THE RELATIONSHIP BETWEEN TEST TAKERS' PERFORMANCE ON THE TEM4 AND THEIR KNOWLEDGE OF THE RELEASED TEST SPECIFICATIONS

BY

XIAOWAN ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in the Teaching of English as a Second Language
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Master's Committee:

Visiting Assistant Professor Scott Walters, Chair
Professor Emeritus Fred Davidson

ABSTRACT

The role that released test specifications can play during test preparation is often neglected by test takers, and even researchers. Focusing on the Test for English Majors-Band 4 (TEM4) in a Chinese EFL setting, this paper investigates the preparation effects associated with the use of TEM4 Syllabus, or its released specifications. Data collection involved 48 test takers of the TEM4 recruited from a large university in central China, where the experimental group was given a tutorial session on the TEM4 Syllabus as the treatment. Specifically, the study measured the effects associated with the TEM4 Syllabus by using a quantitative metric of score improvement and a qualitative metric informed by a framework adapted from the work of Messick (1982) and Xie (2013). Along with its exploration of possible preparation effects, this paper also discusses the ethicality of different test preparation methods and touches on the issue of specification releasability (Davidson, 2012).

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Chinese people have been using tests to make decisions since ancient times. Most high-stakes tests in China, such as the College Entrance Test, are norm-referenced and intended to spread out students along a scoring scale. Since it is often the case that little or no test guidance is provided for these norm-referenced tests (NRTs), Chinese students tend to perceive test preparation as doing sample tests, published tests of previous years, or practice tests. Standardized criterion-referenced tests (CRTs), with the Test for English Majors (TEM) being a representative, emerged rather recently in the testing market of China and account for a rather small share of high-stakes tests. Because this fresh shift in the constitution of the testing market was not accompanied by a similar shift in testing culture, the historically prevalent norm-referenced paradigm has extended its power to the new territory of CRTs, and as a result, students naturally carry over their old habits of preparing for NRTs into their preparation for CRTs. Specifically, they often ignore the guidebook on a CRT and devote most of their time to practicing sample items. However, is such an attitude that Chinese students have towards CRT guidebooks reasonable? Should their test preparation behavior be advocated? And how can test takers' performance during test preparation become different if their attention is intentionally directed to the guidebook of a CRT? This paper will look into these questions by focusing on one of the high-stakes CRTs in China—the Test for English Majors-Band 4 (TEM4).

CHAPTER 2: LITERATURE REVIEW

The Difference between CRM and NRM

The distinctions between criterion-referenced measurement (CRM) and norm-referenced measurement (NRM) were first clarified by Glaser and Klaus in 1962 (as cited in Davidson & Lynch, 2002). In the following year, Glaser (1963) identified two types of information that an achievement test can provide differentiated by “the standard used as a reference” (p. 6). CRM, according to Glaser (1963), takes “the desired performance at any specified level” (p. 6) as its reference standard, whereas NRM draws attention not so much to what a student can/cannot do as to his/her “relative standing along the continuum of attainment” (p. 6).

Over time, the definitions of NRM and CRM have evolved and have been enriched with the interpretations contributed by different researchers. In order to maximally contrast the two measurement paradigms, only the most conventional views will be presented in the discussion below, but readers should be aware that many of these aspects that were historically believed to distinguish between NRM and CRM, according to some scholars (e.g., Davidson & Lynch, 2002), no longer constitute a clear border at the present time.

To summarize traditional beliefs about CRM and NRM, they differ in terms of their reference standards, score associated meanings, purposes of score use, and test construction principles. At the most fundamental level, CRM and NRM rely on different reference standards for score interpretation: for NRTs, “test results are interpreted with reference to the performance of a given group” (Bachman, 1990, p. 72), while for CRTs, test results are interpreted “with reference to a criterion level of ability or domain of content” (Bachman, 1990, p. 74). Test scores of NRTs and CRTs, as interpreted against different reference standards, convey contrastive meanings: we typically obtain from NRTs the relative ranking of a test taker without knowing his/her actual capability in a particular domain, whereas with CRTs, we typically trade rankings for more absolute information about the level of ability/the degree of mastery a test taker has regarding a particular domain of knowledge or skills. The divergent score interpretations attainable from NRTs and CRTs then qualify their use in practice. Typically, we use the score distribution (often bell-shaped) of NRTs for selecting the best portion of candidates from a pool; in contrast, cut-off scores are often set for CRTs with reference to an expected ability level regardless of the percentage of pass or fail. In order to optimize test use, testers develop NRTs and CRTs following distinct test construction principles. Variability in test scores is especially important to an NRT, and is usually guaranteed by selecting items that adequately and appropriately discriminate test takers. The construction of a CRT, on the contrary, takes little account of score variability; items are often chosen subsequent to a meticulous task sampling and analysis to ensure high degree of representativeness of the target domain or certain levels of ability.

Besides what have been stated above, another important aspect that historically discriminated between an NRT and a CRT, and that is particularly relevant to this paper, is a descriptive "document" that guides the test development process—test specifications ("test specs"). Although test specs had been facilitating the development of NRTs long before the onset of CRTs, it was CRM that elevated the primacy of specs to the height where precision of score interpretation was assured (Popham, 1978). Since the main research questions of this study were formed around released test specs, next section will be devoted to discussion of the definition, the format, and the importance of test specs.

Test Specs and CRM

The definition of test specifications. Test specifications are not a recent invention. The earliest use of this term can be traced back to 1929 in the work of Ruch (as cited in Davidson & Lynch, 2002). Under the influence of behaviorism, Popham (1978) regarded test specs as the *descriptive theme* that provides specifics of the test items measuring a particular behavior. From a functional perspective, Davidson and Lynch (2002) defined test specs as "generative blueprints from which test items or tasks can be produced" (p. 3). Bachman and Palmer (1996) differentiated between two levels of specs, namely *test task specifications* and specifications for a complete test, or a *blueprint*. Though not the exact same term has always been used in the literature, the notion of test specs is quite well established. To avoid ambiguity, here and in the rest of this paper I will stick to the term of *test specs* and apply it to any levels of descriptive language that guides the development of a test.

Formats of test specifications. Test specs tend to vary in the formats they take and the levels of detail they provide. The format adopted by Davidson and Lynch (2002) in their illustration of test specs was slightly adapted from the classical Popham-style specification. This particular format of test specs is featured by its five components: the general description (GD), the prompt attributes section (PA), the response attributes section (RA), the sample item (SI), and the specification supplement (SS). A complete spec is led by the GD that gives a brief summary of what is to be assessed in the test. The GD may also include "a statement of purpose, the reason or motivation for assessing the particular skills" (Davidson & Lynch, 2002, p. 21). The PA and the RA closely follow the GD. While the PA clarifies "what will be given to the test taker", the RA describes "what should happen when the test taker responds to the given" (Davidson & Lynch, 2002, p. 25). The SI component is quite self-explanatory, as it concretizes and illustrates abstract descriptions by providing sample items and tasks. The last component—the SS—is optional to a spec; it supplements the information included in the previous four components to make a spec more complete and organized.

One alternative test spec format can be found in Bachman and Palmer (1996). As mentioned earlier, they used the term *blueprint* to refer to a complete set of test specs, which is composed of *test task specifications* and the *structure of the test*. The typical components of a test-task spec are the purpose of the test task, the definition of the construct to be measured, the characteristics of the setting of the test task, time allotment, instructions for responding to the task, characteristics of and relationship between input and response, and scoring method. Multiple test-task specs are then assembled into the final blueprint consistent with the structure of the test.

Alderson, Clapham, and Wall (1995) approached test specs from several different perspectives. By putting themselves into the shoes of different readers, they distinguished among at least three versions of test specs, with each of them targeted at a particular audience. In their opinion, specs for test writers, test validators, and test users should vary in both contents and structures. The version of a spec that they proposed for testers contains a general statement of purpose, test battery, time allotted, test focus, source of text, test tasks, item types, and rubrics. Alderson et al. (1995) also emphasized the importance of having adequate test information available to test candidates through publishing partial test specs.

> The intention of such specifications for candidates should be ensured that as far as possible, and as far as is consistent with test security, candidates are given enough information to enable them to perform to the best of their ability.
>
> (Alderson et al., 1995, p. 21)

This quote presages an important issue in language testing—test spec releasability (Davidson, 2012a), which will be addressed later in the section of *test preparation and releasability*.

The importance of test specs to a CRT. In the comparison between CRM and NRM, it was implied that CRM tended to utilize the asset of test specs more thoroughly in contrast to NRM. Though this no longer holds true as the boundary between CRM and NRM becomes fuzzy, it would still be helpful to examine the importance of test specs to CRM from a historical point of view if we want to understand why and how test specs might frame CRM as an alternative form of educational measurement to NRM.

The most classic, and probably also the most appealing, feature of test specs resides in their generative and controlling power. Serving as the blueprint of a test, test specs spell out "the nuts and bolts of how to phrase the test items, how to structure the test layout, how to locate the passages, and how to make a host of difficult choices as we prepare test material" (Fulcher & Davidson, 2007, p. 52). These detailed instructions, on the one hand, allow test specs to act as an assembly line to consistently generate equivalent test items or tasks (Davidson & Lynch, 2002; Davidson, 2012b; Davidson, 2013; Fulcher & Davidson, 2007; Popham, 1978); and on the other hand, they strictly control the test development process and ensure test reliability through an enhanced level of standardization (Davidson & Lynch, 2002; Moss, 1994).

Nonetheless, the primary goal of test specs in CRM is not to reproduce or control but to describe. Educational reformers advocate the replacement of NRM with CRM chiefly for the heightened descriptive power that test specs possess. The *descriptive theme* incarnated in the test specs of a CRT, according to Popham (1978), essentially distinguishes the test from a traditional NRT in that the test taker's performances and test scores are associated with a lucid and precise meaning. In the old era dominated by behaviorism, when language learning was defined in terms of rote-learned behaviors, the emergence of CRM, under the help of test specs, provided explicit definitions for the behaviors being tested and addressed the defects prevalent in NRM by informing teachers and other potential test users of what students can/cannot do, as well as how effective an education program has been in improving students' performance (Popham, 1978).

Test specs help CRTs to realize their potential in providing precise score interpretation mainly through two functions, as illustrated below in Figure 1 (summarized from Popham, 1978). While the connection between test specs and test items communicate to test writers specific rules to be followed in developing a test, the other connection between test specs and the *test criterion* (by which Popham referred to a "well-defined behavioral domain" (p. 94)) communicate to test users with high clarity the domain of behaviors to be assessed in a test:

Test Items—Test Specs—Target Behavioral Domain

Figure 1. The relationship between test specs, test items, and criterion

At the present time, with the role of specs shifted from defining a behavioral domain to operationalizing an underlying construct, the figure could be modified in the following way:

Test Items—Test Specs—Target Domain of Knowledge, Skills, or Processes

Figure 2. The relationship between test specs, test items, and construct

Davidson (2012b) echoes Popham's (1978) stance in his nomination of test specs as the chief legacy of CRM. Given the importance of test specs to CRM, it is not surprising that they play a more or less central role in the frameworks proposed for the development of a CRT (e.g. Davidson & Lynch, 2002; Hudson & Lynch, 1984; Lynch & Davidson, 1994; Mislevy, Sternberg & Almond, 2003).

Over time, the refinement of CRM models has extended test specs' descriptive power to a more fundamental dimension of testing—validity (e.g. Davidson, 2012b; Davidson, 2013;

Fulcher & Davidson, 2007; Lynch & Davidson, 1997). Because test specs compel testers to articulate every aspect of a test, for example, the purpose of it, the conceptualization of the target construct, the operationalization of the target construct (including test format, task configurations, and scoring procedures, etc.), and all the underlying rationales, they help to achieve the ultimate clarity of a test by demanding fine-grained validity evidence. Moreover, specs grow and change with the influx of new data, theories, feedback and discussions, funding situations, and so on (Fulcher & Davidson, 2007). Along with the evolution of test specs grows the understanding of the construct, as well as the relationship between teaching, learning, assessment and the real-world context. Chapelle, Enright, and Jamieson (2008) provided a good demonstration of the evolution of test specs in the context of the TOEFL development. Li (2006), on the other hand, illustrated with the conceptual diagram below (Figure 3) how changes to different versions of specs could be tracked through audit trails and thereby serve as valuable validity evidence. We can see that test specs evolve until they are deliverable by incorporating feedback data from various sources, and it is from this iterative process that we obtain enhanced validity. The relationship of specs to validity is an ongoing exploration; more research is needed to further disclose how specs, preferably of varied levels of control, can affect validity in different measurement contexts.

