AN INVESTIGATION INTO A CHINESE PLACEMENT TEST'S SCORE INTERPRETATIONS AND USES

By

Wenyue Ma

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2023

ABSTRACT

Foreign language placement testing, an important component in university foreign language programs, has received considerable, but not copious, attention over the years in second language (L2) testing research (Norris, 2004), and it has been mostly concentrated on L2 English. In contrast to validation research on L2 English placement testing, the discussion on tests in languages other than English is limited (e.g., Mozgalina & Ryshina-Pankova, 2015). Additionally, these studies have been constrained by two main methodological limitations. First, the importance of item-level data analysis is largely overlooked. While the researchers have highlighted the value of examining total test scores in validation research, defensible score interpretations and uses should not be assumed without further evidence showing all test items function as intended by test developers. Second, the validity evidence reported in these studies falls into a narrow range: the evidence mainly focused on generalization (e.g., reporting test reliability), explanation (group performance comparisons), and extrapolation (correlational studies on the relationship between test scores and other criterion) inferences, and validation needs more than that (Chapelle, 2021). In contrast, the documentation of empirical results supporting domain description (content representation and relevance), evaluation (examination of item quality), and utilization (stakeholders' perception of score usefulness) has been limited.

The primary goal of my dissertation is to provide a comprehensive examination and evaluation of the test score uses and interpretations for the listening and reading sections of an in-house, college-level Chinese placement test. For my dissertation, I collect and evaluate quantitative (placement test scores, item responses, ACTFL proficiency test scores) and qualitative (interviews, focus group, questionnaires) validity evidence in an argument-based validation framework that was conceptualized by Kane (2006), and was further expanded by

Chapelle et al. (2008): domain description, evolution, generalization, explanation, extrapolation, and utilization (see Chapelle, 2021, for a review). Employing mixed-methods, I aim to (1) study the functioning of test items by identifying and revising psychometrically problematic items, if any; (2) utilize the empirical results to inform test revisions; (3) demonstrate how the collected quantitative and qualitative results serve as strong or weak evidence or counterevidence for the claims within the validity argument; and (4) provide an overall evaluation of the intended interpretation and use of the placement test scores. With the study I hope to contribute to the larger discussion of the practices of foreign language assessment and argument-based test validation, and at the same time, offer insight into the ongoing development of validity research.

ACKNOWLEDGEMENTS

As I reflect on the journey of my doctoral studies, I am humbly grateful for the wealth of guidance, support, and encouragement I have received from numerous individuals who have had profound impacts on my life and work.

First and foremost, I would like to extend my heartfelt gratitude to my advisor, Dr. Paula Winke. Our journey began seven years ago when I first embarked on my graduate studies at Michigan State University (MSU), and her initial course was my introduction to this fascinating academic world. Her steadfast guidance, patience, and contagious enthusiasm for language testing have been an illuminating beacon throughout my pursuit. I remain forever indebted to her.

My appreciation extends to several faculty members whose guidance and insights have been instrumental in shaping my research and academic growth. Dr. Dan Reed provided patient and meticulous review for my Qualifying Research Paper (QRP) twice. His invaluable input has left a profound impact on the development of my research. I was privileged to work with Dr. Koen Van Gorp as a Graduate Assistant at the Center for Language Teaching Advancement (CeLTA). This opportunity presented a significant milestone in my academic journey, granting me hands-on experience in data analysis and invaluable guidance that has deeply influenced my doctoral studies. Equally deserving of special mention is Dr. Ryan Bowles, whose Applied Measurement course became the foundation upon which my dissertation was built. The knowledge and skills gleaned from his course have been instrumental throughout my study, for which I hold enduring gratitude. Furthermore, I wish to extend my gratitude to Dr. Steven Pierce from the Center for Statistical Training and Consulting (CSTAT). His mentorship has equipped

me with a wealth of skills and knowledge that will undoubtedly prove invaluable in my future endeavors.

My gratitude also extends to the faculty of the Chinese program at MSU, particularly, Ho-Hsin Huang, Xuefei Hao, and Wenying Zhou. Their assistance and insightful contributions have been critical to the successful completion of my dissertation project. The value of their input to my research cannot be overstated.

