concrete example is provided during the interview, it would not be specifically quoted. The interview should take about 60 to 90 minutes.

(**Transition:** 在进入正题之前，我们先来聊聊您的工作经验。Let me begin by asking you some questions about your work experience.)

II. Body

Segment A: Work/schooling experience

Questions for interviewees other than students

Q1. 请问您现在在哪里工作？主要负责的项目是什么？

Where are you working? Could you tell me more about your job?

Q2. 可否请您简单叙述一下您过去的工作经验？

Could you describe briefly your previous work experience?

Q3. 请问您曾参与国家级考试的设计或执行工作么？请说明工作内容。

Have you been involved in the design and execution of any national examination?

Could you elaborate on the nature of the work involved?

Questions for student interviewees

Q1. 请问你现在在哪里念书？可否简单叙述一下你过去学习华文的经验？

Where are you studying? Could you tell me more about your experience of learning Chinese?

Segment B: The reading construct

Q1. 华文作为第二语文的新加坡考生具有什么特点？有哪些特点是我们在设计 1162 试卷时所必须注意的？

Are there any unique characteristics of Singapore's CL2 readers that have to be taken into consideration when designing the GCE 1162 reading examination?

Q2. 您认为我们的学生在完成了中小学教育后须具备什么样的华文阅读能力和思维能力？

In your opinion, what is the level of CL2 reading proficiency a student is expected to attain after completing secondary education? What about a student's cognitive ability?

Segment C: Cognitive parameters of the GCE 1162 reading examination

Q1. 可否请您谈谈您对 1162 阅读试卷的总体印象？Could you tell me about your general impression of the GCE 1162 reading examination paper?

Q2. 您认为 1162 阅读试卷的考试目标为何？

In your opinion, what are the assessment objectives of the GCE 1162 reading paper?

Q3. 您认为 1162 试卷的成绩有哪些用途？

In your opinion, what are the purposes of the GCE 1162 reading paper?

Q4. 您觉得这份试卷考核了什么样的阅读技能和思维能力？您可以按项目逐步分析（综合填空、阅读理解一选择题、阅读理解二简答题）。

What are the reading skills and cognitive processes involved in answering the GCE 1162 reading paper? You may wish to look at the different sections in turn, namely, multiple-choice cloze, reading comprehension multiple-choice and reading comprehension constructed response.

Q5. 除了阅读技能和思维能力，这份试卷还考核了什么要素？

Besides reading and cognitive skills, does the GCE 1162 reading paper elicit other aspects of learning?

Q6. 考试实际测量的构念与《中学华文课程标准》和《考试纲要》中列出的目标是否契合？为什么？

Does the GCE 1162 reading examination paper measure what it proposes to measure? Why?

Q7. 您在较早前指出新加坡学生在完成中小学教育后所应该具备的华文阅读能力和思维能力。您认为 1162 阅读试卷能否有效地测量这些要素？为什么？

How effective is the examination paper in eliciting the reading and cognitive skills which, as you mentioned earlier, are critical to a student after completing secondary education?

Q8. 您认为 1162 试卷能否有效地区分出读者的优劣？为什么？

Is the GCE 1162 reading examination able to differentiate between the competent and less competent reader? Why?

Segment D: Contextual parameters of GCE 1162 reading examination

Q1. 1162 阅读试卷的题型有三种，即综合填空、选择题和简答题。您认为这些题型是否真实有效？

The GCE 1162 reading examination has three item formats, namely, multiple-choice cloze; reading comprehension multiple-choice and reading comprehension constructed response. What do you think of these item formats? Are they effective and authentic?

Q2. 您可否举出其他考查学生阅读能力和兴趣的题型或方式？

Can you think of other types of items or ways of assessing reading ability and interest?

Q3. 您所列举的这些题型和考核方式可否包括在中学生阅读能力的终结性评估中？倘若可以，应该如何融入？

Can the existing GCE 1162 reading examination include a wider range of item formats and assessment methods? How can they be incorporated?

