Chapter 2
The Development of the Test of Chinese as a Foreign Language (TOCFL)

Li-ping Chang

Abstract This chapter describes the development of the Test of Chinese as a Foreign Language (TOCFL), which is a proficiency assessment tool for Chinese learners. It is divided into four major sections. First, a brief background of Teaching Chinese as a Second Language in Taiwan is provided together with the historical development of the Chinese Proficiency Test (CPT) and the Test of Proficiency-Huayu (TOP), which were the predecessors of the TOCFL. The second section discusses issues that the TOCFL research team faced in its effort to map the test to the Common European Framework of Reference (CEFR). The third section then discusses four challenging issues that the TOCFL research team has faced over the years in its effort to develop and maintain a standardized CSL/CFL test. The final section presents some washback effects of the TOCFL and work in progress.

Keywords Proficiency test · Chinese as a second/foreign language · CSL/CFL testing · TOCFL · CEFR

Introduction

Teaching Chinese as a Second Language (TCSL) began in the 1950s in Taiwan as a branch of the Yale University system of teaching Mandarin Chinese. At that time, most Mandarin Chinese centers used textbooks from the Yale University series, for example, Tewksbury's *Speak Chinese* (1948), and offered small-group classes of two to three people for two hours per day, five days per week. In response to the global increase in the study of Chinese language, in 1995, National Taiwan Normal University (NTNU) founded the first graduate institute to offer an MA program in the field of TCSL. As TCSL slowly but steadily became an important academic discipline, a test designed for Chinese as a Second Language was necessary. In 2001, some professionals and scholars from the Mandarin Training Center (MTC),

L. Chang (✉)
National Taiwan Normal University, Taipei, Taiwan
e-mail: lchang@ntnu.edu.tw

D. Zhang and C.-H. Lin (eds.), *Chinese as a Second Language Assessment*,
Chinese Language Learning Sciences, DOI 10.1007/978-981-10-4089-4_2

the NTNU Graduate Institute of TCSL, and the NTNU Research Center for Psychological and Educational Testing (RCPET) formed a research team that started to construct a Chinese Proficiency Test referred to as the CPT (華語文能力測驗), which later became what is now the Test of Chinese as a Foreign Language (TOCFL).

This chapter focuses primarily on the early development of the TOCFL from 2001 to 2011, the 10-year period that witnessed the test's advancement from the embryonic stage to the forming stage. It is divided into four sections. First, a brief overview is provided on the historical development of the Chinese Proficiency Test (CPT) and the Test of Proficiency-Huayu (TOP), which were the predecessors of the TOCFL. The second section discusses issues that the TOCFL research team faced in its effort to map the test to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). The third section then discusses four challenging issues that the research team has faced over the years in developing and maintaining a standardized test of Chinese as a Second/Foreign Language (CSL/CFL). The fourth section presents some washback effects of TOCFL and work in progress. We believe that the issues discussed in this chapter will be useful to others who are interested in either the TOCFL or CSL/CFL test development and research in general.

Overview of Historical Development

Literature Review

To develop and deliver a high-quality proficiency test, the CPT research team started by searching for previous literature on testing Chinese as a Second/Foreign Language. We found that the body of research in this area was very limited. In Taiwan, Ko and her team pioneered in research in this area. For example, Ko and Chang (1996) and Ko et al. (1995, 1997) conducted some experiments with a few test item types and task types to develop a Test of Chinese as a Second Language. However, there were several limitations with those studies in that (a) only one level of the test was developed, (b) some test tasks seemed to be too difficult for foreign language learners, and (c) most pilot samples were Chinese native speakers studying in local elementary schools rather than adult foreign language learners. Therefore, other than gathering some ideas about test item types and task types, the CPT research team was not able to follow the line of research conducted by Ko and her team.

In addition to searching the literature, the CPT research team also investigated existing Chinese proficiency tests created in other countries (Chang 2002), such as the *Hànyǔ Shuǐpīng Kǎoshì* (HSK; 漢語水平考試) in mainland China (HSK Test Center) (Liu 1997; Teng, this volume), the *Zhōngguóyǔ Jiǎndìng Shìyàn* (中國語檢定試驗) in Japan (The Society for Testing Chinese Proficiency), the Scholastic Aptitude Test—Chinese (SAT II—Chinese) (The College Board; Liu, this volume), the Chinese Proficiency Test (CPT) and the Preliminary Chinese Proficiency Test (Pre-CPT) developed by the Center for Applied Linguistics (CAL) in the USA. The investigation revealed the following 5 points:

1. Most proficiency tests set the target examinees at above 18 years of age.
2. Except for the HSK's cloze section of the basic/intermediate level, which included 16 items that required examinees to write Chinese characters, and the test administrated in Japan, which included a translation task and a sentence writing task, most proficiency tests used multiple-choice questions in reading and listening tests.
3. All proficiency tests investigated emphasized grammar structures either by testing grammar as one separate part such as the grammar section of SAT II—Chinese or by including grammar items in the reading section such as the reading section of CAL's CPT.
4. With the exception of the HSK, which used only Chinese in the test booklet, the tests developed in Japan and the USA used Japanese and English (the test takers' native languages), respectively, for test instructions, multiple-choice options, and some test questions.
5. It seems that each test defines the notion of proficiency differently, making it hard for the research team to interpret the tests' results in relation to one another. That means, for example, level one of one test is not equal to that level of another test.

The First Version of Chinese Proficiency Test: Listening and Reading Tests

After the review of literature, the research team created the first version of the CPT. To meet learners' demands and in consideration of the limited human and financial resources available at that time, the research team focused solely on designing tests of listening and reading comprehension targeting learners who had studied Chinese for more than 360 h (approximately 9 months) in Taiwan. To help with the creation of different test levels, the research team decided to use the Chinese proficiency guidelines developed by the American Council on the Teaching of Foreign Languages (ACTFL 1987). The ACTFL guidelines were used because they were the only international framework for Chinese teaching and testing available in Taiwan at that time, and more importantly, ACTFL's five major levels of proficiency—novice, intermediate, advanced, superior, and distinguished—were readily accepted by local teaching specialists.[1] The advantages of the ACTFL guidelines

[1] The distinguished level did not exist in speaking and writing guidelines until 2012. For detailed information, please visit the Website: http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012.

were that they spelled out linguistic characteristics for each level of four language skills: listening, speaking, reading, and writing. For example, a description of the intermediate-high level of listening proficiency stated that the candidate at this level can "understand major syntactic constructions, e.g., *Shì-de* (是…的)-focus, *bǎ* (把)-disposal, and *bèi* (被)-passive" but may still make mistakes with "more complex patterns, e.g., *jiào/ràng* (叫/讓)-passive, *lián* (連)-emphasis/contrast, *chúle* (除了)-exclusion/inclusion" (ACTFL 1987, p. 477). Although these structures are mainly defined by the authority of scholars rather than empirically validated (De Jong 1990; Liskin-Gasparro 2003), the concrete descriptions were helpful in setting the levels for the tests.