Special recognition goes to my cherished peers from the SLS program who have been my support system throughout this journey. Dylan Burton, Yingzhao Chen, Bronson Hui, Matt Kessler, Jongbong Lee, Shinhye Lee, Myeongeun Son, Michael Wang, Monique Yoder, and Xiaowan Zhang, thank you for your steadfast camaraderie and academic companionship. Your friendship has made my life at MSU both enjoyable and rewarding. In addition, my Seattle friends, Anqi Chen, Shjjia Chen, Corie Weijia Dai, Bixi Zhang, Hou Wang, and Liwei Wang, have enriched my remote working experience during the pandemic with excitement and fun.

Lastly, but most certainly not least, my heartfelt thanks extend to my parents, Feng Ma and Rebecca Wei Wang. Their boundless love, patience, and understanding during my prolonged academic journey far exceed what words can adequately express. I eagerly anticipate the joy of our imminent reunion. I would also like to acknowledge my boyfriend, Kevin Zhai Zihao. Your support throughout my doctoral journey, coupled with your remarkable efforts in helping me strike a work-life balance during these challenging times, has been nothing short of extraordinary. Thank you.

This dissertation is a culmination of the efforts and contributions of all those mentioned and many more not mentioned. I am deeply thankful for each of you.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Foreign language placement testing, an important component in university foreign language programs, has received considerable, but not copious, attention over the years in second language (L2) testing research (Norris, 2004). Unlike many ESL (English as a Second Language) programs, where standardized language proficiency test scores are often available for admission processes and thus can be used to inform placement decisions, foreign language programs at many universities in the United States often internally develop their own local placement tests (e.g., Georgetown University, the University of Wisconsin, Michigan State University). These tests aim to create groups of newly enrolled foreign language learners with homogeneous language abilities. The goal is to use the test scores to place students into courses which are at the appropriate levels for the students, which maximizes effective instruction. While these local placement tests are typically designed to be aligned with the local curriculum and language needs, validity evidence needs to be collected by the programs and test developers to ensure the alignment, both at or during test creation, and over the lifetime of the test's usage. The validity evidence can be used to justify and confirm the appropriateness of decisions and interpretations that are based on the test scores. However, against the wealth of existing validation research focusing on English placement testing, there has been, until recently, comparatively little discussion regarding the validity evaluations of foreign language placement tests. In addition, most validation research studies on foreign language placement tests conducted so far, as will be shown in the literature review section, have paid special attention to a narrow range of validity evidence, such as test reliability and comparisons of group-level performance, whereas documentations of validity evidence concerning test content relevance, test stakeholders' perception of the test, and item functioning, is comparatively limited.

The primary goal of my dissertation, therefore, is to provide a comprehensive examination and evaluation of the test score uses and interpretations for the listening and reading sections of an in-house, college-level Chinese placement test at a large Midwestern university in the U.S. In this paper, I collected and evaluated the validity evidence in an argument-based validation framework for the placement test. I proposed a set of warrants and their underlying assumptions following a sequence of six inferential steps that were conceptualized by Kane (1992, 2006, 2013), and were later expanded by Chapelle, Enright, and Jamieson (2008) and Chapelle and Voss (2021): (1) domain description, (2) evolution, (3) generalization, (4) explanation, (5) extrapolation, (6) utilization, (7) consequence implication. As Chapelle (2020) argued, "validation research needs to address a variety of different types of claims about scores encompassing such meanings as their real-world relevance, substantive sense, functional role, and stability. Such diverse meanings require research undertaken using a variety of methodologies including both qualitative and quantitative research" (p. 114). Therefore, I undertook a mixed-methods approach to data collection and analysis and integrated the two forms of data and their results to address research questions motivated by the warrants and their underlying assumptions specified in the validity argument, which aim to provide an overall evaluation of the intended interpretation and use of the placement test scores.

With this study I hope to contribute to the larger discussion of the practices of foreign language assessment and argument-based test validation, and at the same time, offer insight into the ongoing development of validity research.