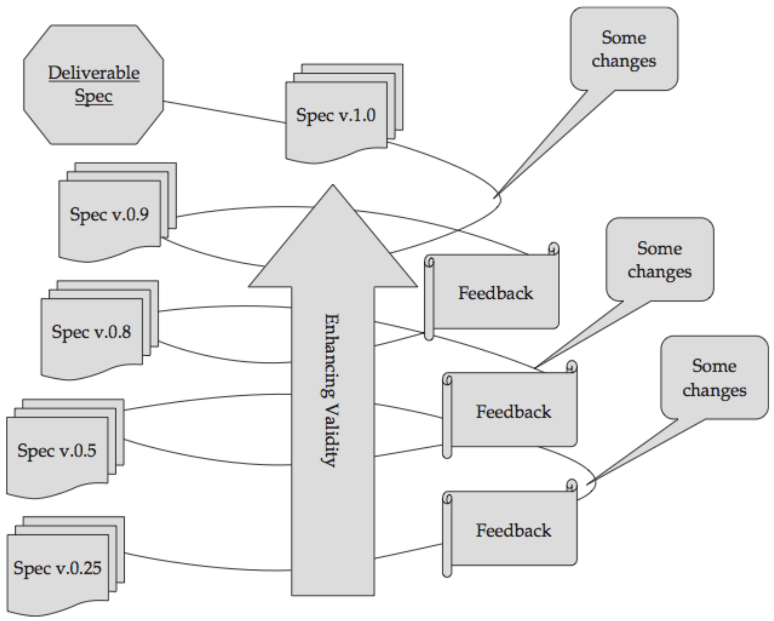


Figure 3. How enhanced validity is achieved through the evolution of test specs.

(Li, 2006, p. 20)

Test Preparation and Releasability

An overview of test preparation. Test preparation is an old topic in the realm of testing. Following the inclusive definition of "coaching" given by Messick (1982), *test preparation* will be used in this study as an umbrella term to refer to "any intervention procedure specifically undertaken to improve test scores" (p. 70). Briggs (2009) identified three key elements that define different types of test preparation behaviors—content review, item practice, and development of testwiseness (p. 11); he also made finer distinctions between "student-driven" or informal test preparation methods and instructor-led "coaching" courses or programs that fall towards the formal end of the continuum. This paper will focus mainly on the informal test preparation methods, though the literature review will cover studies of test preparation at both formality levels.

Researchers have approached test preparation from various perspectives. The two most frequently visited questions are:

1. Does test preparation work?
2. Is test preparation instrumental in guiding teaching and learning?

Around these two questions have developed two major lines of research: one strand attempts to quantify the effects of test preparation on test performance, while the other strand investigates washback, or the impact of a test on teaching and learning (Alderson & Wall, 1993), through revealing various types of teaching and learning activities engaged in the test preparation stage.

Within the first line of research, where the product of test preparation rather than its process is emphasized, a large number of studies have focused on the Scholastic Aptitude Test (SAT), a standardized aptitude test widely used for college admissions in the US (e.g. Briggs, 2009; Messick, 1982; Powers & Rock, 1999). The effects of coaching on the SAT scores, as most prior studies invariably indicated, are relatively small, especially in light of the standard error of measurement of the test. However, controversies have arisen about the influence that score inflation would wield over college admissions decisions. The older literature, represented by Messick (1982), tends to deem the effects as negligible on the basis of reasoning and speculation, whereas more recent literature, based on the interview data of selective postsecondary institutions, believes that chances for a student to be admitted will increase even for a rather small score improvement (e.g. Briggs, 2009). Considering the limitations of the previous studies (such as a small sample size, absence of a control group, and self-selection bias) and their inconclusive findings, more exploration is needed to better understand the relationship between test preparation and the SAT scores.

In contrast, the second line of research focuses on educational improvement in general rather than the degree of score inflation that might result from test preparation. Washback researchers are interested in revealing and analyzing the practices employed in test preparation, on the top of which they often evaluate different practices from an educational perspective by relating them to test impact. Studies in this strand usually investigate regional or national tests, taking teaching methodologies, contents of instruction, preparation textbooks, and learning strategies as their focus (e.g. Smith, 1991; Wall & Alderson, 1993; Wang, Yan & Liu, 2014; Xie, 2013). A few similar studies are also found on internationally administered tests, such as the TOEFL (e.g. Alderson & Hamp-Lyons, 1996; Hamp-Lyons, 1998; Wall & Horak, 2011) and the IELTS (e.g. Badger & Yan, 2012; Mickan & Motteram, 2009; Read & Hayes, 2003), with a large proportion of them funded by test companies in an attempt to improve their product quality. Compared with effect studies that adopt purely quantitative methods, washback studies inspect test preparation at a fine-grained level through close observations and interviews. Along with their detailed reports of findings are often remarks on the desirability of the test washback, and in some cases, explanations are provided as to why certain types of washback, either positive or negative, have been observed.

Several washback studies have also looked into the outcome of test preparation, though their findings remains largely inconclusive. To take the IELTS for illustration, some of the studies agreed with what has been discussed earlier about the SAT coaching and claimed that special test-focused programs did not result in significantly higher score improvement than general English learning programs (e.g. Gan, 2009; Green, 2007), whereas others (e.g. Brown, 1998) maintained that intensive classes on the IELTS were significantly more effective in enhancing test takers' performance than courses on English for Academic Purposes (EAP).

It should be pointed out that when outcome takes a role in washback studies, the target is not so much the outcome itself (whether there is any score improvement and how significant it is) as the reasons behind it, that is, the exact test preparation practices that account for the score improvement. By expanding the scope of effect-only investigations, washback studies take ethics into consideration and evaluate whether or not the score improvement has been achieved through ethically appropriate means.

Different preparation methods and an ethical issue. Messick (1982, p. 81) identified three types of "coaching" programs (or test preparation methods/behavior) and their respectively associated preparation effects:

1. Those that genuinely improve the target skills or abilities;
2. Those that eliminate the construct-irrelevant difficulties of taking a test, such as unfamiliarity with test format; and
3. Those that exploit "testing taking stratagems and answer-selection tricks".

According to Messick, the first two types of test preparation practices are beneficial to test takers' performance without doing any harm to test validity, and the second type will even contribute to enhancing validity. The third type, on the contrary, threatens the validity of test-related inferences, though it might not pose any threat to well-designed tests.

After synthesizing Messick (1982) and Haladyna and Downing (2004), Xie (2013) added a fourth type to Messick's (1982) original list, that is, practices that narrow the curriculum. This fourth type inflates test scores by teaching to the test; specifically, it involves "using test materials, following a test-based curriculum, using similar or identical test items, or focusing exactly on what the test measures" (Xie, 2013, p. 198).

Recall that Chinese students always prepare for a test by doing sample tests or published tests of previous years. Their way of preparation, apparently, falls within the second and the fourth types, and mostly the latter. Teaching to the test, or in this case, practicing to the test is not the innovation of Chinese students; it is, instead, a common phenomenon in response to the pressure of performing well in a test. Though to most test takers teaching/practicing to the test is an unquestionably acceptable and reliable method of test preparation, in the eyes of most researchers, it is ethically inappropriate, educationally indefensible, and detrimental to test validity. Mehrens and Kaminski's (1989) discussion of test preparation placed a list of preparation activities on a continuum of ethicality, where the majority of teaching-to-the-test practices, such as practicing actual test materials and identical test items, were condemned as illegitimate test preparation behavior. To complement the standard of ethical appropriateness proposed by Mehrens and Kaminski (1989), Popham (1991) suggested another standard of educational defensibility and maintained that some teaching-to-the-test practices (such as same-format preparation), though they might be ethically justifiable, are educationally indefensible, because they will not enhance the learner's mastery of the entire target domain in a balanced manner. Popham's (1991) view has been supported by Xie (2013), who argued that the fourth type merely helps learners develop their abilities in those aspects sampled by the test, with the rest of the target domain completely neglected. In terms of the relationship of test preparation to validity, researchers unanimously believe that teaching to the test compromises the validity of the inferences that test users wish to make from test performance to the target domain (Haladyna & Downing, 2004; Mehrens & Kaminski, 1989; Xie, 2013).

Students voluntarily practice to the test usually because they have convenient access to relevant materials. One empirical study on commercial test preparation materials was conducted by Hamp-Lyons in 1998. From the perspective of washback, she examined test preparation textbooks for the TOEFL PBT and found most of them unethical and educationally indefensible. Hamp-Lyons argued that while these preparation textbooks provided students with self-study based practice materials, they failed to truly help them "diagnose problem areas, patterns of need, or even areas of strengths" (p. 332).

Another pertinent study showed how test specs could be inappropriately used for test preparation. Shepard and colleagues (1987) published a lengthy report on a case study of the Texas Teacher Test, where they explicitly questioned the legitimacy of the contents of some test workshops and criticized part of their instruction as inappropriate and unethical exploitation of test specs. For instance, some study guides published by the test agencies attempted to inform test takers of how distractors were constructed in multiple-choice questions and to teach them to rule out the wrong options by using extraneous clues. The unethical usage of test specs exposed in Shepard et al. (1987), though it is not very common, has great implications for how much and in what form test specs should be released, or the issue of test spec releasability, and we will revisit this study in our later discussion.

Teaching to the test has received a great amount of attention from researchers not only because of its questionable nature, but also because of the influence it might exercise on test outcome. Messick (1982) speculated the effects of coaching would be more evident on achievement tests than on aptitude tests considering the former's relatively higher responsiveness to instruction. If Messick's prediction is correct, then teaching to the test, though it might not undermine the integrity of an aptitude test, can become a primary concern for the developers and users of achievement tests. Xie's (2013) findings regarding the College English Test-Band 4 (CET4) actually corroborated Messick's speculation in that among all the preparation practices studied, those that narrowed the curriculum, especially drilling, accounted for one third of the total preparation effects.

The releasability of test specs. Alderson et al. (1995) clarified the distinctions between a test spec and a test syllabus in their discussion of the issue of spec confidentiality. In their opinion, the two terms should not be used interchangeably because they are intended for different readers and are characterized by different levels of confidentiality with respect to content: test specs, as the document circulating within the inner circle of test developers and validators, should provide as much detail as possible, whereas test syllabi, given their broader target audience (including test candidates, test users, and practice exam publishers), should contain no confidential information. What Alderson et al. (1995) referred to as a test syllabus is the partial specs released to the public for promoting a test and is more commonly known to American testers as *released test specs*.

Davidson (2012a) devoted an entire paper to discussion of the issue of test spec releasability, which he defined as "whether a spec should be shared outside of the test development team, and if so, when and in what form." As test security and test validity both depend heavily on the releasability of specs, test developers or companies have to make sure that their competitors will not take advantage of their hard work, and that test candidates are not provided with excessive access to the test development process. The case study conducted by Shepard et al. (1987) has illustrated some negative consequences that could arise from the ignorance of this issue. That is, test takers were literally cheating on the test with the assistance of the over-released test specs.

Apparently, the publication of test specs walks a fine line between "constructive" and "destructive": adequate test information must be made available to test candidates to ensure test validity; yet with some indiscretion the released test information can lead to either vicious competitions or unethical test preparation practices that endanger fairness and validity. Davidson (2012a) underlined that "production of a releasable version of a test spec involves careful editing of the internal spec to remove details that affect study for the test, test security, and other considerations that might alter the test's validity." However, later we will see how tricky it can be to balance the functionality and confidentiality of released specs.

Released test specs, test preparation, and test outcome. As we have seen so far, at least among testing researchers, it is agreed that preparing for tests according to official guides or test syllabi should be advocated, and that to prepare only by teaching/practicing to the test should be discouraged, although school administrators, teachers, or students might not find such arguments to be persuasive. Given the amount of attention drawn to test preparation, it is surprising that not a single controlled study, to the best knowledge of this author, has been conducted in the field of language assessment to inspect the possible effects that desirable and undesirable test preparation methods can have on test performance. A chapter written by Perlman (2003), despite its attractive title—"Practice Tests and Study Guides: Do they Help? Are they Ethical? What is Ethical Test Preparation Practice?"—repeated nonetheless what has been reviewed earlier about teaching to the test and barely touched on test guides or released test specs. In the context of employment screening test, several studies have been conducted looking into the effects of providing pre-test information and preparation materials (e.g. Burn, Siers & Christiansen, 2008), but none of them shed light on the situation of educational measurement.

The foregoing literature review has exposed Chinese students' questionable perception of test preparation, underlined the importance of test specs to CRM, and presented different test preparation effects associated with various preparation behaviors, but it has also revealed a lack of empirical knowledge about the relationship of test specs to test preparation and test performance. This is important because there is a logical connection between the influential role that test specs play during test production and the role that they might play as students prepare for tests: the generative function of test specs suggests that study (of that function) might be beneficial to test-takers, provided that ethical mis-steps can be avoided.

CHAPTER 3: THIS STUDY—SCOPE, FOCUS, AND RESEARCH QUESTIONS

The focus of this study, as informed and inspired by the literature review, is the preparation effects (to be measured quantitatively by score improvements and qualitatively by behavior changes) associated with the use of released test specs. The test of interest has been conveniently chosen to be the fourth band of the TEM (TEM4), because it satisfies the preconditions for this study—it is a CRT and has released test specs and other test preparation materials (including published/practice test papers) available to the public.

While it is not possible or appropriate to control the preparation behavior of test takers in a study (i.e., to make them prepare following one method but not others), the findings might still be interesting if we can intentionally direct test takers' attention to the test syllabus of the TEM4 and observe what effects this intervention might have on their performance during test preparation. Keeping in mind the problematic nature of practicing to the test, to incorporate test syllabi into test takers' preparation experience might be the first step, and probably also the most critical one, to transform their perception of test preparation, and ultimately, to lead to a positive change in their test preparation behavior, that is, from practicing to the test to making genuine improvements in English abilities. The potential educational value associated with the use of released test specs for test preparation also constitutes the very motivation for conducting this study.