Based on the ACTFL scales, the research team created different forms of listening and reading tests covering three levels: basic, intermediate, and advanced, which were designed to reflect the novice-high, intermediate-high, and advanced-low to superior levels of the ACTFL scales. The structure of the three levels of CPT is provided in Table 2.1. As Table 2.1 shows, all levels of CPT included three sections: listening, vocabulary/grammar, and reading. The time allocation for each section was equal across proficiency levels, and the task types were almost the same except for the reading section. These task types reflected the influence of other Chinese proficiency tests investigated by the CPT research team, as reported earlier. In addition to the time allocation and test content, Table 2.1 also includes short descriptions that were shown to prospective test takers indicating the hours of Chinese instruction that they should have received as well as the amount of vocabulary that they should have learned prior to attempting the test.

The Pilot Tests

The pilot tests of the CPT took place during 2002–2003 with a total of 2160 listening and reading test items. Approximately 1000 foreign learners of Chinese in language centers across Taiwan took the pilot tests and completed questionnaires. The aim was to test the suitability of task types and item difficulty. After two years of pilot testing, the research team gathered sufficient evidence to support the division of the test into three proficiency levels (basic, intermediate, and advanced) and the use of the task types listed in Table 2.1. Subsequently, the listening and reading tests, which were comprised of multiple-choice items in a paper-based format, were administered semiannually in Taiwan starting in December 2003. This transformed the CSL situation since prior to 2003 students of CSL in Taiwan had no common tool to measure their proficiency; instead, they had to rely on grades from the various language training centers and documentation of the time spent studying the language.

Table 2.1 Content breakdown of the CPT listening and reading tests

Levels/format	Test of basic	Test of intermediate	Test of advanced
Listening	Sentence (20) Conversation (20) Paragraph (10)	Sentence (15) Conversation (20) Paragraph (15)	Sentence (15) Conversation (20) Paragraph (15)
Vocabulary and grammar	Vocabulary (20) Grammar (20)	Vocabulary (10) Grammar (20)	Vocabulary (20) Grammar (10)
Reading	Sentence (10) Authentic material (20)	Sentence (10) Authentic material (10) Passage (20)	Sentence (10) Passage (20)
Suggested learning hours	360–480 h	480–960 h	Over 960 h
Vocabulary base	1500 words	5000 words	8000 words
Others	1. There are 120 test items in each level. All items are multiple-choice questions 2. The approximate test time is 110 min, with 40 min for the listening test and 70 min for the reading test		

Note The Arabic numerals in parentheses indicate the number of test items. The suggested learning hours have to be doubled for test takers from overseas (i.e., not studying in Taiwan)

The Steering Committee for the Test of Proficiency—Huayu

In 2005, the Taiwan Ministry of Education (MOE) entrusted NTNU with a mission to organize a testing center that promotes an effective CSL assessment system. As a result, the CPT project was funded by the MOE. While its Chinese name remains the same (i.e., 華語文能力測驗), the CPT's English name was changed to Test of Proficiency-Huayu (TOP).[2] The Steering Committee for the Test of Proficiency—Huayu (SC-TOP) (國家華語測驗推動工作委員會) was created to oversee test development and validation.

In 2007, the Taiwan Ministry of Education mandated that the SC-TOP evaluates the possibilities of mapping TOP test results to the CEFR proficiency scale (Council of Europe 2001) (See the next section). The research team began to reexamine and modify the test format and content of the original three levels of the test to meet this challenge. In this revision of the tests, the basic, intermediate, and advanced tests were renamed into Level 3, Level 4, and Level 5, and corresponded to CEFR levels B1, B2, and C1, respectively. TOP was subsequently renamed TOCFL (in 2011) after this revision. In the meantime, the writing and speaking tests were launched. And furthermore, all TOP/TOCFL tests administered in Taiwan since then became computer-based instead of paper-based. Since 2011, the (new) TOCFL test is delivered via the Internet with automatic scoring so that test takers can get their test

[2] "Huayu" is a Taiwanese reference to Chinese as a Second/Foreign Language. For more information, please visit the Website: http://www.sc-top.org.tw.

results right away. In addition, all levels corresponding to CEFR (six levels) were finished in 2013. Participants who pass the test are given a certificate indicating their level of Chinese proficiency, which is used when applying for Taiwanese scholarships and university enrollment.

Mapping the (Old) TOP to the CEFR Can-Do Statements

This section discusses the effort of the SC-TOP on mapping the TOP/TOCFL with the CEFR, and some issues the research team faced during the mapping process. According to Council of Europe (2001), the CEFR is a common framework created to provide an extensive and transparent reference for language users and for teaching and testing specialists to communicate what kind and what degree of language knowledge is certified through a particular examination result, diploma, or certificate. The CEFR adopts the *can-do* statements that describe in general terms what language learners can typically do with a language at different levels of ability. It also raises key issues for the user to consider what language learners have to learn to do in order to use a language for communication. The CEFR levels are presented in three broad categories, A, B and C, with six levels: A1 and A2 for "basic" users; B1 and B2 for "independent" users; and C1 and C2 for "proficient" users. These levels are related to the three meta-categories: communicative activities, strategies employed in performing communicative activities, and aspects of linguistic, pragmatic, and sociolinguistic competence (Council of Europe 2001).

The Steps of Mapping

The first step that the research team took to explore the possibility of mapping the TOP results to CEFR was to follow the draft manual on relating examinations to the CEFR (Council of Europe 2003; Figueras et al. 2005). The research team followed the first three steps suggested in the draft manual: familiarization, specification of examination content, and standardization of judgments (Figueras et al. 2005). First, a basic-level TOP test paper was selected (refer to Table 2.1 for the components of the test). Next, five research team members divided all test items (i.e., 120 items) into 14 communication themes in the CEFR framework (Council of Europe 2001). Then, the team members discussed what communication activity each test item involves (e.g., whether a listening item involves listening to public announcements or listening to media or listening as a member of a live audience). Finally, the team members judged which CEFR level a test taker would need to have in order to answer each item correctly. Results showed that the TOP basic-level test items mostly belong to the CEFR B1 level (86.67%), with less items from the A2 level (12.5%) and B2 level (0.83%) (Gao et al. 2007).