CHAPTER 2: LITERATURE REVIEW

Argument-based validation in testing and assessment

The *Standards for Educational and Psychological Testing* (henceforth called the *Standards*), defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test scores" (AERA et al., 2014, p. 1). The definition is different from the previous notion held by some researchers that validity is a characteristic of tests and that tests can, therefore, be either valid or invalid. The arguments in the *Standards* (p. 1) are that "statements about validity should refer to particular interpretations for specified uses" and "it is incorrect to use the unqualified phrase of 'the validity of the test.'" These notions are in line with Kane's argument-based validity framework (1992, 2006, 2013), which conceptualizes validation as a process of building and evaluating a validity argument within the context of the test's score uses. Kane's approach to validation provides a means for defining intended score-based interpretations and uses, so that the specified interpretations and uses, rather than the test itself, is validated. This thus entails that if a test's score uses are changed (for example, if a test designed for college students is additionally applied for the use of assessing high school students), the validity argument will not apply to the new context, and a new validity argument structure must be formed and evaluated.

Technical terms and explanations

Referring to Toulmin's argument structure (2001, [1958] 2012), Kane's argument-based validity framework employs two kinds of argument. An *interpretive argument* makes claims about the proposed interpretation and uses of test scores by specifying relevant *inferences* with their supporting *warrants* and underlying *assumptions* that are necessary to make such claims. A

validity argument evaluates the interpretive argument based on the *backing* to determine whether the proposed score interpretations and uses are justified or not.

According to Kane (2006), *inferences* are steps denoting the reasoning processing to bridge examinees' observed performance to the claims based on the performance. To identify the types of evidence to support the intended score interpretation and uses, more detail about the inferences is needed. The detail is expressed in warrants and assumptions. A *warrant* is a general rule or an established procedure for inferring claims from observed performance, and *assumptions* underlying the warrant clarify what theoretical and empirical evidence, namely the *backing*, is needed.