Introduction to the TEM4

The Test for English Majors, as indicated by its name, assesses the academic achievements of English major students at institutions of higher education in China (NACFLT, 2000). The TEM test battery comprises two levels, with the TEM4 targeted at sophomores and the TEM8 targeted at seniors (Jin & Fan, 2011). As a typical CRT, the TEM bases its test contents on the syllabi of the core courses taken by English majors nationwide. Ever since its first launch, the TEM has been valued by the English departments as one of the most authoritative ways to measure their teaching and learning effects. At some institutions, degree conferral is contingent on passing the TEM4 in addition to acceptable academic performance at school. Therefore, the TEM, especially the TEM4, is a high-stakes test to English majors in China. Furthermore, because the TEM is the only large-scale standardized test used particularly for evaluating English majors, companies and organizations are also inclined to see the TEM scores an important facet in their selection of English-major employee candidates.

The TEM4 Syllabus and Other Test Preparation Materials

The released specs of the TEM4, or more commonly known as the TEM4 Syllabus, are published by the National Advisory Education Committee for Foreign Language Teaching (NACFLT), which is also the committee that develops and administers the test. Congruent with our earlier discussion about the purpose of released test specs, the TEM4 Syllabus is intended to perform as the liaison between the test developer and stakeholders by communicating to the latter test formats, target skills, assessment requirements, and other test-relevant information. Apart from a preface that sketches out a general picture of the test and a coda that specifies the scoring procedure and rubric, the main body of the TEM4 Syllabus describes test tasks in the order of how the test unfolds, that is, *dictation*, *listening comprehension*, *cloze*, *grammar and vocabulary*, *reading comprehension*, and *writing*. The description of each task is further framed from four aspects—task requirements, task format, task purposes, and material-selection principles. In the appendix of the TEM4 Syllabus, one sample task is provided corresponding to each task type for illustration. The original TEM4 Syllabus is available in Chinese only, but a translated version of the main body of the TEM4 Syllabus can be found in Appendix I.

While it is reasonable to presume that the TEM4 Syllabus is conveniently available to stakeholders, especially given the vast amount of information on the Internet, the truth is that the officially published TEM4 Syllabus is extremely hard to obtain, at least according to the purchasing experience of this researcher. Here is some anecdotal evidence: To begin with, two search engines—Google and Baidu (the biggest search engine in China)—were used to search for the TEM4 Syllabus, and the results loaded on first three pages were examined. The websites that provided information on the TEM4 Syllabus, as expected, were almost all operated by English tutoring/coaching programs. However, a brief scan of their contents could easily lead to the conclusion that these online versions of the TEM4 Syllabus were by no means official, because they were either incomplete or scattered with typos. Next, three largest online book retailers in China were paid a visit, Amazon.cn, Dangdang, and Taobao. Despite the fact that all of them had this brochure in their catalog, none had it in stock. As the two most convenient paths came to a dead end, the researcher was lucky enough to find the website of the publisher based on the book information provided on Amazon.cn. The TEM4 Syllabus was finally obtained through placing an order directly with the publisher, which, obviously, is not what a student would usually do when purchasing for test preparation materials, unless he/she strongly believes that the TEM4 Syllabus can effectively help him/her improve his/her test performance.

In contrast with the TEM4 Syllabus, another major window to the test—published/practice test papers—is much more conveniently “open” to stakeholders. Still, to take the aforementioned online book retailers for an example, the top results shown upon entering the TEM4 as a keyword were a variety of test-paper choices, either published test papers, practice test papers, or a combination of the two. Usually, a packet of published test papers includes the authentic test papers that have been used for the most recent eight to ten years. Practice test papers, on the other hand, are developed by different test coaching teams based on available test-relevant information.

Assumptions and Research Questions

Given the fact of the striking difference in availability of the TEM4 Syllabus and other test preparation materials, and given test takers' typical test preparation behavior (i.e., doing published/practice test papers instead of studying the guide book), it was assumed that students' preferences have driven the sales of different test preparation materials, and further that they often ignore the TEM4 Syllabus as they prepare for the test. The survey data of students' attitudes towards the TEM4 Syllabus to be presented later, will cast some light on this assumption.

Another related assumption was that that being unfamiliar with the TEM4 Syllabus should not prevent students from knowing its information, especially in the context of this study where students had convenient access to and were very likely to rely on prior sample tests for preparation. According to Fulcher and Davidson (2007), it is possible to infer spec-level language by examining sample items using reverse engineering (RE), a mental process critical to test creation. The concept of RE has been taken further in this study to include any attempts to retrieve spec-level information by extracting and synthesizing the common characteristics of individual items, and it is believed that every normal test taker should be able, whether consciously or unconsciously, to undertake some RE when practicing sample items. In other words, the prior sample tests are a window to not only the test itself, but its test specs as well. The accuracy of RE results ought to, presumably, grow along with the increase in students' motivation for the test, analytical abilities, and the amount of time and effort that they devote to test preparation; however, RE can also be wrong due to the limited experience, knowledge, and skills that a person has in doing this activity.

Based on the two assumptions stated above, this study is designed to answer the three questions below:

1. Whether directing test candidates' attention to the TEM4 Syllabus before the test can benefit their test preparation or test performance; and
2. Whether test candidates are able to acquire the meta-knowledge of the TEM4 Syllabus from other resources, such as the published tests of previous years; and
3. Whether the more accurate meta-knowledge of the TEM4 Syllabus can result in better test performance conditionally on other relevant variables.

CHAPTER 4: METHODOLOGY

This study looked into the role played by the released test specs of the TEM4 (or the TEM4 Syllabus) during test preparation by conducting quasi-experiments, survey studies, and statistical analysis of empirical data.

Participants

The participants of this research were 48 English majors from a university in central China. All the participants were sophomores and eligible TEM4 takers who were scheduled to take the test in March 2015, one month after they participated in this research. The original plan of conducting a randomized controlled trial proved to be infeasible due to the factors beyond the investigator's control, such as classroom availability, the program's curriculum schedule, and individual students' preferences. The revised plan was a quasi-experiment where two of the five equivalent classes at Year 2 were randomly assigned into the experimental camp, and the rest three naturally formed the control camp. Without any pre-knowledge about to which treatment condition they were going to be assigned, 18 students in the experimental camp self-selected to constitute the experimental group, whereas 30 students in the other camp voluntarily became the control group. Participation was initiated with volunteers' signing the informed consent form (see Appendix II) and giving consent to have their GPA and TEM4 scores linked with their survey responses.

Because, as mentioned earlier, no external control was imposed on the participants' preparation behavior, the control and the experimental groups did not differ in terms of their access to test preparation materials (including the TEM4 Syllabus, published/practice test papers, and etc.), but rather in terms of the treatment given in this study (i.e., an attention-directing tutorial that will be described later in greater detail). Although the trial carried out was controlled, it was not strictly randomized, and later we will see to what extent this defective randomization compromises our data analysis and interpretation of the results.

Data Collection Procedures

The entire process of data collection was divided into three phases. The first round of data collection started approximately one month before the participants took the TEM4, when both the control and the experimental groups were surveyed on their test preparation plans and progress. The experimental group was also required to sit through a 30-minute tutorial session on the TEM4 Syllabus after they completed the survey. Three weeks later, the participants were given a second survey that measured their meta-knowledge of the TEM4 Syllabus. The third round of data collection involved a third survey that was administered exclusively to the experimental group for their feedback on the tutorial session after they were informed of their TEM4 scores. Moreover, the participants' cumulative GPA (until the first semester of Year 2) and their TEM4 scores were requested from the Office of Undergraduate Affairs at this point in time. Because some participants withdrew from the research before the delivery of the second and the third surveys, in total 18, 17, and 7 valid cases were kept for the experimental group respectively for Survey I, II, and III, and 30 and 24 valid cases were kept for the control group respectively for Survey I and II.

While the first and the third surveys are quite straightforward in terms of their purposes and designs, the tutorial session and the second survey require more explanations. The tutorial session, as the treatment of this study, was intended to perform two missions at one time—to familiarize the experimental group with important test information included in the TEM4 Syllabus and to direct their attention to the role that it might play in their test preparation. To maximally keep students engaged and monitor their learning process, the tutorial was not provided in the form of a student-centered reading session but a teacher-centered lecture session; moreover, in response to the spec-level information attainable from RE activities, a special design was adopted to optimize the usefulness of the tutorial for the experimental group. The material (i.e., a 17-slide PowerPoint document) used for the tutorial was not a copy of the original TEM4 Syllabus but an enhanced version based on a pre-tutorial analysis of the TEM4 Syllabus. The purpose of recreating the TEM4 Syllabus was to avoid presenting information irrelevant to test preparation or repeating information that could be easily obtained from RE. Specifically, the enhanced TEM4 Syllabus was composed according to two principles: (a) to exclude the information that only marginally helps test preparation; and (b) to highlight the information that can hardly be drawn from the published/practice test papers through RE.

As far as the first principle is concerned, not all the statements in the TEM4 Syllabus are relevant to test development, that is, generative in nature and beneficial to test preparation. Some descriptions, while pertinent to the test in general, have little to do with the process of test development. An example of excluded information is the statement as follows: “This test is administered once a year for English major students in their fourth semester of learning.”

The implementation of the second principle is slightly more complicated. RE and a comparison between RE results and the TEM4 Syllabus were carried out by the researcher simultaneously. With the RE, the researcher tried to work out generative spec language by mentally processing the published test papers of the most recent five years; and by doing the comparison, the accuracy of the researcher’s RE specs was determined. The researcher’s judgments about accuracy levels and the efforts involved in retrieving accurate information were then used complementarily to underline syllabus statements that were relatively hard to be inferred from RE. Key words in those underlined statements were later highlighted in the tutorial.

For instance, because specifications on text selection could not be as effectively restored as those on item design, all text selection principles have been underlined.

During the tutorial, students' attention was intentionally drawn to the highlighted information, with explanations given where there was a question. Preparation suggestions were also provided along the way to encourage students to utilize the tutorial to improve their test preparation experience.

Survey II was used as an outcome measure to assess the experimental group's learning results from the tutorial as well as the results of the control group's RE process. Recall that students in the control group were assumed to be able to access the TEM4 specs with the help of RE and sample test papers. Corresponding to the tutorial material, questions in Survey II were designed to assess students' knowledge of the highlighted information, or the information that was believed to be able to maximally distinguish between the control and the experimental group, considering the fundamental difference in their information sources.

Operational Definitions of Released Specs (for This Research)

Three levels of released specs have surfaced based on our discussion so far. The first level is the original version of release specs, or the TEM4 Syllabus. This level of released specs represents the official understanding of the issue of spec releasability. The information and how it is structured in the TEM4 Syllabus convey what the test committee believes to be necessary and sufficient to publicize the test. Specifically, with test security and validity being two prerequisites, this official version of released specs is considered to be able to (a) present in layman terms to test sponsor or supervisor (i.e., relevant governmental departments) the justifications for having this test, (b) inform potential test users (i.e., teachers, administrative staffs, and employers) and test takers (i.e., English majors) of explicit score meanings and intended score uses, and (c) help elicit best performance in test takers by providing enough test information.

The second level is the unofficial specs worked out by test takers through RE. This level of released specs is chiefly inferred from published test papers. It should be noted here that RE specs are very likely to have a broader range of information than the TEM4 Syllabus, although the accuracy of RE specs is not always guaranteed. Because RE specs might contain critical information that is not supposed to be shared outside of the test committee, they pose potential threats to test validity.

The third level is the enhanced version of released specs, or the condensed TEM4 Syllabus, which embodied the investigator's understanding of a more useful version of released test specs for test preparation. This level of released specs was a recreation of the TEM4 Syllabus to optimize its effects as a treatment factor, after taking account of the noise inside the TEM4 Syllabus itself and that might be introduced by RE specs—descriptions irrelevant to test preparation have been omitted and discrepant information from the RE results has been highlighted. One caveat of the use of the enhanced released specs is that the original version can be unfaithfully represented due to the researcher's personal interpretation.

To incorporate the three levels of released specs into the purpose of this study, the interest was not a comparison of preparation effects between having access to the TEM4 Syllabus (experimental) and having no access to the TEM4 Syllabus (control), but rather a comparison between having access to enhanced released test specs (experimental) and having no access to enhanced released test specs (control). In addition, both control and experimental groups had access to and were extremely likely to use the very same convenient preparation materials (i.e., published/practice test papers) and thus were equally exposed to the benefits and dangers of RE.

Later after the findings of this study are presented, we will have some more detailed discussion about the interesting interaction between test preparation and the three levels of released test specs.

Instrumentation

Test preparation questionnaire. The questionnaire used in the first survey (see Appendix III) consisted of eight questions and aimed to collect some basic information relevant to the TEM4 preparation. The first two questions asked about the test preparation methods employed by the participants. The next two questions (with one sub-question under Question 3) were to find out whether the participants had read the TEM4 Syllabus. Question 5 and 6 enquired about the students' test preparation schedules, that is, the amount of time that they spent on test preparation in and out of class. The last two questions required the participants to evaluate their motivation for preparing for the TEM4, as well as the importance that they perceived the TEM4 scores would be to their future along a Likert scale of 1-5. The whole questionnaire was provided in both English and Chinese for clarity reasons.