The Difficulty of Mapping

Despite the high percentage of items belonging to the B1 level, the research team members found that judging which level is needed to complete an item was not straightforward. For example, two items may be identified as involving the personal domain, such as travel or daily life, which is mentioned by both the CEFR A2 and B1 overall listening comprehension scales. The A2 descriptor states the following: "Can understand phrases and expressions related to areas of most immediate priority (e.g., very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated." The B1 descriptor states the following: "Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc., including short narratives" (Council of Europe 2001, p. 66). However, a listening test item that talks about negotiating apartment rental fees with the landlord or interviewing for a job would be harder to understand than another listening test item that simply talks about making a hotel reservation. Therefore, the research team had to create its own criterion and specify that the ability to handle short-term learning or traveling experience would belong to the A2 level and the ability to handle long-term learning or traveling experience (e.g., living abroad) would belong to the B1 level. The draft manual expects test developers to decide by themselves what contextual parameters are appropriate at different levels on the CEFR scales (Weir 2005). This dependence on subjective judgment is a two-edged sword. The negative side is that it poses some difficulty when a test item is borderline between two CEFR levels, but the positive side is that it requires the research team to be very explicit in its distinction of the levels of ability that are required by the test items. This requirement for a clear distinction is useful for future item development.

In addition to the problem judging test items' CEFR levels, another problem that the research team faced in its attempt to link the TOP to the CEFR is related to the difficulty levels of the can-do statements. In order to solve the issue, in 2008, the research team created a self-assessment questionnaire of can-do descriptors based on the European Language Portfolio (ELP) (Council of Europe 2004) for 1500 students learning Mandarin Chinese at MTC, NTNU (SC-TOP 2009a). These statements included five communicative activities covered by the CEFR: listening, reading, spoken interaction, spoken production, and written production.[3] The first purpose of the study was to better understand the relationship between the test takers' language learning backgrounds (e.g., learning hours, teaching materials) and their CEFR levels including listening, reading, speaking, and writing skills. The second purpose was that we wanted to familiarize CSL teachers and learners in Taiwan with the CEFR can-do statements. The questionnaire was translated into 12 languages to ensure that all students would be able to read the statements easily. The students were divided into four groups according to the Mandarin courses they

[3] The speaking and writing tests had not yet been completed, but they were in the stage of pretest at that time.

were studying (i.e., their proficiency[4]). The first group received a Level 1 questionnaire that was composed of 68 statements targeting CEFR A1 and A2 levels. The second group received a Level 2 questionnaire that was composed of 97 statements targeting CEFR A2 and B1 levels. The third group received a Level 3 questionnaire that was composed of 72 statements targeting CEFR B1 and B2 levels. Finally, the fourth group received a Level 4 questionnaire that was composed of 40 statements targeting CEFR B2 and C1 levels. Statements that targeted the same CEFR level sometimes appeared in two questionnaires (e.g., the A2 statements in Level 1 and Level 2 questionnaires) so that they would serve as anchor items to be analyzed by the Item Response Theory (IRT) method.

IRT analysis showed that item difficulty levels of some statements did not necessarily follow the expectation in the ELP (Chang 2010). For example, the statement "I can say what I like and dislike" (Council of Europe 2004, p. 60), which was labeled in the ELP as belonging to the A2 level of spoken interaction, turned out to be much easier (difficulty value = −5.736) than a statement such as "I can indicate time by such phrases as 'next week', 'last Friday', 'in November', 'three o'clock'" (Council of Europe 2004, p. 76), which was labeled in the ELP as belonging to the A1 level (difficulty value = −2.012). Chang (2010) speculated that this result was caused by the large discrepancy between time expressions in Chinese and European languages. To express the specific time, Chinese uses prepositions like *zài* (在) "in" or determiners such as *shàng* (上) "last" and *xià* (下) "next," or localizers like *yǐqián* (以前) "before" or *yǐhòu* (以後) "after"—a complicated list of options for CSL learners. On the other hand, it is easier for learners to express "like" or "dislike" in Chinese by simply using these two verbs without worrying about tense and subject-verb agreement. The results of this questionnaire study were very useful in creating CSL course objectives. More importantly, we learned from this experience that the difficulty level of the ELP can-do statements would need to be empirically tested with CSL learners in Taiwan before they could be used for TOP item development.

The Results of Mapping: The Modified Version of TOP

The experience of mapping each TOP test item to the CEFR scale made the research team realize that the old version of TOP needed to change. The old version of TOP/CPT included many context-free vocabulary and grammar items. The sentence tasks in the reading and listening sections were context-free as shown in Table 2.1. These context-free items posed a difficulty when the research members tried to identify the communication activity that each item involved—a necessary

[4] If we had not divided the students into four groups, every student would have to answer all the statements, the number of which was huge. Our expert knowledge helped us predict which statements would be out of the range of the proficiency of a student; on this ground, grouping was made possbile.

Table 2.2 Content breakdown of the modified TOP listening and reading tests

Levels/format	Test of level 3 (B1)	Test of level 4 (B2)	Test of level 5 (C1)
Listening	Short conversation: Single-round (20) and multiple-round (15) Paragraph (15)	Short conversation (20) Long conversation (15) Paragraph (15)	Short conversation (10) Long conversation (20) Paragraph (20)
Reading	Cloze (20) Authentic material (15) Short essay (15)	Cloze (15) Authentic material (10) Short essay (25)	Cloze (15) Short essay (35)
Other	1. There are 100 test items in each level. All items are multiple-choice questions 2. The approximate test time is 110 min, with 50 min for the listening test and 60 min for the reading test		

Note The Arabic numerals in parentheses indicate the number of test items

step in mapping a test item to the CEFR (Council of Europe 2003; Figueras et al. 2005). It was decided that a new version of TOP would no longer have the vocabulary/grammar section. Instead, grammar, vocabulary, and reading are tested in the context of a cloze test, which is part of the new reading section. A communication theme can first be identified before a cloze passage is written to test grammar and vocabulary. Hung (2009) has studied the design of cloze tests for Chinese reading and suggested that the blanks can be designed based on Bachman's (1985) categorization: (1) within clause, (2) across clause within sentence, (3) across sentences within text, and (4) extratextual. Hung also suggested that the new TOP that corresponds to the CEFR B1 level would include items from categories (1) and (2), the B2 level would include items from categories (2) and (3), and the C1 level would include items from category (3). Her further study has verified the above suggestion (Hung 2013).