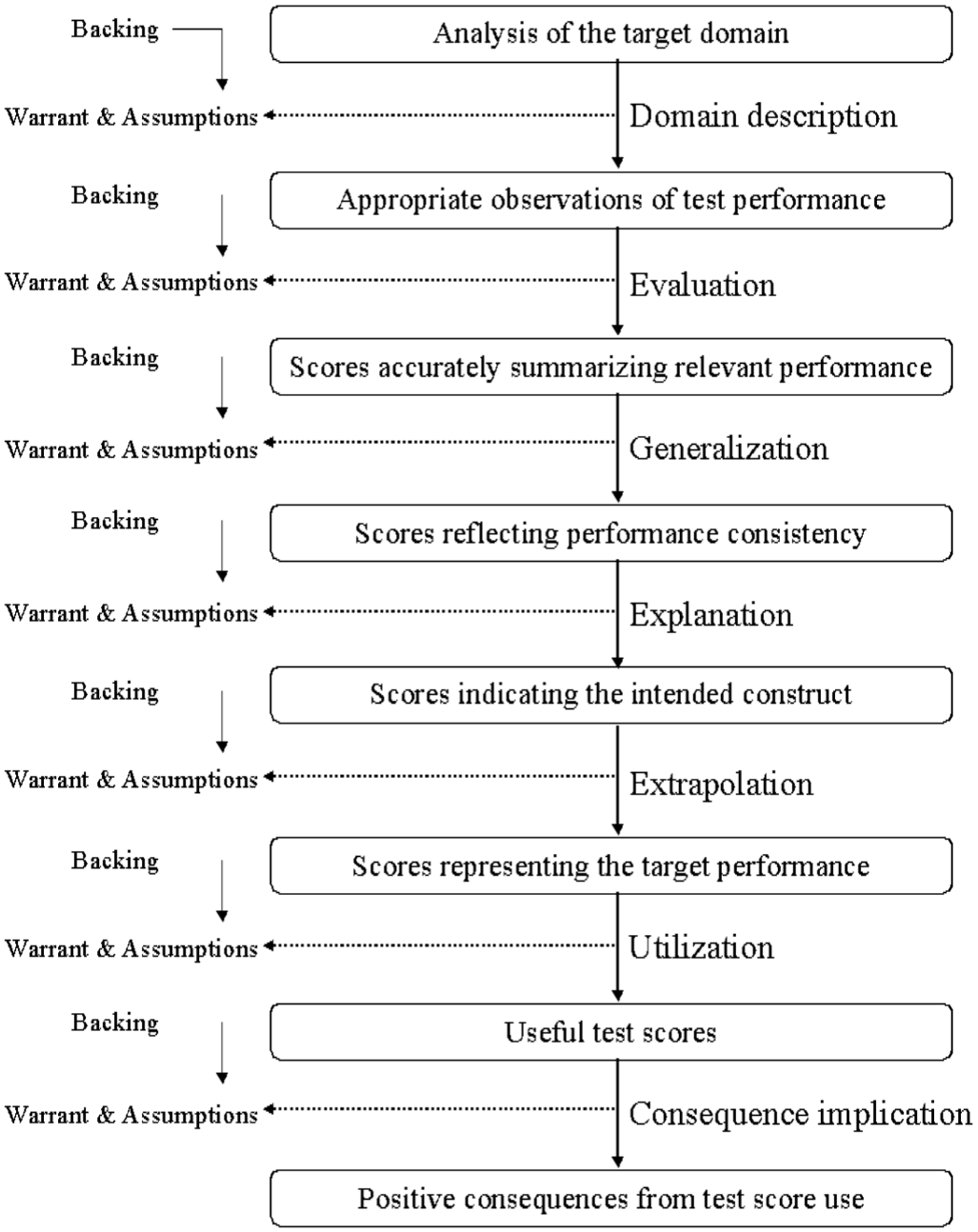
The seven inferences

Kane's argument-based validity framework originally classified score interpretations and uses into scoring, generalization, extrapolation, and decision inferences. Building on Kane's work, Chapelle, Enright, and Jamieson (2008) gathered validity evidence and formulated it into a validity argument to support score inferences for the 2005 revision of the Test of English as a Foreign Language Internet-based Test (TOEFL iBT; https://www.ets.org/toefl.html) through a sequence of six steps: domain description, evaluation, generalization, explanation, extrapolation, and utilization. It is important to note that the consequence implication step was subsequently introduced by Chapelle and Voss (2021), further refining the framework. A major contribution of argument-based validity is the explicit logic that connects the claims about test score interpretations and uses to the score inferences. The logical progression of the validity argument showing how each inference serves to connect these claims is illustrated Figure 1.

A *domain description inference* is made to examine whether the quality of the test development process for obtaining the observed test performance is appropriate for the proposed

test score interpretations and uses. The warrants and assumptions of the domain description inferences make direct reference to Sireci's (1998) research on content validity. The four elements of content validity described in Sireci, domain definition, domain representation, domain relevance, and appropriateness of the test development process, provide detail that is useful for evaluating test content quality. The backing to support the inference may be survey or interview data from content experts about importance, representation, and relevance of prospective test content in relation to the target domain.

Figure 1. An outline of the overall structure of a validity argument.



Note: Revised from Chapelle et al., 2008, p. 18

An *evaluation inference* is made to assess the extent to which the test scores are accurately summarizing relevant performance on test tasks. The quality of test scores can be evaluated from three perspectives, the test administration conditions, the task scoring procedures, and the observed test item quality (Chapelle, 2020). To investigate the evaluation inference, researchers may conduct item analysis to inspect statistical item characteristics, including appropriate difficulty and discrimination, and the existence of items bias; they may examine examinees' test-taking processes to study their cognitive engagement during the test; they could also conduct observation study at test centers to examine whether the required equipment, troubleshooting procedures, and accommodations to certain disadvantaged examinees are in place.

A *generalization inference* addresses an important issue in educational measurement, which is the degree to which score properties and inferences are generalizable to various measurement contexts (Cook & Campbell, 1979; Messick, 1995). More specifically, the inference is concerned with the extent to which the ratings for an examinee are consistent across multiple measurement settings, test forms, test tasks, and raters. The supporting evidence for the generalization inference is normally gathered in generalizability and reliability studies, but there are cases where appropriate scaling and equating procedures may be needed to ensure intended score interpretations and uses.

An *explanation inference* links test performance to the intended construct. More specifically, the inference leads to the question as to whether the observed scores can be attributed to the construct. Qualitative and quantitative research methods can both be used to investigate the inference. Support for the explanation inference is evidenced when (1) observed scores support the theorized position of the construct in relation to other constructs; (2) observed

scores support the internal structure of components of the construct; (3) examinees' test performance varies according to the amount and quality of the measured ability.