Meta-knowledge questionnaire. The questionnaire used in the second survey (see Appendix IV) contained six sections and 17 multiple-choice questions. Following the structure of the TEM4 Syllabus, the six sections were presented in the order of *general questions, questions on diction, questions on listening comprehension, questions on cloze, questions on reading comprehension, and questions on writing*. Respectively, each section had two, four, four, one, three, and three questions. This questionnaire was also provided in both English and Chinese, but unlike the first one, it was graded by the researcher. Each question in the second survey accounted for one point, and the participants only scored when they provided the correct answer to a question. The maximum possible score was 17 and the minimum was zero, in one-point increments; the participants would not be punished for getting any answers wrong.

Post-test feedback questionnaire. The questionnaire used in the third survey (see Appendix V) contained four subjective questions. Target survey takers, or members of the experimental group, were asked to specify whether and in what aspects the tutorial was helpful to their test preparation. The students were also asked to evaluate the usefulness of the tutorial in improving their test scores, upon comparison with other test preparation materials. This questionnaire was also prepared in both languages, and students were allowed to provide answers in Chinese.

The TEM4 tutorial session. Using a 17-slide PowerPoint document (see Appendix VI for some screenshots of the document), the researcher presented the condensed TEM4 Syllabus during the tutorial session and briefly covered each of the six test components, namely, dictation, listening comprehension, cloze, grammar and vocabulary, reading comprehension, and writing. The session provided the attendees with information including scoring standards (for the subjective sections), material selection principles (for the objective sections), test formats (for all sections), and the required knowledge and abilities (for all sections), as well as a hard copy of the sample items given in the original TEM4 Syllabus. The entire tutorial was delivered in Mandarin Chinese for achieving optimal learning effects. To maximally reduce treatment-crossovers, no tutorial attendees were allowed to copy the PowerPoint document or to share any information or materials outside of the session.

The TEM4 test paper and scoring. The TEM4 has six sections and 100 questions (as is shown in Table 1). The test is first graded on a 0-140 score scale, and then the raw scores are converted into their corresponding final scores on a 0-100 score scale. Cut-off scores are set at

60, 70, and 80, with 80-100, 70-80, 60-70, and 0-60 respectively indicating the levels of good, fair, pass, and fail. More information about the test can be found in the TEM4 Syllabus in Appendix I.

Table 1

The TEM4 Components

Part	Question Number	Part Name	Format	No. of Items	Scores	Percentage	Time (min)
I		Dictation	Subjective	1	15	15	15
II	1-30	Listening Comprehension					
		A Conversation	Objective	10	30	15	15
		B Passage	Objective	10			
		C News	Objective	10			
III	31-50	Cloze	Objective	20	20	10	15
IV	51-80	Grammar and Vocabulary	Objective	30	30	15	15
V	81-100	Reading Comprehension	Objective	20	20	20	25
VI		Writing					
		A Essay	Subjective	1	15	15	35
		B Note	Subjective	1	10	10	10
Total	100			103	140	100	130

Note. Translated and extracted from the TEM4 Syllabus

Data Analysis

STATA was used for data preparation and all the statistical tests (i.e., *t*-tests, multiple linear regression analyses, and expectancy table analyses) involved in answering the three research questions. The first research question enquires about the main effects of the tutorial session on participants' test preparation process. As suggested earlier, the effects were measured by applying both quantitative and qualitative metrics. Quantitatively, the focus was score improvement—the TEM4 scores were regressed onto the predictor variables of treatment status, GPA, and the interaction term between the two. Those independent variables were selected because, as is to be shown in the next section, the two groups were roughly equivalent with respect to all relevant predictors other than GPA and treatment status, and because it was assumed that substantial interaction could occur between GPA and treatment status given the fact that students who chose to attend the tutorial session tended to be those who were less successful in coursework. To complement the regression analyses, responses to the third survey, after being translated into English, were also analyzed for detecting score improvement. Qualitatively, the focus was behavior changes. Because no actual observations have been undertaken in this study, we relied on students' self-reported behavior to determine whether they were still practicing to the test or they have worked their way up to genuine improvements of English abilities. Again, translated responses to Survey III were examined for investigating behavior changes.

The second question looks into the effects of the tutorial on students' meta-knowledge of the condensed TEM4 Syllabus. This question was primarily addressed by a *t*-test that compared the mean scores obtained by the two groups on Survey II. A finer-grained analysis also looked into the group-disaggregated error rates of each question, as well as the contents of those questions that maximally discriminated between groups.

The third research question investigates the effects of familiarity with the condensed TEM4 Syllabus on test performance. The meta-knowledge scores that the participants achieved on Survey II were used as an index of familiarity level. The data of the experimental group and the control group have been combined into a single dataset for answering this question. A correlational analysis was first performed for a general exploration of the relationships between variables. Two statistical techniques—multiple linear regression (MLR) and expectancy graphing—were later adopted for analyzing the relationship between participants' meta-knowledge of the condensed TEM4 Syllabus and their test scores.

In terms of MLR, two models were constructed with the TEM4 scores first regressed on meta-knowledge scores, GPA, test preparation time, motivation level, and the level of perceived test importance (the Grand Model), and then regressed on a combined variable of motivation and perceived test importance along with all other independent variables in the Grand Model (the Combined-info Model). The association between meta-knowledge scores and test scores was examined by holding all other variables constant. Reasons for running the second model will be provided later when the results of the correlational analysis are discussed.

For the expectancy graph, a reduced dataset was built by taking out the data of the participants whose GPA fell in the top 10% or bottom 10% of the GPA distribution. To exclude these extreme GPA-possessors from this analysis is because students who perform extremely desirably or undesirably at school are very likely to receive extremely high or low scores in the TEM4 regardless of their meta-knowledge of the condensed TEM4 Syllabus, whereas students whose GPAs stay in the middle range are exposed to a larger room for uncertainty. Valid cases were further divided into six subgroups (i.e., top 25%, middle 50%, and bottom 25%) according their relative standing on meta-knowledge scores and the TEM4 scores. Once the subgroups were created, expectancy graphs were drawn by cross-tabulating the TEM4 subgroups and the meta-knowledge subgroups.

CHAPTER 5: RESULTS

Descriptive statistics are first reported to examine statistical assumptions and to provide an overview of test-preparation-and-Syllabus-relevant information, GPA, meta-knowledge scores, and the TEM4 scores. The results of the three research questions will be followed.

Descriptive Statistics

Based on the results of Survey I, there was no significant difference between the two groups in terms of their *most frequently used test preparation method(s)*, *familiarity with test syllabus*, *average test preparation time*, *level of perceived test importance*, or *motivation for test preparation*. Description statistics of each of the items in Survey I will be presented below.

Table 2 summarizes the experimental and the control groups' responses regarding their uses of different test preparation methods, with both frequencies and percentages given. We can see that an overwhelmingly large number of participants in both groups indicated that they prepared for the test by practicing both published and practice tests, with the former being an especially popular choice among the participants of this study. While three students in the control group also selected *reading the TEM4 Syllabus*, nobody in either group registered for commercial coaching courses or hired private tutors.

Table 2

The Use of Test Preparation Methods Sorted by Group

	Experimental		Control	
Prep Methods	Freq.	Percent	Freq.	Percent
Published Tests	18	100	28	93.33
Practice Tests	13	72.22	21	70.00
Test Syllabus	--[a]	--	3	10.00
Coaching Course	--	--	--	--
Private Tutoring	--	--	--	--

Note. For the experimental group, total n = 18, and for the control group, total n = 30.

[a] No participants were found for the category.

Given the popularity distribution of different test preparation methods, it was not surprising that almost every participant self-reported *practicing published tests* as his/her most frequently used preparation methods, followed by *doing practice tests*. In terms of students' familiarity with the TEM4 Syllabus, only six participants, all of whom were members of the control group, indicated that they had read the TEM4 Syllabus. Except that one student suggested that the reading of the TEM4 Syllabus was useful, the remaining were quite reserved about the effects of the TEM4 Syllabus on test preparation or test performance.

The survey results of average test preparation time have been sorted into four levels, as displayed in Table 3. Over a half of the experimental group spent 8-12 hours or more on the TEM4 preparation every week. Though the control group seemed to spare less time for test preparation, a chi-square test of independence showed that there was no statistically significant relationship between test preparation time and group, X^2 (3, *N* = 41) = 4.98, *p* = .173.

Table 3

Average Test Preparation Time Reported by Both Groups

	Experimental		Control	
Prep Time	Freq.	Percent	Freq.	Percent
<5 hours	1	5.88	7	29.17
5-8 hours	6	35.29	9	37.50
8-12 hours	8	52.94	6	25.00
>12 hours	1	5.88	2	8.33

Note. Cases kept after the third round of data collection were used to create this table. For the experimental group, total *n* = 17, and for the control group, total *n* = 24.

Summary statistics for *motivation* and *perceived test importance* are reported below in Table 4 along with participants' *GPA*, *meta-knowledge scores*, and *TEM4 scores*. Both the experimental group (*M* = 3, *SD* = 1.27) and the control group (*M* = 3.29, *SD* = 1.00) perceived the TEM4 scores to be moderately important to their future, and accordingly self-evaluated themselves to be somewhat motivated towards preparing for the test (*M* (E) = 3.18, *SD* = 1.19; *M*

(C) = 3.21, *SD* = 0.98). No significant group difference has been found in the mean score of these two measures.

Table 4

Test Relevant Variables: Descriptive Statistics

	All		Experimental		Control		*t*-test	
	M	SD	M	SD	M	SD	Difference	*p*-value
Prcvd Imprt[a]	3.17	1.12	3	1.27	3.29	1	-0.29	0.417
Motivation	3.19	1.05	3.18	1.19	3.21	0.98	-0.03	0.926
GPA	3.51	0.21	3.35	0.19	3.62	0.13	-0.27	<0.001
Meta-Kwnldg[b]	13.54	2.16	14.41	1.94	12.92	2.12	1.49	0.027
TEM4	71	5.6	67.82	6.03	73.25	4.07	-5.43	0.001

Note. Cases kept after the third round of data collection were used to create this table. For the experimental group, total *n* = 17, and for the control group, total *n* = 24.

[a] Perceived test importance
[b] Meta-knowledge scores of the enhanced TEM4 Syllabus

GPA, as the indicator of the participants' prior academic achievements, however, statistically distinguished the two groups from each other. Specifically, participants in the control group demonstrated higher academic achievements than their peers in the experimental group. Recall that this study did not involve a randomized controlled trial but a quasi-experiment, where students self-selected to participate in the study. One plausible reason for such an initial difference between groups was that students who were less successful in coursework would more likely be attracted to the treatment (i.e., the tutorial session) in hopes that they could perform better on the test with its help. Hence, instead of having two perfectly equivalent groups, we ended up with two groups that satisfactorily comparable in all aspects of interest except for GPA. Considering that students' prior achievements play an extremely influential role in affecting their test scores (Messick, 1982), if not the most important one, the subsequent analysis will have to account for the group difference in GPA.

The other two variables with which significant group differences have been found are meta-knowledge scores and the TEM4 scores. While the control group achieved higher test scores on the TEM4, the experimental group outperformed them in the meta-knowledge survey of the condensed TEM4 Syllabus. Since the meta-knowledge survey and the TEM4 are both outcome measures of this study, group disparities in those aspects will be subject to more sophisticated analysis later by taking account different predictor variables.

Research Questions

Whether directing test candidates' attention to the TEM4 Syllabus before the test can benefit their test preparation or test performance. In terms of score improvement, the main effect of the tutorial was not significant. The regression model accounted for a significant 54.95% of the variance in the TEM4 scores (*F* [3, 37] = 15.04, *MSE* = 15.29, $p < 0.001$). Among the three predictors, only GPA was found to be significant ($\beta = .74$, $p < .001$). After checking Cook's Distance and DFBETAs, one outlier was identified. It was decided to be removed from the dataset because its TEM4 score was merely 55, which was very atypical for students at the university in question, where the pass rate on the TEM4 is usually 99%. It was speculated that this person was simply unmotivated for test preparation, as confirmed by his perceived test-importance level at one and motivation level at three. The same model without the outlier accounted for a significant 52.16% of the variance in the TEM4 scores (*F* [3, 36] = 13.08, *MSE* = 15.29, $p < 0.001$). Though GPA was still the only significant predictor ($\beta = .65$, $p < .001$), the regression parameter estimates for treatment status and the interaction term have been raised by 400% and 300% respectively, as shown in Table 5.

Table 5

Comparison of regression coefficients for models with and without the outlier

	b		*SE b*		*p*		β	
Predictors	1[a]	2[b]	1	2	1	2	1	2
Group	3.01	15.25	28.29	26.70	0.92	0.57	0.27	1.50
GPA	19.79	16.31	5.13	4.95	<.001	0.00	0.74	0.65
Group*GPA	0.83	4.30	8.04	7.59	0.92	0.57	0.27	1.53

[a] Model with the outlier
[b] Model without the outlier

The output of the two MLR analyses indicates that receiving the tutorial session failed to confer a statistically distinguishable advantage to the experimental group over the control group in taking the test, with GPA and interaction being held constant. Post-test survey data reinforced these findings in a sense but meanwhile provided a more thorough picture of the experimental group's evaluation of the condensed TEM4 Syllabus and how they made use of it during their test preparation process.