The single-sentence task is also removed from the new TOP listening section. The conversation task is expanded to include three item types: one-turn conversation, two-turn conversation, and longer conversation. These three item types provide more contexts to assist with the identification of CEFR communication activity. The content of the modified version (TOP) is presented in Table 2.2. The major difference between Tables 2.1 and 2.2 is the test items designed within broader language contexts instead of in a single sentence in the listening and reading tests.

Empirical Studies

After modifying the test, the research team began to conduct empirical studies to validate this modified version (i.e., new TOP). Chang's (2009) study is one of the first efforts. In her study, 55 CSL students at the Mandarin Training Center at NTNU took the listening section of the new TOP that was designed to correspond to the B1 level; and 56 students took the new B1 reading section. Simultaneously, 127 CSL students took the new B2 listening section and 126 students took the new

B2 reading section. The students who took the listening and reading sections at the same level were mostly the same people, but some students took only one section due to absence. These students' ability levels were also evaluated by 45 CSL teachers. The Spearman's rank correlation of the teachers' ratings and the students' performance on the B1 listening test was moderate (0.553). On the other hand, the correlation of the teachers' ratings and the students' performance on the B1 reading test was relatively low (0.296). The correlations of the teachers' ratings and the students' performance on the B2 listening and reading tests were also moderate and low, respectively (0.389 and 0.262). Thus, from the results of this study, it seemed that the B1 and B2 listening tests functioned a bit better than the reading tests, but both tests still required revision.

The results of K.C. Chang's study also informed the team that there are limitations when teachers' ratings are used to compare test results as teachers may not be able to evaluate the same constructs being tested by the new tests. In other words, unless classroom activities include the same tasks that are on the listening and reading tests, teachers may not be the best judge of the extent to which students can perform these tasks. In fact, in another study by a SC-TOP researcher that compared students' self-assessment with teachers' evaluation, Li (2009) found that students and teachers had many disagreements. For example, more than 40% disagreement was found in two descriptors: "Can understand the best part of a narrative passage or a well-structured contemporary literary text with a dictionary if need be" and "Can understand mail I receive well enough to be able to correspond regularly." Li speculated that this discrepancy resulted from the fact that some descriptors are hard for teachers to observe in class activities. Thus, although self-assessment and teachers' assessment are important data sources for a test validation project, they also have limitations. More training is needed for teachers to assess their students in terms of the CEFR.

Summary

To sum up, due to the need to map the test to the CEFR, we analyzed the TOP items using the CEFR framework and encountered several issues. In the process, we also realized that new test formats are needed to facilitate mapping. However, perhaps the most valuable gain for the research team in this whole experience is the chance to become familiar with the CEFR. Some scholars may question the idea of mapping the TOP to the CEFR because, after all, Chinese is not a European language and Taiwan is not a European country. A simple answer might be that a test development project funded by the government is part of a nation's language policy, which needs to be considered by test developers. On top of any consideration of that kind, the research team, however, also recognized the inherent importance of the work of mapping TOP to the CEFR, as we believed that it is important to have a common scale for CSL teaching and learning, and that it is useful for test users to be able to interpret different test results using the same scale.

On the other hand, we would like to argue that any effort to map tests to the CEFR will not be useful unless language teachers are well-informed about the CEFR itself. Without adequate knowledge of the CEFR, CSL teachers will not fully understand the test results or be able to use them in their teaching. The research team has conducted several workshops for CSL teachers familiarizing them with the CEFR, and over the years, it is clear to us that teachers appreciate the clarity of the ELP can-do statements and the CEFR domains of language use (Chou and Chang 2007). However, more efforts must be made to help teachers apply the CEFR and the test scores in their teaching.

Issues Related to the Listening and Reading Tests

Apart from the aforementioned issues of aligning the TOP test to the CEFR, over the years some other issues have been challenging to the development of the new TOP/TOCFL (thereafter, TOCFL). They include the training of test item writers, the research of word/grammar lists to help design the test items, the suggestion of learning hours for test takers, and the discrepancy in the use of Mandarin Chinese between mainland China and Taiwan. This section of the chapter will discuss each of these four issues.

Issue 1: Training of Test Item Writers

An important issue relating to the development of TOCFL involves the training of test item writers. In Taiwan, there are two types of Mandarin teachers: those who had training in CSL (referred to as “the L2 teachers”) and those who were trained to teach Mandarin Chinese to native Chinese speakers at elementary, junior high, and senior high schools (referred to as “the L1 teachers”). At the beginning of the CPT project, it was thought that any Chinese teachers can be trained to write standardized proficiency test items. However, our experience showed that L1 teachers were not the best item writers for a CSL test due to their limited experience teaching Chinese to non-native speakers. The main problem was that they had less sense of how to create useful distracters in a multiple-choice item. Take a vocabulary question as an example:

他右手 __________ 著手錶。

Tā yòushǒu ______ zhe shǒubiǎo.

He right hand wear ASP watch.

“He wore a watch on his right hand.”

Four possible options written by an experienced item writer would be (A) *dài* (帶) “bring,” (B) *dài* (戴) “wear,” (C) *chuān* (穿) “wear,” and (D) *dài* (袋) “bag.”

The correct answer is option (B), with options (A) and (D) serving as phonological distractions and option (C) serving as a semantic distraction (the word *chuān* (穿) "to wear" is used in Chinese when talking about clothes and shoes but the word *dài* (戴) "to wear" is used in Chinese when talking about hats and wristwatches). Native speakers of English might choose option (C), which would be an error caused by direct English-Chinese translation. This direct translation problem is well-known by CSL teachers, who could use this type of item as a diagnostic test. L1 teachers, on the other hand, were less intuitive about creating these distracters. This is why the research team decided to invite only CSL (L2) teachers to write items.

To qualify as a TOCFL item writer, a CSL teacher has to have at least two years of CSL teaching experience, be currently teaching CSL, and pass the item writing training. The first stage of training is a six-hour training course that familiarizes the CSL teachers with the test specifications. The teachers are then asked to write 10 items of a specific task type per week. These items are reviewed by the research team and an outside panel of experts. After a few weeks, qualified item writers are selected. To improve their item writing skills, after their items are used in a trial, the writers have meetings to discuss item qualities such as difficulty values, discrimination values, and distracter analysis results. The item writers then must revise their items for another round of trials.