Q4. 有的老师认为，应该恢复过去阅读试卷中填写汉字和造句的题型。对此您有何看法？

Some teachers recommend the revival of item formats used in older versions of the GCE 1162 reading examination, such as filling in the Chinese character and sentence construction. What is your opinion on this?

Q5. 您认为考评局根据什么标准选择考试篇章？

What do you think the selection criteria for the GCE 1162 reading passages are?

Q6. 您觉得篇章的数目、长度和类型一般是否适合？为什么？

Are the number, length and genre of the passages generally appropriate? Why?

Q7. 您觉得题目的顺序和权重一般是否合适？为什么？

Are the order and weightage of the items generally appropriate? Why?

Q8. 您认为一个半小时的作答时间是否足够？为什么？

The duration of the GCE 1162 reading paper is one and a half hours. Do you think the time given is sufficient? Why?

Segment E: Evaluation

Q1. 您认为可以通过什么方式提升 1162 的质量？

What improvements can be made to enhance the quality of the GCE 1162 reading examination paper?

Q2. 您认为维持或提高一套试题的质量是否需要持续性的监督与审查？在这个过程中，考评局、教育部、教育学院、学校和家长又能扮演什么样的角色？

Are ongoing evaluation and validation needed to ensure the quality of an examination?
What roles can the different actors, e.g. the Singapore Examinations and Assessment Board (SEAB), Ministry of Education (MOE), National Institute of Education, schools and parents play in these evaluation and validation processes?

(**Transition:** 非常感谢您的宝贵意见。让我在结束访问前简单地总结一下这次访谈的要点。Thank you for your invaluable inputs. Let me briefly summarize the main points of our interview.)

III. Closing

Segment A: Summarize

您提到以下几点，我觉得非常关键.....

Here are some salient points that you have raised.

Segment B: Maintain rapport

请问您还有什么要补充的？或是这次访谈有什么可以改进的？

Is there anything else you would like to add? Or do you have any suggestions for improving this interview?

再次感谢您在百忙之中抽空接受我的访问。

Once again, thank you for your time and willingness to speak with me.

Segment C: Action to be taken

如果在转写的过程中有疑问的话，可否再向您请教？

I should have all the information I need. May I contact you if I have any doubts or questions?

谢谢！再见！我们保持联系！

Thank you and keep in touch!

Appendix H

Excerpts of raw data transcribed and coded using NVivo10

Transcript_Gamma

03....

Quick Access

- Files
- Memos
- Nodes

Data

- Files
- 01. Alph
- 02. Beta
- 03. Gam
- 04. Delt
- 05. Epsil
- 06. Zeta
- 07. Eta_
- 08. Thet
- 09. Iota_
- 10. Kap
- 11. Lam
- 12. Mu_
- 13. Nu_
- 14. Xi_C
- 15. Omi
- 16. Pi_W
- 17. Rho_

needed something in between for their students. And it looks like it's a little bit easier than the first language Chinese paper. It is interesting to some extent, that as the standard gets harder with the reading, ...to some extent, it may be generalisation, but to some extent, the content seems to become a bit more traditional. ^^^ I was aware, looking through one of the papers, I was comparing 1160 to 1162, you know the new syllabus, and thinking presumably 1160 is a little bit shorter some of the passages aren't they but 1162, the paper that I have got on my machine here, there is certainly the traditional picture of a dragon.

Comment [C16]: Choice of passages, gets more traditional with increase in level of difficulty.

Comment [C17]: New and old reading papers

Comment [C18]: Traditional components

03:19 I: They tried to introduce some authentic components into it, so the authentic components are mainly in the form of advertisements...

03:23 G: Yes.