Of those who responded to the post-test survey, only one person believed that the enhanced familiarity with the condensed TEM4 Syllabus had truly inflated his test score, whereas the others held relatively neutral opinions towards the effects of the tutorial session on improving their test scores. Nonetheless, everyone acknowledged the positive help that the tutorial had lent to his/her overall test preparation process. Their answers can be summarized into two general categories: first, by acquainting them with the test contents, the tutorial allowed them to identify their weaknesses in the target knowledge or skills and thereby to prepare with a focus; secondly, their anxiety on the test day had been greatly reduced thanks to having learned about the test procedure and task flows from the condensed TEM4 Syllabus. While it seemed at the first sight that the condensed TEM4 Syllabus not only facilitated English learning and produced the first type of preparation effects (see the literature review), but also fostered testwiseness and brought about the second type of preparation effects (see the literature review), both instrumental for language learning and hospitable for test validity, the actual preparation effects associated with the use of the condensed TEM4 Syllabus, however, manifested a deviated pattern after the participants' responses were subject to a closer examination, as will be explained below.

When asked to compare the usefulness of the condensed TEM4 Syllabus and other test preparation methods, the respondents showed a clear preference for the latter, remarking that the former only pointed out a general direction for test preparation, and that practicing sample items was the only method that worked for test preparation and guaranteed high scores. Given their responses to this question, the word *practice* in the statement, "the tutorial helped them to practice with a focus," was re-interpreted to have a literal meaning of *doing exercises* instead of a figurative meaning of *improving English abilities*.

Therefore, the intervention of the tutorial did not reverse the tendency of the experimental group to practice to the test; instead, it seemed to have further aggravated the situation. Regarding the relationship between reading the TEM4 Syllabus and English learning, several respondents commented that the two did not have a solid connection with each other. Their distrust of the TEM4 Syllabus as a potential guide for improving English abilities through test preparation further implied that they might not have engaged in any activities (e.g., listening to BBC and VOA radios to prepare for the listening comprehension section) recommended during the tutorial for them to make genuine improvements.

Given the analysis above, the function performed by the condensed TEM4 Syllabus in the experimental group's test preparation process was closer to a guide that specified what to practice instead of something that truly facilitated and improved English learning. The previous speculation about preparation effects has then been modified into the finding that the effects associated with the tutorial session, instead of being Type One and Type Two, were actually Type Two and Type Four; more specifically speaking, the enhanced version of the TEM4 Syllabus equipped students with testwiseness as well as encouraged them to practice to the test in a more efficient manner.

Whether test candidates are able to acquire the meta-knowledge of the TEM4 Syllabus from other resources, such as the published tests of previous years. A *t*-test has been used here in spite of the group difference in GPA because it was believed that GPA only peripherally associated with students' abilities to comprehend the tutorial session, apply RE, or understand questions in the meta-knowledge survey. Such assumptions about GPA were supported by the following facts: First, both the tutorial and the meta-knowledge survey were in Chinese with neither dealing with subject-matter knowledge; and second, cognitive skills rather than language skills are primarily deployed during the process of RE. While the GPA of English majors may effectively index students' language skills, it does not necessarily reflect students' cognitive skills, such as making inferences or syntheses.

The results of the *t*-test have also indirectly confirmed the above assumption regarding GPA in that the experimental group scored significantly higher in the meta-knowledge survey than the control group (*t* [39] = 2.30, *p* = .027). If GPA exerted similar influences on the meta-knowledge survey scores as it does on the TEM4 scores, we would not expect such considerable positive differential effects for the treatment, given that the experimental group had a significantly lower GPA than the control group. Because the two groups have been demonstrated to be equivalent in their test preparation behavior, and since no external control was laid on their use of test preparation resources, the difference found in the meta-knowledge survey scores between groups could be reasonably attributed to the treatment.

To obtain a finer-grained picture of where the group difference occurred, a post-hoc analysis of error rates was undertaken, whose results can be found in Table 6. Each percentage in the rows labeled *Experiment* and *Control* represents the proportion of the survey takers who answered a question wrong in the corresponding group, and percentages in the rows labeled *Difference* denote the subtraction of the control group's error rate from that of the experimental group. From the signs of the percentages in the *Difference* rows, we can easily tell that the experimental group performed better than its counterpart in 12 out of the 17 questions. Considering that some questions only subtly distinguished the experimental group from the control group, the question range of interest has been narrowed down by marking out those questions whose differential percentages had an absolute value over 10%, as shown below.

Table 6

Error Rates by Question

	Q1*	Q2	Q3	Q4	Q5	Q6*
Experiment	41.18%	0%	5.88%	5.88%	0%	5.88%
Control	70.83%	0%	8.33%	12.50%	4.17%	16.67%
Difference	-29.65%	0.00%	-2.45%	-6.62%	-4.17%	-10.79%
	Q7	Q8	Q9	Q10*	Q11	Q12
Experiment	29.41%	5.88%	29.41%	5.88%	5.88%	35.29%
Control	37.50%	4.17%	25.00%	16.67%	0%	37.50%
Difference	-8.09%	1.71%	4.41%	-10.79%	5.88%	-2.21%
	Q13	Q14*	Q15*	Q16*	Q17	
Experiment	5.88%	35.29%	0%	23.53%	11.76%	
Control	8.33%	54.17%	25%	66.67%	20.83%	
Difference	-2.45%	-18.88%	-25.00%	-43.14%	-9.07%	

Note. For the experimental group, n = 17, and for the control group, n = 24.

* Questions that maximally discriminate between the two groups (i.e., |Difference| > 10.00%).

Those questions marked out by a star were believed to maximally discriminate between the two groups, and we can see that their differential percentages are unanimously negative, which means that the experimental group gained a higher average score than the control group on all of these questions. Evidently, the six starred questions (which are listed below) played the most crucial role in raising the experimental group above the baseline performance of the control group. Therefore, a further examination was carried out to see how they managed to do so. It was found that except for Q1, all the other questions enquired about information almost unobtainable from RE; in other words, preparation materials other than the TEM4 Syllabus itself would hardly contain hints to their answers. To take Q2 as an example, unless the practice exercises used by test takers provided detailed scoring rubrics strictly based on the TEM4 Syllabus, which they often do not, there would be little chance for students who have not attended the tutorial to answer the question correctly.

Q1. How many parts are included in the TEM-Band 4?
Q6. Is it possible for you to get a lower score because of punctuation errors?
Q10. Is American English only tested in this part of the test?
Q14. Is it true that instructions and charts could appear in this part of the test?
Q15. What is the desirable length of the essay?
Q16. Could you be asked to write an exposition in the section of essay writing?

To summarize the foregoing analysis on research question two, the experimental group has demonstrated more accurate meta-knowledge of the condensed TEM4 Syllabus thanks to the treatment, which thereby implies that the information provided in the TEM4 Syllabus cannot be fully interpreted by students from other resources.

Whether the more accurate meta-knowledge of the TEM4 Syllabus can result in better test performance conditionally on other relevant variables. Table 7 shows the correlations between different predictor variables and the outcome variable. Treatment status had been excluded from the list of predictor variables because here the interest was no longer the main effect of the treatment but instead that of the meta-knowledge scores. In the first column of the table we can see that the TEM4 scores and GPA have a strong, positive, and linear relationship. However, we also note those negative correlation coefficients of the TEM4 scores with meta-knowledge scores and test preparation time. Moreover, the correlations between the

TEM4 scores and perceived test importance and motivation were extremely low. One plausible explanation for these "unusual" coefficients is the sampling problem mentioned earlier. Because GPA had the strongest association with the TEM4 scores among all the five predictor variables, and because those who had lower GPA, most of whom were in the experimental group, tended to have higher meta-knowledge scores and to report stronger motivation, higher perceived test importance, and longer average test preparation time, the correlation coefficients of these four variables with the TEM4 scores were unexpectedly low or even negative. In addition, the ratings of motivation level and perceived test importance could be unreliably represented due to the limitations of self-report data; that is, the students might interpret the same scale in distinctively different ways. It should be noted that though test preparation time was also reported by the students instead of being observed by the researcher, it was less susceptible to the limitations described above because it attempted to measure a concrete behavior rather than an abstract construct such as motivation.

The second highest correlation coefficient in the matrix was the one between perceived test importance and motivation. Such a high correlation between the two dependent variables implied that they should measure the same aspect and would probably be redundant if both were included in a regression model.

Table 7

Correlations between Dependent and Independent Variables

	TEM4	Meta Scores	GPA	Prcvd Imprt	Motivation	Prep Time
TEM4	1.00					
Meta Scores	-0.11	1.00				
GPA	0.74	-0.25	1.00			
Prcvd Imprt[a]	0.07	-0.08	0.31	1.00		
Motivation	0.11	0.00	0.36	0.67	1.00	
Prep Time	0.18	0.03	-0.24	0.12	0.34	1.00

[a] Perceived test importance

Based on the correlation matrix, two MLR models were constructed. A Grand Model was first run with all the predictor variables discussed above. Since the same outlier had been identified as the one in the model built for research question one, the Grand Model was run again without the outlier, and Table 8 displays a comparison of the regression coefficients between the analyses with and without the outlier. While the analysis with the outlier shows that the model has explained a significant 60.02% of the variance in the TEM4 scores (*F* [5, 35] = 10.51, *MSE* = 14.35, *p* < 0.001), the one without the outlier shows that the model has explained a significant 62.92% of the variance (*F* [5, 34] = 11.54, *MSE* = 10.84, *p* < 0.001). In both analyses, GPA appears to be the only significant predictor of the TEM4 scores.

Table 8

A Comparison of Regression Coefficients for the Grand Model with and without Outlier

	b		*SE b*		*P*		β	
Predictors	1[b]	2[c]	1	2	1	2	1	2
Meta Scores	0.26	0.29	0.29	0.25	0.37	0.27	0.10	0.12
GPA	23.85	22.13	3.53	3.10	<.001	<.001	0.89	0.89
Prep Time	0.65	1.21	0.80	0.72	0.43	0.10	0.10	0.20
Prcvd Imprt[a]	-0.34	-1.22	0.74	0.69	0.65	0.08	-0.07	-0.26
Motivation	-1.09	-0.55	0.89	0.79	0.23	0.49	-0.20	-0.12

[a] Perceived test importance
[b] The analysis with outlier
[c] The analysis wihout outlier

Given the high correlation between motivation and perceived test importance, scores of the two variables were combined to avoid including redundant or losing important information. Table 9 presents the results of the Combined-info Model. A vertical comparison within the Combined-info Model reveals a similar pattern as the Grand Model—with the outlier, the Combined-info Model has explained a significant 60.02% of the variance in the TEM4 scores (*F* [4, 36] = 13.34, *MSE* = 14.05, *p* < 0.001), whereas without the outlier, the model explained a significant 62.64% of the total variance (*F* [4, 35] = 14.67, *MSE* = 10.61, *p* < 0.001). As with the

Grand Model, the removal of the outlier in the Combined-info Model has considerably bumped up the regression parameter estimate for each of the dependent variables.

Table 9

A Comparison of Regression Coefficients for the Combined-info Model with and without Outlier

	b		*SE b*		*p*		β	
Predictors	1[b]	2[c]	1	2	1	2	1	2
Meta Scores	0.24	0.30	0.28	0.25	0.40	0.23	0.09	0.13
GPA	23.44	22.54	3.40	2.96	<.001	<.001	0.87	0.90
Prep Time	0.51	1.30	0.75	0.69	0.50	0.07	0.08	0.22
MPCom[a]	-0.67	-0.92	0.35	0.31	0.06	0.01	-0.24	-0.36

[a] A combined variable of perceived test importance and motivation
[b] The analysis with outlier
[c] The analysis wihout outlier

Table 10

Adjusted R^2 and Standard Error of both Models

Model	Outlier	Adjusted R^2	*se*
Grand	w	54.30%	3.79
	w/o	57.46%	3.29
Com-Info	w	55.24%	3.75
	w/o	58.37%	3.26

A cross-model comparison of adjusted R^2 and standard error can be found in Table 10. As the Combined-info Model produced larger adjusted R^2 (Adjusted R^2 = 58.37%) and smaller standard error (*se* = 3.26), the predictor variables included in the Combined-info Model have provided a better prediction than those in the Grand model. In addition, three predictor variables in the Combined-info Model were found to be significant—GPA and the combined variable both had a significant regression parameter estimate at the *.05* level, and the effect of preparation time was significant at the *.1* level.

Based on the model outputs above, the two models, especially the Combined-info Model, seemed to satisfactorily approximate the relationship between the variables of interest, although meta-knowledge scores remained an insignificant predictor of the TEM4 scores across the models or analyses.

To further explore the relationship between the TEM4 scores and meta-knowledge scores from the second survey instrument, the data within the middle 80% of the GPA distribution was analyzed using an expectancy graph. Figure 4 shows what percentage of participants within each of the three meta-knowledge groups (i.e., the upper 25%, the middle 50%, and the lower 25% of meta-knowledge scores) received a TEM4 score in each of the three TEM4 groups (i.e., the upper 25%, the middle 50%, and the lower 25% of the TEM4 scores). The three meta-knowledge groups are represented respectively by the three bars in the graph, with percentages of the three TEM4 groups represented by different colors in each of the bars.