While there are more things that we can do to improve item writers' training, we realize that the most challenging problem is not the training of item writers but how to keep a well-trained item writer to work for us on a long-term basis. The TOCFL project does not have adequate funding to hire full-time item writers. Our item writers are full-time CSL teachers who have their own teaching responsibilities while helping us part-time. Given the workload that the teachers face, the turnaround rate of our item writers is very high and we have to constantly train a new group of item writers. This is a real-world problem that the TOCFL project cannot avoid because of funding issues.

Issue 2: Word/Grammar List Used to Help Design the Items

The second issue also relates to item writing. To enable the TOCFL test item writers to distinguish different levels of test items, it is necessary to rank the difficulty levels of lexical and syntactic patterns in Mandarin. From August 2003 to July 2005, I conducted a two-year research project funded by Taiwan's National Science Council (NSC-92-2411-H-003-04; NSC-93-2411-H-003-066) to rank both the vocabulary and syntactic patterns of the Chinese language. The vocabulary pool was obtained from various sources, including the Academia Sinica Corpus (Chinese Knowledge and Information Processing Group [CKIP] 1998) that contains five million words, CSL textbooks often used by Taiwanese language centers and American colleges, and a vocabulary list constructed by the HSK office (Liu and Song 1992). The research team used the frequency information from the

CKIP (1998) and the weighting method to rank each word. Each word was tagged with its frequency and source (i.e., the corpus, the textbooks, or the HSK). If a word was listed in all three sources, it received greater weight. Chang and Chen (2006) reported the details of the procedures involved. Based on the frequency and the weight of the information, 1500 words with the highest weights were set for the basic level, an additional 3500 words were set for the intermediate level, and another 3000 words were set for the advanced learners (Chang and Chen 2006). These numbers are also listed in Table 2.1. The decisions on these numbers (1500; 5000; and 8000) were based on our surveys of the students' learning hours and teaching materials (Chang 2002). Cheng (1998) also suggested that knowledge of approximately 8000 words is needed to function in a Chinese academic setting.

With the development of the TOCFL, we need to reexamine this issue of the number of words needed to pass each test. Recently, Hanban (2010) suggested 600 words for a CEFR B1 proficiency level and 1200 words for a CEFR B2 level (see also Teng, this volume). On the other hand, the Fachverband Chinesisch e.V. (Association of Chinese Teachers in German Speaking Countries) (2010) has proposed a 2500-word threshold for the CEFR B1 proficiency level and a 5000-word threshold for the CEFR B2 level. The latest research was done by Chang (2012). It is a corpus-based study, and the results suggest the vocabulary size for A2 level is 1000; B1 level is 2300–3000; B2 level is 4500–5000; C level is 8000–10000.

Further research is needed to investigate whether the current setting of words will be adequate for students to pass the TOCFL tests that are linked to the CEFR.[5] One option is to recruit students who have completed the learning hours specified by the TOCFL. Those students' test performance as well as a survey of their learning hours and instruction materials will provide important information not only about the required number of words but also about the required minimum number of learning hours.

Issue 3: Suggestion of Learning Hours for Test Takers

The issue of a minimum number of learning hours required to pass each level of the TOCFL is an important one for test takers. From 2003 until the present, the number of test takers has increased gradually to more than 10,000 annually. So far, the test results can be used for various purposes. Foreign students wishing to study at Taiwanese universities can use the test results to apply for academic programs at Taiwanese universities either through the "Taiwan Scholarship (台灣獎學金)," a scholarship given by the Taiwanese government, or through their own funding

[5] The current setting of words is that A1 level is 500; A2 level is 1000; B1 level is 2500; B2 level is 5000; C1 level is 8000. The word lists can be retrieved freely from http://www.sc-top.org.tw/english/download.php (8000 Chinese words).

sources. Overseas students of the Chinese language can use the test results for University Entrance Committee for Overseas Chinese Students (海外聯招). The TOCFL certificates can also be used as a proof of Chinese language proficiency for employment.

As more students have the need to take the test, more and more language centers and teachers are aware of this test and are willing to use it to measure their students' abilities. To help students pass the TOCFL, some language centers have adjusted their curriculum design. When this occurs, CSL learners gradually become familiar with the relationship between performance on the TOCFL and their learning hours and learning materials. Chang and Chen (2008) analyzed the relationship between the results of the TOCFL and the learning hours that the students had taken at the Mandarin Training Center. They found that higher numbers of hours completed correlated positively with higher TOCFL scores. The students with around 360 h of learning can correctly answer 80% of the items in the TOCFL B1 level test; students with around 610 h can correctly answer 67% items in the B2 level test.

To check the relationship between the suggested learning hours and the pass rate, a chi-square test of association was conducted with survey data collected from TOCFL test takers during 2008–2009. Table 2.3 reports chi-square results of test takers from various Mandarin centers in Taiwan (N = 3505), and Table 2.4 reports the chi-square results of overseas test takers (N = 1229). Both showed that for all three levels, the pass rates of test takers whose learning hours were more than the hours suggested in Table 2.1 were significantly higher than for those whose learning hours were less than suggested. Table 2.3 shows that 47.4% of CSL learners in Taiwan who completed 360 learning hours passed the basic-level test while only 39.4% of learners who had not yet completed those hours passed the basic-level test. The percentages were 46.2% versus 38.8% at the intermediate level, and 67.9% versus 55.5% at the advanced level. For Chinese learners from overseas, the SC-TOP suggests that the minimum learning hours are doubled. Table 2.4 shows that 44.6% Chinese learners from overseas who had completed 720 learning hours passed the basic-level test while only 24.3% of learners who had not yet completed those hours passed the basic-level test. The percentages were 51.8% versus 37.1% at the intermediate level, and 84% versus 64.4% at the

Table 2.3 Learning hours versus pass rates for Chinese as a second language (2008–2009)

Suggested hours = 360, 480, 960	Basic (B1 level)		Intermediate (B2 level)		Advanced (C1 level)	
	N = 1321		N = 1492		N = 692	
	Failure	Pass	Failure	Pass	Failure	Pass
Less than suggested hours	341 (60.6%)	222 (39.4%)	296 (61.2%)	188 (38.8%)	133 (44.5%)	166 (55.5%)
More than suggested hours	399 (52.6%)	359 (47.4%)	542 (53.8%)	466 (46.2%)	126 (32.1%)	267 (67.9%)
χ^2	8.245**		7.248**		11.186**	

**p < 0.01

Table 2.4 Learning hours versus pass rates for Chinese as a foreign language (2008–2009)

Suggested hours = 720, 960, 1920	Basic (B1 level)		Intermediate (B2 level)		Advanced (C1 level)	
	N = 664		N = 274		N = 291	
	Failure	Pass	Failure	Pass	Failure	Pass
Less than suggested hours	321 (75.7%)	103 (24.3%)	78 (62.9%)	46 (37.1%)	99 (35.6%)	179 (64.4%)
More than suggested hours	448 (55.4%)	361 (44.6%)	289 (48.2%)	311 (51.8%)	77 (16.0%)	403 (84.0%)
χ^2	48.993**		8.928**		37.818**	

**$p < 0.01$

advanced level. These results provide supporting evidence for the suggested minimum learning hours presented in Table 2.1.