An *extrapolation inference* in the argument-based validity framework moves the argument from the intended construct to examinees' expected scores in the *target domain,* which is defined as "the full range of performances included in the [test score] interpretation" (Kane, Crooks, & Cohen, 2005, p. 7). The inference can be evaluated using two kinds of evidence, (1) the evidence collected through criterion-related studies supporting the relationship between examinees' performance on the test and other indicators in the target domain, (2) the evidence showing the quality of test performance, if it can be examined qualitatively (e.g., linguistic features), is comparable to other target domain performance.

A *utilization inference* is used to examine whether the test produces results that are useful for making appropriate decisions and can be well communicated to stakeholders. The inference can be evaluated from two perspectives, the intended uses for the test (i.e., utility) and the actual decision rules adopted by test users (i.e., decision). For the utility aspect of the inference, researchers need to provide evidence showing that the test scores are judged to be useful for the intended educational purposes (e.g., admission, placement, performance prediction, instruction effectiveness evaluation) by test stakeholders. As for the decision aspect, empirical evidence needs to be presented showing that the cut-off scores used for making decisions or the score bands used to describe either an individual examinee or groups of examinees are set appropriately.

A *consequence implication inference* connects the test use with its impact on stakeholders. Specifically, the inference examines whether the test uses have a positive influence on language teaching and learning. Empirical evidence supporting the inference can be backed

by examining diverse stakeholders' perspectives of the impacts of the test results on examinees' enhancement of language skills and curriculum development of language courses. The evidence can be collected through individual interviews and focus groups.

Advantages of the argument-based framework of validity

Chapelle, Enright, and Jamieson (2010) identified four advantages for the adoption of an argument-based approach to test validation over alternative approaches. First, the argument-based approach has shifted the prominent role of *construct* in validation research. In contrast to positioning construct as the foundation for test score interpretation (e.g., Messick's unitary validity framework, 1989, 1995), the argument-based validity is considered a more practical and efficient approach to test validation. This process, as Chapelle (2012) put it, "downplays, but does not eliminate, the need to define the construct" (p. 19), which has proven to be a daunting and difficult task in language assessment. Second, the interpretive argument and validity argument are linked by the research questions and their supporting evidence that are prompted by the particular warrants and assumptions. Therefore, the way in which the interpretive argument and validity argument is specified makes clear what research needs to be conducted and what types of validity evidence is required. Third, the internal logic among the argument is made apparent by showing how test performance and test score uses and interpretations are connected through a series of inferences. This allows validation to be completed through a systematic process of examining inferential steps rather than reviewing a list of types of potential validity evidence, some of which may not be directly relevant in the testing context of interest. Fourth, clear warrants and their underlying assumptions provide a place for counterevidence. The way the validity argument is specified creates an opportunity to challenge and question the proposed interpretation by presenting evidence for rival hypotheses. Given the advantages noted above, the

argument-based validity framework has thus been increasingly used in the field of second language testing and assessment in the recent ten years (e.g., Becker, 2018; Chapelle, Cotos, & Lee, 2015; Knoch & Chapelle, 2018; LaFlair & Staples, 2017; Winke et al., 2022; Yan & Staples, 2020; Youn, 2015) and has provided language testing researchers with useful guidance on how to collect and organize validity evidence following a logical structure to justify and support score-based interpretations and uses.

Validation research on foreign language placement tests

Over the past several decades, there has been a gradual increase in research focusing on placement testing (Long, Shin, Geeslin, & Willis, 2018). Within this body of work, the topic of placement test validity has garnered significant interest from researchers, as validity is a fundamental consideration in test development and test evaluation. However, in contrast to the richness of validation research on English placement tests, the discussion on the tests in languages other than English is relatively limited. Of the limited validation research on foreign language placement tests, most studies so far have mainly focused on gathering validity evidence by investigating examinees' performance at the test level or the group level and/or often relied on a single source of data, test scores, to claim validity for a given test (see Bernhardt et al., 2004; Eda et al., 2008; Heilenman, 1983, Long et al., 2018; Mozgalina & Ryshina-Pankova, 2015, Norris, 2004). Specifically, in addition to presenting the evidence of satisfactory test reliability, researchers claimed that a validity argument for the use of test scores was supported when (a) the mean total test scores were found to increase from examinees enrolled in lower-level courses to those in higher-level courses; (b) the total scores of the placement test were found to be strongly correlated with examinees' performance assessed by another measurement instrument assessing similar skills (e.g., a reading proficiency test, the oral proficiency interview

test); and (c) examinees' performance in the placement test was found to improve after they had received instruction and practice. Table 1 provides a brief summary of the validity evidence provided in these validation studies. Below I described these studies in more detail.