03:23 I: and newspaper articles, so they are shorter and less...traditional

Coding Density

- T-MOTIV
- COGNI
- S-PURPOLICY
- S-COMMUN
- S-COGNI
- P-STN
- X-COMMOTH

In Nodes Code At Enter node name (CTRL+Q)

Quick Access

- Files
- Memos
- Nodes

Data

- Files
- 01. Alph
- 02. Beta
- 03. Gam
- 04. Delt
- 05. Epsil
- 06. Zeta
- 07. Eta_
- 08. Thet
- 09. Iota_
- 10. Kap
- 11. Lam
- 12. Mu_
- 13. Nu_
- 14. Xi_C
- 15. Omi
- 16. Pi_W
- 17. Rho_

Search Project

Nodes

Name	Files	References
COGNI	14	128
CONTEXT	8	20
PURPOLICY	16	201
SCORE	14	52
SUGGEST	0	0
TAKER	11	63
WBASH	2	3
WORK	17	60

Drag selection here to code to a new node

Transcript_Gamma COGNI PURPOLICY

<Files\01. Alpha 23122014\1.Transcript Alpha> - 5 7 references coded [1.12% Coverage]

Reference 1 - 0.05% Coverage

I: 我是说现在的考试方式还是非常地传统，对吗？ A: 嗯嗯

Reference 2 - 0.04% Coverage

所以考评局对我来说是一个很神秘的机构，只是去

Reference 3 - 0.08% Coverage

那边改，而改我也只是改我那一题。我也只跟我那一组，而且又不准这个，不准那个，所以非常非常地局限！

Reference 4 - 0.34% Coverage

I: Erh,可以不可以从这个角度来想呢？可能我们在metacognition方面, our thinking is good but the thinking about thinking part not so familiar, 我们不知道怎么去整理，我们这套到底是怎么运作的。没有透明嘛，所以很少去跟人家讲，我们背后的结构是怎样的。就是说 ok we have a good system, 这个制度是好的，

Reference 5 - 0.08% Coverage

Section 1 of 2

Appendix I

Excerpt from Excel evaluation spreadsheet used for expert judgement

File Home Insert Page Layout Formulas Data Review View Add-Ins

Normal Page Layout Page Break Preview Custom Views Full Screen

Workbook Views

Ruler Formula Bar Gridlines Headings

Show

Zoom 100% Zoom to Selection

New Window Arrange All Freeze Panes Unhide

Split Hide View Side by Side Synchronous Scrolling Reset Window Position Window

Save Workspace Switch Windows

Macros

C3 Strength and overcoming depression

	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	Theme		Genre		Literary value	Q11					Compre 1B	Theme		Genre		Literary value	Q12					Q13				
2						Local/Global	Expeditions/Careful	Cognitive	C-specific								Local/Global	Expeditions/Careful	Cognitive	C-specific		Local/Global	Expeditions/Careful	Cognitive	C-specific	
	Local news and culture		Functional		Not applicable/Low	Global	Careful	Lower	L: Understand	Summarise gist and intention	A lady who likes singing	Aesthetic appreciation		Narrative		Not applicable/Low	Global	Careful	Lower	L: Understand	Infer	Global	Careful	Lower	L: Understand	Infer
3	Local news and culture		Functional	Newspaper article	Not applicable/Low	Global	Careful	Lower	L: Understand	Summarise gist	a lady whose life revolved around music	Aesthetic appreciation		Narrative		Not applicable/Low	Global	Careful	Lower	L: Understand	Summarise gist	Global	Careful	Lower	L: Understand	Identify information interpret
4	Local news and culture		Functional	Newspaper article	0	Global	Careful	Lower	L: Understand	Summarise gist and intention	A lady whose life revolves around music	Aesthetic appreciation		Narrative		0	Global	Careful	Lower	L: Understand	Infer	Global	Careful	Lower	L: Understand	Infer
5	Local news and culture		Functional	Newspaper article	0	Local	Careful	Lower	L: Understand	Identify information and paraphrase	Helping others (Violinist who gives his violin to a pauper)	Values and attitudes		Narrative		Not applicable/Low	Local	Expeditions	Lower	L: Remember	Identify information (repeat key phrases in answer)	Global	Careful	Lower	L: Understand	Understand and infer nature of character (more than one correct answer)