Despite the significant chi-square test results in Tables 2.3 and 2.4, a close look at the percentages indicates that less than 50% of the test takers who completed 360 h in Taiwan passed the TOP basic level, and less than 50% of the test takers who completed 480 h in Taiwan passed the TOP intermediate level. Similarly, less than 50% of the test takers from overseas who completed 720 h passed the TOP basic level. Chang (2011) further examined the suggested learning hours for learners of different L1s, including Japanese, Korean, and English. She found only the data of Japanese learners showed significance. Additionally, in Guder and Kupfer's (2005) report, the Association of Chinese Teachers in German Speaking Countries estimated that between 1200 and 1600 h of instruction (plus private study time) are required to attain oral and written proficiency in Chinese, which is comparable to CEFR level B2. Taken together, these findings suggest that the issue of suggested learning hours for test takers could be complex, and depends on individual differences, especially learners' native language background.

Issue 4: Mainland China Versus Taiwan's Use of Chinese

Another issue that the tests have to deal with is the discrepancy in the use of Mandarin Chinese between mainland China and Taiwan, which occurred due to the political separation in 1949. This discrepancy in language use has created several problems for language learning and testing alike (also see Shang and Zhao, this volume). The first problem is the fact that some high-frequency words have a variety of forms across the straits. For example, a taxicab is called *jìchèngchē* (計程車) (literal translation "mileage counting car") in Taiwan but *chūzūchē* (出租車) (literal translation "rental car") in China. A taxi driver is referred to as *sījī xiānshēng* (司機先生) (literal translation "Mr. Driver") in Taiwan but *shīfù* (師傅) (literal translation "master") in China. Computer software and hardware are called *ruǎntǐ/yìngtǐ* (軟體/硬體) (literal translation "soft/hard body") in Taiwan but they are called *ruǎnjiàn/yìngjiàn* (軟件/硬件) (literal translation "soft/hard device") in China. The list goes on. While this discrepancy may not cause problems for test takers who are above the intermediate proficiency level, who can guess the meaning of a word from its context, or for CSL learners in Taiwan, it often causes difficulties for basic-level overseas learners of Chinese because they may use textbooks published by mainland Chinese publishers. In most cases, the TOCFL uses words that are common in Taiwan. However, we also do our best to avoid idiosyncratic words that are used only in Taiwan. For example, we do not use the word *jiéyùn* (捷運) to refer to subway train even though this word is commonly used in Taiwan. Instead, we use the word *dìtiě* (地鐵), which is more commonly used in the Chinese-speaking world outside of Taiwan.

Another difference between mainland China and Taiwan lies in the use of Chinese characters. These days, learners may be exposed to either simplified characters used in mainland China or traditional characters used in Taiwan. The current strategy employed by SC-TOP is that when administering the TOCFL outside of Taiwan, test takers are allowed to choose either the simplified or traditional version of the test. However, this strategy is a post hoc method to deal with this practical issue and not a part of the initial test development plan. Therefore, it is essential to the SC-TOP to ensure that the different character versions are similar in difficulty. Using IRT, the team analyzed the difficulty of all items in the same advanced level for traditional and simplified character versions of the test that had been taken by more than 200 candidates and found a similar level of difficulty (see Fig. 2.1) (SC-TOP 2009b).

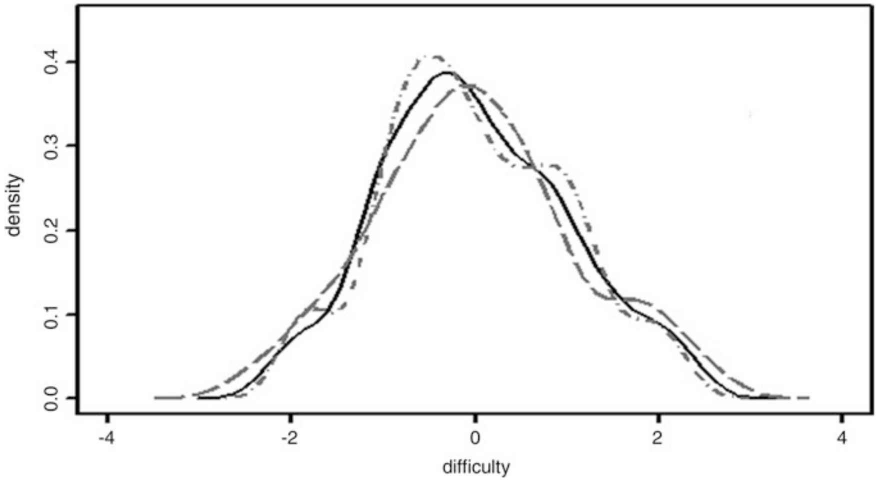

Fig. 2.1 Distribution of item difficulty in the reading section (67 items). *Note* ——— indicates the difficulty for all the testees, – - – indicates the difficulty of the traditional version for the testees, — — indicates the difficulty of the simplified version for the testees

To sum up, all the issues raised in this section are related to the development of a standardized CSL/CFL test that the research team has faced over the years. Further studies are needed to explore issues such as the analysis of item difficulty for traditional and simplified character versions in each TOCFL level, the suggested learning hours for test takers of different L1 backgrounds, and the training and maintaining of item writers.

Concluding Remarks

The previous sections of this chapter described the motivation for creating the listening and reading tests of the then Chinese Proficiency Test (CPT) in 2001 and provided a brief overview of the historical development of the TOCFL and its predecessors (i.e., CPT and TOP). They also discussed SC-TOP's alignment of the TOP/TOCFL to the CEFR and some issues that the team faced in that effort. A few research studies conducted by SC-TOP researchers were also reviewed. In addition, these sections also highlighted four issues that the research team had faced over the years in its effort to develop and maintain a standardized test of CSL/CFL. In what follows, I present a brief account of the most recent development of the test. It is beyond the scope of this chapter to provide details of validity and reliability evidence of the four components of the TOCFL (i.e., listening, reading, speaking, and writing). Interested readers can refer to SC-TOP's technical reports (2014a, b, c, d).