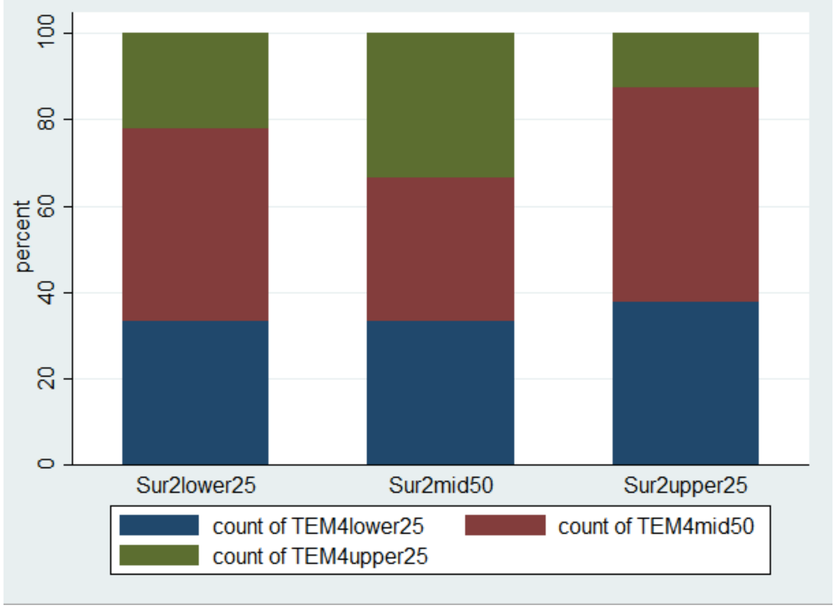


Figure 4. Percentage of the participants earning upper 25%, middle 50%, and lower 25% TEM4 scores by meta-knowledge groups.

The expectancy graph provides no compelling evidence to the predictive power of meta-knowledge scores, either. We can see no increasing trend in the likelihood for a student who had a particular standing in meta-knowledge survey to have a corresponding standing on the TEM4.

Instead, the students in the top range of the meta-knowledge scores were even shown to be less likely to get a higher score on the TEM4 than those in the middle range. Relating the ambiguous pattern in the expectancy graph with the previous regression analyses, we can summarize that the more accurate meta-knowledge of the enhanced released test specs has failed to lead to better test performance.

CHAPTER 6: DISCUSSION

No Significant Effects or Positive Behavior Changes Associated with the Condensed TEM4 Syllabus

The preparation effects associated with the condensed TEM4 Syllabus have been examined from two perspectives using a variety of analytical methods. While a significant difference has been found in meta-knowledge scores between groups, directing students' attention to the condensed TEM4 Syllabus did not lead to a significant inflation of test scores; nor did the familiarity level of the condensed TEM4 Syllabus prove to be a significant predictor of test scores. Although the evidence at hand has not supported our null hypothesis that closer attention to and higher familiarity with the condensed TEM4 Syllabus would lead to better test performance, a discussion is worthwhile on why this was the case.

One plausible explanation is that one month is probably too short for a one-shot tutorial session to manifest significant effects. Admittedly, it might be difficult for any type of educational treatments (even a repeated tutorial series) to demonstrate significant effects within such a short period of time; however, even if the experimental group had been provided with sufficient time for test preparation, there might still remain no significant treatment effects, because their test preparation behavior did not lead to genuine improvements in their English abilities, and what they had gained were only practice effects. As stated in the results part for research question one, no evidence has suggested that the experimental group engaged in activities that could genuinely improve their English abilities; instead, the situation of practicing to the test became aggravated for them upon the intervention of the tutorial session. That is, the students were not only practicing to the test as the control group; they developed it into a more systematic practice with the guidance of the tutorial session. Since the direction of their behavior changes went completely against the original intention of the investigator, it might be reasonable for us to have observed no significant treatment effects, given the fact that the intended comparison of effects between spec-guided English learning and practicing to the test had been changed into the one between spec-enhanced practicing to the test and spec-free practicing to the test.

This then brings up another question as to why students failed to take advantage of the released test specs as expected. One direct answer yielded from this study was that students did not trust the usefulness of the TEM4 Syllabus. Responses to Survey I disclosed that most students held a doubtful attitude towards the usefulness of the TEM4 Syllabus even without reading it. Their distrust should have, to some extent, exerted negative influences on their evaluation of its contents and their willingness to actually follow its suggestions over the test preparation process. The data of Survey III confirmed the students' reluctance to experiment with the advice provided in the TEM4 Syllabus, even after they were encouraged to do so during the tutorial session.

However, to go below the surface of this trust crisis, I see culture and the quality of the released specs working interactively to drive students away from the TEM4 Syllabus. We should note that the students' anti-syllabus attitude and behavior cannot be simply attributed to individual preferences. At a more fundamental level, there is no historical or cultural imperative for test takers to trust the test syllabus. The culture directs them to prior sample tests and might also suggest that it is far more important to rely on RE than undertaking detailed analysis of the test syllabus, despite how beneficial the syllabus might claim to be. To relate our present discussion to the operational definitions of released specs provided earlier, at least in this study, Chinese students were revealed to find RE specs much more attractive than the official syllabus, though unfortunately, their over-reliance on RE specs is just another teaching-to-the-test practice, which is not ethical, either.

On the other hand, the form of language in TEM4 Syllabus might not be generative or precise enough for it to be trustworthy or utilizable to students during test preparation. I used *might* here because to determine the truthfulness of this statement we would need to look inside a black box—the releasability of specs, which varies across tests, test developers, and contexts in a mysterious manner, because no underlying rationales are known to outsiders. Yet some evidence I have gathered by examining the TEM4 Syllabus can lend some support to what I am trying to argue here. To begin with, some of the descriptions in the TEM4 Syllabus are over-vague and of no practical value. For example, the following sentence appears many times in the TEM4 Syllabus, "the content of this passage is at the intermediate level," often without further information given explaining how the intermediate level is defined. Another example is that the TEM4 Syllabus is inclined to align its assessment requirements with the learning requirements set by the National College English Teaching Syllabus for English Majors (hereinafter the Teaching Syllabus, NACFLT, 2000), such as the difficulty level of grammar points and vocabulary items, in spite of the fact that the majority of students are not familiar with the Teaching Syllabus.

Secondly, the TEM4 Syllabus includes some information that unfaithfully represents the actual test contents. A good example can be found in the descriptions of the listening section. Even though it reads, "students should be able to deal with listening materials at the intermediate level, such as the mini-lectures in the TOEFL," no tasks similar to the TOEFL mini lectures have been discovered upon a close examination of the test contents. Since the vague or misleading information might confuse students if they do not have much experience with the test, students are very likely to be attracted to the published test papers and RE specs.

A Tension between Usefulness and Releasability for Test Specs

Previous discussions have focused on some possible reasons that might have rendered the TEM4 Syllabus an unpopular test preparation material for Chinese students. In this section, we will first review how the popular practice of teaching to the test can be problematic if evaluated against different standards, then suggest some solutions towards making a positive change, and finally discuss the difficulties for such solutions to take effect by returning back to the issue of spec releasability.

Teaching/practicing to the test, as represented by doing authentic and sample test items, has been criticized by generations of testing scholars using different standards (e.g., Haladyna & Downing, 2004; Mehrens & Kaminski, 1989; Popham, 1991). Briefly speaking, the application of a standard of ethical appropriateness will classify such test preparation behavior as indirect cheating, and another standard of educational defensibility will expose its impediment to the development of students' ability in the long run. As far as test integrity is concerned, practicing to the test will not only undermine the validity of test-score related interpretations and decisions, but also raise concerns about test security. Before we move on to suggestions, I would like to make a possibly bold extrapolation and to extend these evaluations of teaching-to-the-test practices to RE activities and their product of RE specs. To some extent, the impact of students' RE might be worse than practicing to the test itself because if they are able to hack into the actual test specs that are not supposed to be released, the good intention of having this test no longer exists.

Therefore, how do we protect a test from the destructive attacks from published test papers and systematic RE? To me, this question is an alternative way of asking how to make test syllabi more trustworthy and useful. The key to solving this problem lies on either the first or the third level of released specs (see operational definitions for released specs). For the first level of released specs to better serve test preparation, test developers should tailor its contents to better suit the needs of test takers. In the context of this study, the current version of TEM4 Syllabus might cater to the demands of the test sponsor and supervisor, but it does not necessarily please test users or test takers. What test takers would find more helpful is a form of language more generative and precise in nature, compared with what we have seen in the preceding section. As long as it is consistent with test security and validity, the test committee should make their best attempt at adequately preparing test takers for the test, and such an attempt at refining the test syllabus, in return, will not only earn students' trust, but more importantly reciprocate test security and validity by discouraging teaching-to-the-test practices and over-reliance on RE specs.

Alternatively, the third level of released specs can be further explored and employed. Sometimes, it might be hard for the first level of release specs to serve all its target readers equally well. Under such circumstances, the third level of released specs can be presented to test takers in an appropriate format to complement the first level. For instance, test developers can authorize a group of testing professionals to be their spokesmen to explain the test syllabus. This study chose to conduct a one-shot tutorial session, and as we have seen, it has not achieved much success in gaining students' trust. A better way of doing this is probably to provide sustained support to students, that is, to make the tutorial a workshop series where participants can analyze the TEM4 Syllabus together with the professional, and meanwhile, when the workshop is not offered, the professional can always be available to answer the questions students have on the TEM4 Syllabus. If students can feel secure and confident enough from the relying on the third level of released specs, it is very likely that they will voluntarily turn away from sample test papers and put more effort into improving their English abilities.

However, these solutions are apparently much easier said than done, because both lead us back to the issue of spec releasability (Davidson, 2012a), which, as I have said, is a massive black box. Extant literature has informed us of a fine line between what is releasable and what is not, but it has not provided us with hard and fast rules that we may follow to resolve this issue, and probably there will never be considering the great diversity we have in the world of testing. When test developers possess absolute power over the releasability issue, it is understandable if they are unwilling to follow the solutions suggested above but to opt for a more conservative path out of their concerns about test security and validity. The only downside is that their concerns may backfire, as we have seen in the case of TEM4, because when stakeholders, especially test takers, cannot be truly benefited from the released test specs, they will always resort to unethical test preparation behavior and launch an even stronger attack on test validity or security, and a long-run consequence is that a culture where test syllabi are trusted can never be successfully fostered.

CHAPTER 7: LIMITATIONS

The biggest limitation lies in the sampling procedure of this study. In particular, students self-selected when deciding to participate, and as a result, the experimental group was significantly biased against in terms of their GPA, the most influential factor that could affect test performance. More accurate causal inferences would have been drawn if a randomized sample had been recruited.

Moreover, the sample size of this study is relatively small for obtaining trustworthy results. As a rule of thumb for conducting a multiple regression analysis requires, a minimal sample size should be at least 15 times as large as the number of predictors in the model, which means that in the case of this study, we should have over 100 students in the sample for the results to be trustworthy.

The last problem with the sampling procedure is representativeness. Since all the participants involved in this study came from a single university in China, they were not believed to be representative enough of the entire population of English majors. Particularly, the test pass rate for students in this university has been much higher than average universities in China, which might thereby negatively influence the participants' motivation for test preparation.

The second limitation concerns the self-report method widely employed in this study. To begin with, the self-reported motivation level and perceived test importance are not totally reliable. As previously mentioned, students' motivation towards test preparation and how important they perceived the test results were to their future were measured in Survey I using two five-point Likert-scale questions. While this self-report measure was easier to implement, especially in the situation of an international study, problems nonetheless arose around the reliability of the responses collected. For instance, students could interpret and therefore use the

Likert scales differently. It was very likely that two students at the same level of motivation ended up rating themselves differently on the scale, simply because the points conveyed different meanings to them. On the other hand, since it was a five-point scale, point three might have attracted those “midpoint hoggers” even before they gave serious thoughts to the questions.

In addition, to rely solely on self-reported data for analyzing preparation behavior can be risky. For one thing, human memory is fallible. It was likely that the experimental group suffered from inaccurate memory when they tried to report their preparation behavior on Survey III, because almost half a year had passed since the test administration. Therefore, on-site observations of students’ behaviors are necessary to ensure more trustworthy results.

CHAPTER 8: CONCLUSION

Focusing on the role of released specs during test preparation, this study examined the pattern of test preparation for the TEM4, the effects associated with directing students' attention to the TEM4 Syllabus during test preparation, the possibility of obtaining important syllabus information from other test preparation materials, and the predictive power of familiarity with the TEM4 Syllabus. As far as the entire sample is concerned, the most frequently used preparation method was practicing published test papers, and the TEM4 Syllabus was often ignored as the students prepared for the test. Moreover, familiarity with the TEM4 Syllabus was not shown to be a significant predictor of the TEM4 scores. In terms of the data disaggregated by group, while the tutorial on the condensed TEM4 Syllabus significantly raised the experimental group's meta-knowledge of the TEM4 Syllabus, it failed to result in significant test score improvement; neither did it elicit any positive changes in students' preparation behavior—the experimental group was still practicing to the test, and this time with greater efficiency and clearer focus under the guidance of the tutorial session.

Such unexpected behavior changes in the experimental group not only revealed practicing to the test as an inherited cultural habit of Chinese students, but also pointed to having higher-quality released specs as a possible solution for transforming students' problematic preparation behavior. However, this solution path is obstructed by the issue of spec releasability, because what information in test specs is releasable and how to evaluate a set of released specs remain, at best, a definite black box in theory at this moment. Any advancement we make in unlocking this black box will be rather critical for fostering a healthy testing culture where teaching-to-the-test practice is avoided and released test specs are trusted. More empirical research is needed for us to understand this issue better, and in light of the findings of this study, one recommendation for future studies is that a qualitative instead of a quantitative methodology should be adopted, such as conducting a focus group.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

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

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Badger, R., & Yan, X. (2012). To what extent is communicative language teaching a feature of IELTS classes in China? *IELTS Research Reports, 13*. Retrieved 4/11/2016 from https://www.ielts.org/~/media/research-reports/ielts_rr_volume13_report4.ashx.