From 2011 to Present

Since 2011, the SC-TOP has paid much attention to research on test validity and reliability, mainly including setting the passing grade for the new test and standard setting research of TOCFL in corresponding relationship with CEFR and ACTFL proficiency guidelines. The new version of TOCFL became available in 2013. In terms of task types, the old TOCFL and the new one are similar; the only difference is the scoring method. The new TOCFL results are presented in the form of scale scores instead of raw scores (the number of test items answered correctly). The scale score is more objective since it is not affected with the different item difficulties of each test. The new TOCFL has three proficiency bands: Band A, Band B, and Band C. Each of the bands has two levels. Therefore, there are a total of six levels: Levels 1–6. Test takers do not choose from the six levels, but only one of the three bands. In other words, two proficiency levels are distinguished in each band test to be taken by a test taker. The advantage of the new version is not only to simplify the administration of the test but also to save the quantity of test items in the item bank. Before 2013, each level of listening and reading test used 100 test items. After 2013, two levels use the same number of items. Furthermore, the SC-TOP is developing the computerized adaptive test (CAT), which can significantly enhance test efficiency and accuracy by automatically adapting item difficulty to examinees' ability levels. In the near future, the test will require fewer test items to arrive at equally accurate scores; and test takers also need not choose which level of test they should take. Lastly, to provide more useful information to test takers and widen the scope of the use of the test, SC-TOP also did a series of standard setting research during 2012–2014 in order to let test takers or school administrators know the relationship of a TOCFL proficiency level with that of CEFR and ACTFL (e.g., TOCFL level 3 equivalent to CEFR B1 and ACTFL intermediate-high level). More information about the research can be found on the SC-TOP Website (http://www.sc-top.org.tw/english/LS/test4.php).

Washback Effects of TOCFL and Future Directions

With the increasing number of test takers, positive impact on teaching and learning has been observed. Firstly, in order to help students pass the levels of the TOCFL, some language centers have adjusted their curriculum design and paid attention to learners' language performance, including abilities in practical writing. Second, the SC-TOP team has spent a lot of time conducting trial tests in language centers across Taiwan and overseas. During the process, learners gradually know how many hours they should take their Chinese lessons and what they can do with the Chinese language after they finish their formal study. A clear sense of learning and professional targets could make learners' CSL teaching and learning better targeted and more effective. Lastly, a lot of undergraduate and graduate programs in Taiwan adopt the TOCFL certificate as the evaluation of an applicant's Chinese proficiency and a requirement for admission.

Unlike its predecessors, the current TOCFL targets both CSL and CFL learners in a global context. To address the diverse backgrounds of learners in different places of the worlds, the TOCFL team has been cooperating with oversea universities and colleges, providing free tests to their students in order to collect and analyze the Chinese proficiency information of global learners and better support these learners in the future. Additionally, to accommodate the downward extension of CFL curriculums to young learners, a test named Children's Chinese Competency Certification (CCCC) for 7–12-year-old children was also developed by the SC-TOP and launched in 2009. How to best meet the needs of test takers from different backgrounds or with different purposes of taking a test is certainly a challenging task that the SC-TOP needs to address continuously in the future. It is hoped that this chapter has achieved its purpose of informing the international language testing community on the development of the TOCFL and that the issues discussed in this chapter that have challenged the SC-TOP can shed light on CSL/CFL test development and research.

Acknowledgements Some research studies described in this chapter were funded by the Ministry of Education, Taiwan. The author would like to express sincere thanks to the colleagues who worked in the SC-TOP, especially Ling-ying Lin and Pei-jun Lan. The author is also greatly indebted to Dr. Shou-hsin Teng, who has guided her in this field over the years, and Dr. Viphavee Vongpumivitch, who assisted her in the writing of this article. The author would also like to thank anonymous reviewers who gave valuable suggestions that have helped to improve the quality of the manuscript. Of course, all remaining errors are my responsibility.

References

American Council on the Teaching of Foreign Languages [ACTFL]. (1987). ACTFL Chinese proficiency guidelines. *Foreign Language Annals*, *20*(5), 471–487.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535–556.

Chang, K. C. (2009, November). *Xīnbǎn huáyǔwén nénglì cèyàn xiàobiāo guānlián xiàodù yánjiù* [*The research of criterion-related validity of updated test of proficiency-Huayu*]. Paper presented in American Council on the Teaching of Foreign Languages (ACTFL) Annual Meeting, San Diego, CA.

Chang, L. P. (2002). *Huáyǔwén nénglì cèyàn lǐlùn yǔ shíwù* [*Theoretical and practical relevance of Chinese proficiency test*]. Taipei: Lucky Bookstore.

Chang, L. P. (2010, December). *Huáyǔwén nénglì zhǐbiāo nándù fēnxī* [*Scaling descriptors of Chinese proficiency*]. Paper presented in Annual Conference of the Association of Teaching Chinese as a Second Language, Fu-Jen Catholic University, Taiwan.

Chang, L. P. (2011, January). *Duìyìng yú Ouzhou gòngtóng jiàgòu de duì wài hànyǔ xuéshí jiànyì* [*Suggested CSL learning hours based on the CEFR scale*]. Paper presented at the First East Asian Forum for Graduate Students of Teaching Chinese as a Second Language. Taipei: National Taiwan Normal University.

Chang, L. P. (2012). Duìyìng yú Ouzhou gòngtóng jiàgòu de huáyǔ cíhuìliàng [The study of the vocabulary size at the CEFR levels for CFL/CSL learners]. *Journal of Chinese Language Learning*, *9*(2), 77–96.

Chang, L. P., & Chen, F. Y. (2006). Huáyǔ cíhuì fēnjí chūtàn [*A preliminary approach to grading vocabulary of Chinese as a second language*]. In X. Su & H. Wang (Eds.), *Proceeding of 6th Chinese Lexical Semantics Workshop (CLSW-6)* (pp. 250–260). Singapore: COLIPS publications.

Chang, L. P., & Chen, F. Y. (2008, November). *Nénglì kǎoshì yǔ xuéxí zhījiān de guānxì* [*Relationship between proficiency test and learning*]. Paper presented in American Council on the Teaching of Foreign Languages (ACTFL) Annual Meeting, Florida.

Cheng, C. C. (1998). Cóng jìliáng lǐjiě yǔyán rènzhī [*Quantification for understanding language cognition*]. In B. K. T'sou, T. B. Y. Lai, S. W. K. Chan, & W. S.-Y. Wang (Eds.), *Quantitative and computational studies on the Chinese language* (pp. 15–30). Hong Kong: City University of Hong Kong.