Briggs, D.C. (2009). *Preparation for College Admission Exams*. NACAC Discussion Paper. Arlington, VA: National Association for College Admission Counseling.

Brown, J. (1998). An investigation into approaches to IELTS preparation, with particular focus on the Academic Writing component of the test. *IELTS Research Reports, 1*, 20-37.

Burn, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pre-test information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment*, *16* (1), 73-77.

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

Davidson, F. (2012a). Releasability of language test specification. Paper based on the keynote presentation at the 15th Annual Meeting of the Japan Language Testing Association, Osaka, 29 October 2011.

Davidson, F. (2012b). Test specifications and criterion referenced assessment. In Davidson and Fulcher (Eds.), *The Routledge Handbook of Language Testing.* London and New York: Routledge, 197-206.

Davidson, F. (2013). Test specifications. In C. Chapelle (General Editor). *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications.* New Haven, CT: Yale University Press.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Gan, Z. (2009). IELTS preparation course and student IELTS performance: A case study in Hong Kong. *RELC journal: A Journal of Language Teaching and Research, 40*(1), 23-41.

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy & Practice, 14*, 75-97.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-52 1.

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

Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly, 32*, 329-337.

Hudson, T. D., & Lynch, B. K. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing, 1,* 171-201

Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing, 28*(4), 589-596.

Li, J. (2006). Introducing audit trails to the world of language testing (Unpublished MA thesis). Division of English as an International Language, University of Illinois at Urbana-Champaign.

Lynch, B. K., & Davidson, F. (1997). Is my test valid? Paper presented at TESOL Convention, Orlando, Florida.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: Issues and Practices, 8*, 14-22.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist, 17,* 67-91.

Mickan, P. & Motteram, J. (2009). The preparation practices of IELTS candidates: case study. *IELTS Research Reports, 10*. Retrieved 4/11/2016 from https://www.ielts.org/~/media/research-reports/ielts_rr_volume10_report5.ashx.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003) On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

NACFLT. (2004). *Syllabus for TEM4*. Shanghai: Shanghai Foreign Language Education Press.

Perlman, C. (2003). Practice tests and study guides: Do they help? Are they ethical? What is ethical test preparation practice? In *Measuring up: Assessment Issues for Teachers, Counselors, and Administrators*. Retrieved 11/20/2015 from http://files.eric.ed.gov/fulltext/ED480062.pdf

Popham, W. J. (1978a). *Criterion Referenced Measurement*. Eaglewood Cliffs, NJ: Prentice Hall.

Popham, W. J. (1978b). Well-crafted criterion-referenced tests. *Educational Leadership, 36*(2), 91-95.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practices, 10* (4), 12-16.

Powers, D. & Rock, D. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Education Measurement, 36*, 93-118.

Read, J., & Hayes, B. (2003). The impact of the IELTS test on preparation for academic study in New Zealand. *IELTS Research Reports, 5*, 153-206.

Shephard, L. A. (1987). *A case study of the Texas Teacher Test: Technical report*. Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal* 28, 521-42.

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

Wall, D., & Horak, T. (2011). The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook; Phase 4, Describing change. *TOEFL iBT Research Report, 17*. Retrieved 3/22/2015 from https://www.ets.org/Media/Research/pdf/RR-11-41.pdf.

Wang, C., Yan, J., & Liu, B. (2014). An empirical study on washback effects of the Internet-based College English Test Band 4 in China. *English Language Teaching, 7* (6), 26-53.

Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*, 196-218.

APPENDIX A

A Translated Version of the Tem4 Syllabus

Preface

1. The purpose of the test
   This test aims to fully measure the ability of English major students who have accomplished all the courses at the elementary level. This test helps to decide whether the students have met the requirements set by the Teaching Syllabus by assessing their basic skills as well as their mastery of grammar and vocabulary.
2. The features and the domain of the test
   This is a criterion-referenced test, and its domain covers the four skills and grammar and lexical knowledge that are required by The Syllabus.
3. The time, the target test-takers, and the development of the test
   This test is administered once a year for English major students in their fourth semester of learning. The test is developed and administrated by The National Advisory Committee for Foreign Language Teaching (NACFLT).
4. The format of the test
   This test adopts several different formats in order to measure the students' English skills more efficiently, to guarantee the scientificity and the objectivity of the test, and to take into consideration the practicality of the test and the features of language assessment at the elementary level. The reliability and the validity of the test are also guaranteed in this way.
5. The content of the test
   This test consists of 6 parts, namely dictation, listening comprehension, cloze, grammar and vocabulary, reading comprehension, and writing. The test takes about 130 minutes.

Part one: Dictation

1. Requirements
   a) Be able to do a word-for-word dictation on the basis of a comprehensive understanding of the whole passage.
   b) Be able to spell every word correctly and to use punctuation in an appropriate way. To receive a full score, the errors should not account for more than 8%.
   c) The test lasts about 15 minutes.
2. Format
   This part is subjective. The passage will be read for 4 times in total. The first time is read at normal speed (120 words/min) so that the test taker can get a general idea of the passage. The second and the third time will be read slower with a 15-second pause left between sense groups, clauses and sentences. The test taker should finish writing by the end of the third time. The fourth time switches back to the normal reading speed and the test taker can check his/her answer.
3. Purpose
   This part aims to measure the test taker's listening comprehension ability, spelling, as well as the use of punctuation.
4. Material-choosing principle
   a) Genres and topics vary.

b) The material used is at the intermediate level and is within the requirements set by The Syllabus
c) The passage contains around 150 words.

Part two: Listening comprehension

1. Requirements
   a) Be able to understand daily conversations between the native speakers of English. Be able to deal with the listening material at the intermediate level (such as the mini lectures in TOEFL). Specifically, be able to grasp the general meaning of the material, as well as the speaker's attitude, feeling, and real intention.
   b) Be able to understand the general idea of VOA and BBC news.
   c) Be able to distinguish between different variations of English (such as American English, British English, and Australian English).
   d) This part lasts around 15 minutes.
2. Format
   Multiple-choice questions are used in this part and are divided into three sub-sections according to the nature of the material.
   Section A: Conversations
   There are several conversations in this section. Each of them contains about 200 words, and is followed by a couple of questions. There are 10 questions in total.
   Section B: Passages
   There are several monologues in this section. Each of them contains around 200 words, and is followed by a couple of questions. There are 10 questions in total.
   Section C: News Broadcast
   There are several pieces of VOA and BBC news in this section. Each of them is followed by a couple of questions, and there are 10 of them in total. A 5-second pause is left between every two news items. The reading speed is around 120 words per min.
3. Purpose
   This part is to measure the test taker's ability to acquire oral information.
4. Material-choosing principles.
   a) Some of the conversations and monologues talk about daily life.
   b) The topics of VOA and BBC news are those familiar to the test taker.
   c) In principle, words used in the material do not exceed the vocabulary requirement set by The Syllabus.

Part three: Cloze

1. Requirements
   a) Be able to select the best answer to complete the passage on the basis of a comprehensive understanding of the corrupted passage content.
   b) This part lasts 15 minutes.
2. Format
   Multiple-choice questions are used in this part. 20 words or phrases in a 250-word passage are left in blank. The topic of the passage is familiar to the test taker and the content is at the intermediate level. Every blank itself stands as a question, and the test taker should choose the best answer from four given choices. This part tests both grammar and vocabulary.

3. Purpose
   To measure the test taker's integrated English knowledge and skills.

Part four: Grammar and Vocabulary

1. Requirements
   a) Master and be able to use the grammar knowledge required by The Syllabus for band-1 to band-4
   b) Master the vocabularies (5500-6000) required by the Syllabus for English majors at the elementary level and be able to correctly use the most of them (3000-4000).
   c) This part lasts about 15 minutes.
2. Format
   Multiple-choice questions are used in this part. There are 20 questions in total, and each is provided with 4 choices. About 50 percent of the questions test the use of words and the rest test grammar.
3. Purpose
   This part is to measure the test taker's mastery of vocabularies and of basic grammar concepts.

Part five: Reading Comprehension

1. Requirements
   a) Be able to understand the intermediate-level articles and material published in the US and Britain.
   b) Be able to understand the international news reports at the similar difficulty level with *Newsweek*.
   c) Be able to understand the unabridged literature works at the similar difficulty level with *Sons and Lovers*.
   d) Be able to grasp the main idea of the material and to understand the facts and details that support the main idea; be able to not only comprehend the literary meaning of the article but also to make judgments and inferences; be able to get the meaning of individual sentences as well as the logic relationship between lines.
   e) Be able to adjust the reading speed according to needs.
   f) This part lasts around 25 minutes.
2. Format
   Multiple-choice questions are used in this part. This part comprises several 1800-word excerpts, and each of them is followed by a couple of questions. The test taker should select the best answer from the four given choices based on his/her understanding of the excerpts.
3. Purpose
   This part measures the test taker's ability to acquire information by using relevant reading strategies. Both accuracy and speed are emphasized in this part. The desired reading speed is 120 words per min.
4. Material-choosing Principles
   a) The topic varies to include areas such as sociology, technology, culture, economy, daily life, and bibliography, etc.
   b) The genre varies to include narratives, descriptive essays, expositions, argumentative essays, advertisements, instructions, and charts, etc.

c) The material is at the intermediate level, and key words do not exceed the vocabulary requirement set by The Syllabus.

Part six: Writing

1. Requirement
   a) Essay
   Be able to write a 200-word essay with the given title, outline, chart or figure. The essay should be on topic, complete in structure and meaning, grammatically correct, coherent in logic, and properly composed. This section lasts about 35minutes.
   b) Note
   Be able to write a 50-to-60-word note (including an invitation letter and an announcement) with the given hints. The note should be in correct format and proper language. This part lasts about 10 minutes.
2. Format
   This is a subjective part with two sub-sections.
   Section A: Essay
   The title of the essay is already given. The genres involved are largely expositions, narratives, and argumentative essays.
   Section B: Note-writing
   The test taker is going to write a note in this section.
3. Purpose
   To measure the test taker's writing ability according to the requirements set by The Syllabus.

Answering and Scoring Rubrics

The essay and the dictation should be written on the subjective answer sheet. The answers to the multiple-choice questions should be marked out on the objective answer sheet. The notes or answers given on the test papers are not eligible to be graded.

The essay and the dictation should be written with a pen or a ball pen. Answers written inside the gutters are invalid. The test taker is supposed to select only one correct answer; otherwise the answer is invalid. Please use a 2B pencil. No scores will be deducted if the test taker answers a question wrong. Objective questions are scored through an automated process.

No dictionaries or other reference books are allowed in the test.

First, a raw score is computed by taking the sum of the scores of different parts. Then it is converted to a scaled score of 0 to 100, with 60 being the cut score.

More detailed information on each part (such as the format, the number of items, the score percentage, and the test time, etc.) is given in the chart below:

Part	Question Number	Part Name	Format	Number of Items	Scores	Percentage (%)	Time (min)
1		Dictation	Subjective	1	15	15	15
2	1-30	Listening Comprehension			30	15	15
		A Conversation	Objective	10			
		B Passage	Objective	10			
		C News	Objective	10			
3	31-50	Cloze	Objective	20	20	10	15
4	51-80	Grammar and Vocabulary	Objective	30	30	15	15
5	81-100	Reading Comprehension	Objective	20	20	20	25
6		Writing					
		An Essay	Subjective	1	15	15	35
		B Note	Subjective	1	10	10	10
Total	100			103	140	100	130

APPENDIX B

Informed Consent Form

You are invited to participate in a research project conducted by Xiaowan Zhang from the Department of Linguistics of the University of Illinois. It investigates the relationship between students' performance on the Test for English Majors-Band 4 (TEM4) and their approaches of preparing for that test. You are being asked to participate in this experiment because you are a sophomore student from the English Department of Wuhan University and are eligible to take the TEM4. You must be at least 18 years old to participate in this study.

You will be randomly assigned to either a treatment group or a control group. If you are in the control group, you will be asked to complete two short surveys regarding the TEM4. The first survey is for you to report the progress of your preparation for the TEM4. You are going to answer questions regarding the materials that you currently use for preparing for the test, the average time you spend every week on preparation, and other test preparation-relevant issues. This survey will take no more than 10 minutes and will be administered about one month prior to the test. The second survey evaluates your knowledge of the TEM4. This survey will take no more 15 minutes and will be administered a week prior to the test. You will take both surveys via Google forms and will be able to access them with the given URL.

If you are in the treatment group, you will be asked to take one tutorial session on the TEM4 and a third survey in addition to the completion of the two surveys mentioned earlier. The tutorial session on the TEM4 will be open immediately after you take the first survey. You will be asked to sit through the session to get familiar with the TEM4, and the session lasts about 40 minutes. The third survey will be sent to you via email after you receive your TEM4 scores; it aims to seek your feedback on the tutorial sessions on TEM4 and will take no more than 10 minutes.

All surveys used in this study will be provided to you in English and Chinese. More detailed instructions on how to answer the questions can be found in the survey.

For both control and treatment groups, your GPA and TEM4 score will be requested from the Academic Affairs Office of the English Department at Wuhan University. However, no personal identifiers will be used here and in the rest of the study. Rather, all of your information will be entered into the database under an assigned a subject number, which will be the only identifier.

Your survey responses and all other information collected in this experiment (including your GPA and TEM4 score) will be saved on the researcher's working computer, to which only the researcher will have access. The results will be analyzed and might be published. However, your name and identity will not be associated with any types of data. All the raw data will be permanently deleted from any device after the research is done. The researcher will keep the information that you provide confidential. However, we cannot guarantee the same for the service hosting the surveys. Google or your email service provider may have access to the data you submit and to the IP number of the computer on which you complete this test. In order to minimize the risk of your information being shared, please avoid writing your name on any of the surveys and just use the subject number designated to you.

There is no significant risk in participating in this research except that it might slightly increase feelings of anxiety towards taking the TEM4. However, the performance of the treatment group may benefit from the tutorial session, and both groups might become more motivated in preparing for the test, which conceivably will result in better test performance. Moreover, we hope the results will shed light on the nature of the test-preparation process, which may help future test takers.

You will be compensated with an Amazon.cn gift card (worth about $5) for your participation. In order to be qualified for getting the compensation, you must submit the two surveys, and the treatment group members should also attend the tutorial session and finish the third survey. Your head teacher will give you the compensation right after you finish this research. To participate in this research is completely free and voluntary. You may discontinue at any time without any penalty or loss of benefits to which you are otherwise entitled. The decision to participate, decline, or withdraw from participation will have no effect on your grades at, status at, or future relations with Wuhan University.

By clicking the *I agree* button below, you acknowledge that you have read and understand the above consent form and voluntarily agree to participate in this study.

If you have any questions or would like to find out later about the results of the study, you can contact Xiaowan Zhang, at 1-217-819-7913, or at xzhng124@illinois.edu. You may also contact Dr. Randall Sadler, at rsadler@illinois.edu. If you have questions about your rights as a participant in this research, you can also contact the University of Illinois Institutional Review Board in the United States at 1-217-333-2670, or at irb@illinois.edu. The IRB may be called collect.

APPENDIX C

A Survey on the Tem-Band 4 Preparation Status
关于英语专业四级备考状况的调查

Thank you very much for participating in this survey! You are going to answer 8 questions regarding your preparation progress of the TEM- Band 4. Several of them are related to your experience of using the Test Syllabus of the TEM-Band 4. Please answer each question according to your actual test preparation experience with the TEM-Band 4. All the information in this survey will be kept absolutely confidential by the researcher.

非常感谢您参与这项调查！您将会在本调查中回答 8 个关于专四备考状况的问题。其中的一些是询问您使用考试大纲备考的经历的。请根据您的真实备考情况回答以下的每一个问题。您在本项调查中提供的所有信息我们将绝对保密。

1. How do you prepare for the TEM-Band 4 (please select all that apply)
   你如何备考专业四级（请选择所有符合实际情况的选项）
   A) Practicing published tests
   做真题
   B) Doing practice tests
   做模拟题
   C) Reading the published Test Syllabus
   阅读考试大纲
   D) Taking commercial coaching courses
   上辅导课
   E) Receiving private tutoring
   请家教
   F) Others ______________________________(please specify)
   其他__________________________________(请说明）

2. Of the choices above in Question 1, please indicate the ONE that you use most often to prepare for the test.
   在第一题的所有选项中，请选出一个你最常用的备考方式。
   A) Practicing published tests
   做真题
   B) Doing practice tests
   做模拟题
   C) Reading the published official Test Syllabus
   阅读考试大纲
   D) Taking commercial coaching courses
   上辅导课
   E) Receiving private tutoring
   请家教
   F) Others ______________________________(please specify)

其他 ____________________________ （请说明）

3. Have you read the published official Test Syllabus?
   你读过官方发行的专四考试大纲吗？
   A) Yes
      是

   B) No
      不是

   * If the answer is yes, how much more knowledge have you gained about the test, compared to before you read the official Test Syllabus.
   如果你的答案是肯定的，那么你觉得阅读大纲的收获有多大呢？
   A) Little
      很少
   B) Some
      一些
   C) Most
      很多
   D) Everything
      我找到了我需要的所有信息

4. Do you think the official Test Syllabus will help you better prepare for the test?
   你觉得官方的考试大纲对于你备考会有用吗？
   A) Yes
      有用
   B) Not sure
      不确定
   C) No
      没用

5. Does the head-teacher of your class give you instructions on the TEM-Band 4? If s/he does, approximately how many class hours are spent each week on the test?
   你的班主任有在课上给你做专四复习吗？如果有，大概每周有多少课时用于此呢？
   A) Yes ______________________________(please specify the hours)
      有 __________________________ （请给出具体小时）
   B) No
      没有

6. How many hours do you spend each week on preparing for the test?
   你每周要花多少小时复习专四呢？
   ________________________________ hours/小时

7. Please use the scale below to indicate the importance that you feel the TEM-Band 4 is to your future. (1 stands for ‘not important at all’ and 5 ‘very important’)

请用下面的量度尺来说明你认为专四的成绩对于你今后的发展有多重要。
（1代表”根本不重要”，5代表”非常重要”）

1	2	3	4	5

8. Please use the scale below to indicate how motivated you are to prepare for the TEM-Band 4. (1 stands for ‘not motivated at all’ and 5 ‘highly motivated’)
请用下面的量度尺来说明你备考专四的积极程度。
（1代表”根本不积极”，5代表”非常积极”）

1	2	3	4	5

APPENDIX D

A Survey on Meta-Knowledge of the Official Tem4 Syllabus
关于英语专业四级考试大纲元知识的调查

Thank you very much for participating in this survey! This survey is intended to measure how accurate your understanding of the TEM4 Syllabus is, for example, the test format, the number of test items, and the scoring rubrics. Please answer the following questions according to your current test preparation experience with the TEM-Band 4. All the information in this survey will be kept absolutely confidential by the researcher.

非常感谢您参加这项调查！本调查旨在测试您对于专四考试了解的准确程度，例如，专四考试的形式，考试题目的数量，以及评分标准。请您根据自己的真实备考经验来回答本调查中提出的问题。您在本调查中提供的所有信息我们将绝对保密。

General Questions: Please select only one answer.
概括性问题：请您选择一个正确答案。

1. How many parts are included in the TEM-Band 4?
   专四考试一共有几个部分？
   a) 4
   b) 5
   c) 6
   d) 7

2. Which of the following part is in the TEM-Band 4?
   以下哪一个选项是专四考试的一部分？
   a) Grammar and Vocabulary
      语法与词汇
   b) Proofreading
      改错
   c) General knowledge
      常识
   d) Translation
      翻译

Dictation: Please select only one answer.
听写部分：请您选择一个正确答案。

3. What kind of dictation test does the TEM-Band 4 have?
   专四考试中的听写部分是怎样的？
   a) Word-for-word
      逐字听写
   b) Free but meaning-based
      建立在意思基础上的自由听写

4. How many times is the dictation passage read?

听写的文章一共阅读几遍？

a) 2
b) 3
c) 4
d) 5

5. For each time of the reading, is the speed the same or different?
   每一遍阅读的速度都是相同的吗？
   a) Same
      相同
   b) Different
      不同

6. Is it possible for you to get a lower score because of punctuation errors?
   你可能因为标点错误在听写部分被扣分吗？
   a) Yes
      是的
   b) No
      不是

Listening Comprehension: Please select only one answer.
听力理解部分：请您选择一个正确答案。

7. How many sections is this part comprised of?
   听力理解一共有几个部分组成？
   a) 1
   b) 2
   c) 3
   d) 4

8. Which of the following is NOT a section in this part?
   以下哪一个选项不是听力理解的一部分？
   a) Conversation
      对话
   b) Monologue
      独白
   c) News Broadcast
      新闻
   d) Lecture
      讲课

9. Where are the news excerpts selected?
   新闻是从哪里节选的？
   a) VOA and BBC
      VOA 和 BBC

b) CNN and BBC
CNN 和 BBC
c) VOA and CNN
VOA 和 CNN
d) NPR and BBC
NPR 和 BBC

10. Is American English only tested in this part of the test?
是不是只有美式英语才在听力中出现？
a) Yes
是
b) No
否

Cloze: Please select only one answer.
完型填空：请您选择一个正确答案。

11. What is NOT tested in this part?
以下哪一个选项没有在完型填空中考到？
a) The use of vocabulary
词汇的运用
b) Phrases
短语
c) Collocations
固定搭配
d) Spelling
拼写

Reading Comprehension: Please select only one answer.
阅读理解：请您选择一个正确答案。

12. Will excerpts of unabridged literature work be possibly used in this part of the test?
未删节的文学作品有可能在阅读理解中出现吗？
a) Yes
有
b) No
没有

13. Is it possible for you to come across personal bibliography in this part of the test?
名人自传有可能在阅读理解中出现吗？
a) Yes
有
b) No
没有

14. Is it true that instructions and charts could appear in this part of the test?

说明书和图表有可能在阅读理解中出现吗？

a) Yes
   有
b) No
   没有

Writing: Please select only one answer.
写作：请您选择一个正确答案。

15. What is the desirable length of the essay?
    短文的理想长度是？
    a) Around 100 words
       100 词左右
    b) Around 200 words
       200 词左右
    c) Around 300 words
       300 词左右
    d) The longer the better
       越长越好

16. Could you be asked to write an exposition in the section of essay writing?
    你可能会被要求在本部分写一篇说明文吗？
    a) Yes
       是
    b) No
       不是

17. How long should be note be?
    留言条要写多长？
    a) Around 10 words
       10 个词左右
    b) Around 50 words
       50 词左右
    c) Around 100 words
       100 词左右
    d) The shorter the better
       越短越好

APPENDIX E

A Survey on the Effects of the Tutorial Session
关于专四大纲讲座的调查

Thank you for participating in the study on the TEM4 conducted by Xiaowan Zhang from the University of Illinois at Urbana-Champaign. This is a post-test survey that aims to get to know more about your test-preparation and test-taking experiences. You had been given a tutorial session on the TEM4 Syllabus about one month before the test. Would you please share with us how you feel about that tutorial session and how useful you think the content of the TEM4 Syllabus has been for your test preparation?
Your response to this survey is completely voluntary. Thank you very much for taking time to answer the questions below. Your response is very important to this research. You can respond in either Chinese or English.

非常感谢您参与由张晓菀 (伊利诺伊大学香槟分校) 负责的关于英语专四考试的研究。您现在所收到的这项调查是本研究的一部分，旨在更全面地了解您的备考经历。您大概考前的一个月参与了我们组织的关于专四大纲的复习讲座。能不能请您分享一下您对这个讲座的评价？另外，您认为专四大纲的内容对您的备考是否有帮助？
您对于本项调查的回答完全自愿。非常感谢您花时间填写下面的问题。您的答案将对我们的研究非常重要。您可以用中文，或者英文来填写这项问卷。

1. Do you think the content of the TEM4 Syllabus was helpful for your preparation for the test?
您认为专四大纲的内容对您备考有帮助吗？

2. If your answer to the first question is positive, in what aspects do you think it has been helpful? Please specify.
如果您对于第一道题的答案是肯定的，那么，您能否具体说明专四大纲对您的帮助体现在哪些方面？

3. If you had not been given the tutorial session, do you think your test score would have been better, worse, or about the same?
如果您没有参与我们的讲座，您觉得您的专四成绩会比实际成绩更好、更差、还是差不多?

4. Compared with other means of test preparation you used, do you think receiving the tutorial session on the TEM4 Syllabus was more beneficial to your overall English learning? Why?
与您使用的其他备考方式相比，您觉得了解考试大纲对您的英语学习更有益吗？为什么？

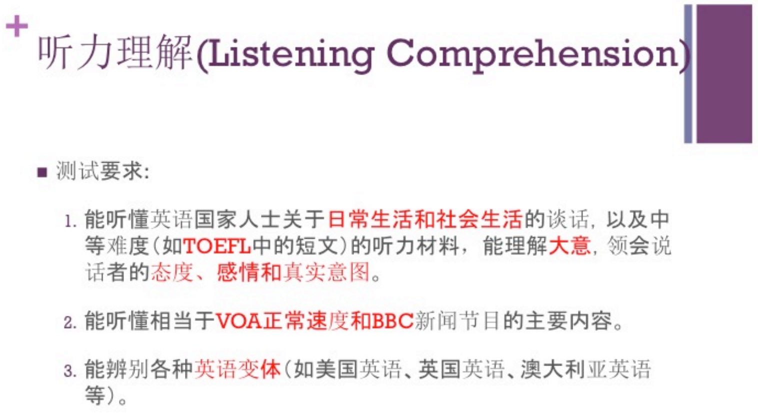
APPENDIX F

Screenshots of Some Material Used for the Tutorial Session

TEM-Band 4
英语专业四级考试

Xiaowan Zhang
张晓菀

Figure 5. The title page of the PowerPoint document.



听力理解(Listening Comprehension)

- 测试要求:

1. 能听懂英语国家人士关于日常生活和社会生活的谈话，以及中等难度(如TOEFL中的短文)的听力材料，能理解大意，领会说话者的态度、感情和真实意图。
2. 能听懂相当于VOA正常速度和BBC新闻节目的主要内容。
3. 能辨别各种英语变体(如美国英语、英国英语、澳大利亚英语等)。

Figure 6. One slide used to explain the section on listening comprehension.