Chinese Knowledge and Information Processing Group [CKIP]. (1998). *Accumulated word frequency in CKIP Corpus* (Tech. Rep. No. 98-01). Taipei: The Association for Computation Linguistics and Chinese Language Processing.

Chou, C. T., & Chang, L. P. (2007, October). *Huáyǔwén nénglì fēnjí zhǐbiāo zhī jiànlì* [*Proposal of a framework of Chinese language competence scal*e]. Paper presented in the Forum for Educational Evaluation in East Asia: Emerging Issues and Challenges. NTNU, Taipei.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2003). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEF). Manual: Preliminary pilot version*. DGIV/EDU/LANG 2003, 5. Strasbourg: Language Policy Division.

Council of Europe. (2004). *A bank of descriptors for self-assessment in European language portfolios*. Strasbourg: Language Policy Division. Retrieved August 1, 2010 from http://www.coe.int/T/DG4/Portfolio/documents/descripteurs.doc

De Jong, H. A. L. (1990). Response to masters: Linguistic theory and psychometric models. In H. A. L. De Jong & D. K. Stevenson (Eds.), *Individualising the assessment of language abilities* (pp. 71–82). Cleveland: Multilingual Matters.

Fachverband Chinesisch e. V. (2010). *Statement of the Fachverband Chinesisch e. V. (Association of Chinese teachers in German speaking countries) on the new HSK Chinese proficiency test*. Retrieved August 1, 2010, from http://www.fachverband-chinesisch.de/fachverbandchinesischev/thesenpapiereundresolutionen/FaCh2010_ErklaerungHSK.pdf

Figueras, N., North, B., Takala, S., Verhest, N., & Avermaet, P. V. (2005). Relating examinations to the common European framework: A manual. *Language Testing, 22*(3), 261–279.

Gao, S. H., Chen, L. P., & Lan, P. J. (2007). *TOP yǔ CEFR chūbù duìyìng jiéguǒ* [*The preliminary study of relating TOP to CEFR*]. Unpublished manuscript. Taipei: SC-TOP.

Guder, A., & Kupfer, P. (2005). *Empfehlungen des Fachverbands Chinesisch e.V. zur Stellung der Fremdsprache Chinesisch in chinawissenschaftlichen Studiengängen* [*Suggestion of TCFL in Germany University from the Fachverbands Chinesisch e.V.*]. Retrieved August 1, 2010 from http://www.fachverband-chinesisch.de/fachverbandchinesischev/thesenpapiereundresolutionen/resolution%20erlangen%20zweisprachig.pdf

Hanban (Confucius Institute Headquarters). (2010). *HSK*. Retrieved March 15, 2010, from http://english.hanban.org/node_8002.htm#nod

Hung, X. W. (2009, November). *Duì wài hànyǔ cèyàn kèlòuzì (wánxíng cèyàn) wénběn fēnxī chūtàn* [*A study of cloze design for the TOP reading test*]. Paper presented in American Council on the Teaching of Foreign Languages (ACTFL) Annual Meeting, San Diego, CA.

Hung, X. W. (2013, March). *Yǐngxiǎng duì wài hànyǔ yuèdú cèyàn nándù de yīnsù* [*The factors of item difficulty in TOCFL reading test*]. Paper presented in CLTAC Spring Conference, Stanford University, CA.

Ko, H. W., & Chang, Y. W. (1996). *Huáyǔwén nénglì cèyàn biānzhì yánjiù II* [*Research of Chinese Proficiency Testing II*]. National Science Council Research Report (NSC85-2413-H194-005).

Ko, H. W., Li, J. R., & Chang, Y. W. (1995). *Huáyǔwén nénglì cèyàn biānzhì: yǔfǎ shìtí nándù cèshì* [*Research of Chinese proficiency testing: Difficulty of grammatical items*]. National Science Council Research Report (NSC83-0301-H194-050).

Ko, H. W., Li, J. R., & Chang, Y. W. (1997). Huáyǔwén nénglì cèshìtí de biānzhì yánjiù [Research of Chinese proficiency testing]. *World Chinese Language, 85,* 7–13.

Li, C. C. (2009, December). *Cóng yuèdú píngliáng tàntǎo duì wài huáyǔ kèchéng guīhuà* [*An investigation of learners' self-assessment in reading Chinese as a second language*]. Paper presented in Annual Conference of the Association of Teaching Chinese as a Second Language, Miaoli, Taiwan.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(4), 483–490.

Liu, L.-L. (Ed.). (1997). *Hànyǔ shuǐpīng cèshì yánjiù [The research of Chinese proficiency test]*. Beijing: Beijing Language Institute Press.

Liu, Y. L., & Song, S. Z. (1992). Lùn hànyǔ jiāoxué zìcí de tǒngjì yǔ fēnjí [Calculating and ranking of Chinese Characters and words]. In The Office of Chinese Language Council (Ed.), *Hanyu Shuiping Cihui yu Hanzi Dengji Dagang* (pp. 1–25). Beijing: Beijing Language College Press.

Steering Committee for the Test of Proficiency—Huayu [SC-TOP]. (2009a, January). *2008 TOP Annual Report.* Taipei: SC-TOP.

SC-TOP. (2009b, October). *Monthly meeting handout*. Taipei: SC-TOP.

SC-TOP. (2014a). *Technical Report of TOCFL 2012 (1): Reliability and Validity of the Listening Test*. Linkou: SC-TOP.

SC-TOP. (2014b). *Technical Report of TOCFL 2012 (2): Reliability and Validity of the Reading Test.* Linkou: SC-TOP.

SC-TOP. (2014c). *Technical Report of TOCFL 2012 (3): Reliability and Validity of the Speaking Test.* Linkou: SC-TOP.

SC-TOP. (2014d). *Technical Report of TOCFL 2012 (4): Reliability and Validity of the Writing Test*. Linkou: SC-TOP.

Tewksbury, M. G. (1948). *Speak Chinese*. New Haven, CT: Far Eastern Publications, Yale University.

The College Board. (n.d.). *Scholastic Aptitude Test-Chinese*. Retrieved September 30, 2009, from http://www.collegeboard.com/student/testing/sat/lc_two/chinese/chinese.html?chinese

The Society for Testing Chinese Proficiency, Japan. (n.d.). *Zhōngguóyŭ Jiăndìng Shìyàn* [*Chinese Proficiency Test*]. Retrieved September 30, 2009, from http://www.chuken.gr.jp/

Weir, C. J. (2005). Limitations of the Common European framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.