Eda and her colleagues (2008) assessed the reliability and construct validity of the Japanese Skill Test (JSKIT), which comprises various single-skill tests along with a grammar test. This evaluation was conducted to determine the test's effectiveness as a placement tool for a nine-week summer intensive program. Out of the 250 students enrolled in the summer Japanese program, 136 participated in the study, taking not only the JSKIT, but also an internal placement test and an oral proficiency interview test at the beginning and end of the program. After comparing the results from the three tests, the researchers assessed both the reliability and effectiveness of the JSKIT in differentiating learners at various proficiency levels, the researchers concluded that the JSKIT served as a reliable and effective placement tool for students with lower levels of proficiency, specifically those with first and second-year language abilities.

The assessment of the Japanese Skill Test (JSKIT) demonstrates the importance of validation in test development. Following this notion, Cronbach (1971) described test validation as an "ever-extending inquiry" (p. 452), which necessitates an ongoing research program rather than a single empirical study. This concept is exemplified in the research conducted by Norris (2004) and Mozgalina and Ryshina-Pankova (2015), who documented the assessment development, validation process, test revisions, and validity evidence evaluations for the placement test used in the Georgetown University German program. The placement test comprised three sections: a cloze test (C-test) with five progressively more difficult texts, a reading comprehension test, and a listening comprehension test. In Norris's study, a total of 193

students enrolled in the German program completed all three parts of the placement test, including previously placed and non-placed students. The placed students are those who entered the course based on their placement test results, while the non-placed students are those who progressed through the lower-level courses without taking the placement test. Additionally, 124 of the previously placed students also took both a semester-beginning and a semester-end test administration (of the same test). Through the analysis of multiple data sources, including students' placement test scores, course grades, scorers' marking sheets, and instructor interviews, Norris determined that the C-test produced more reliable test scores and assessed a broader range of student abilities compared to the listening and reading comprehension tests, making the C-test more suitable for placement purposes.

Building on Norris's initial validation efforts, Mozgalina and Ryshina-Pankova (2015) conducted a validity evaluation of a revised C-test, which was part of the placement test in Georgetown University's German program. The test revisions were implemented to better align with the updated German curriculum following Norris (2004). Administered at the beginning and end of the semester, the researchers reported results from a total of 222 examinees across various course levels, with 66 of them taking the test at both administrations. The findings indicated that the test effectively distinguished between examinees of varying abilities and successfully tracked progress for upper-level students.

The importance of test validation in the context of Georgetown University's German program highlights the significance of selecting the appropriate assessment methods. Heilenman (1983) conducted a study with a larger sample size, examining the C-test scores of 388 students enrolled in French at Northwestern University to determine whether the test was a valid measure of language proficiency and could effectively differentiate students at various instructional

levels. In contrast to Norris (2004) and Mozgalina and Ryshina-Pankova (2015), where the C-test was supported as a placement measure, Heilenman concluded that the C-test should be used cautiously as an alternative or supplement to other placement measures. This caution stems from the considerable overlap in scores obtained by students at different instructional levels, resulting in significant discrepancies between students' actual course assignments and those predicted by the C-test scores.

While the studies discussed so far focused on paper-based tests, web-based language testing, also known as computer-based testing, has gained considerable attention over the past 30 years in second language assessment research. This is due to its potential to greatly enhance the flexibility and logistical efficiency of test delivery and scoring processes (Long et al., 2018; Ockey, 2006). Two empirical validation research studies have specifically examined the practicality and efficiency of web-based language placement tests. Bernhardt et al. (2004) assessed the utility and validity of two web-based language tests as placement tools for college-level German and Spanish programs. The test score reliability and validity of the two placement tests were evaluated using data from 78 students in the German program and 679 students in the Spanish program, with 14 German learners and 41 Spanish learners retaking the placement test after three quarters of target language instruction. The results suggested that the test scores were reliable, and evidence of validity was supported by the trend of students' improved performance in the second administration of the tests.

In a similar vein, Long et al. (2018) investigated the reliability and validity of a newly developed web-based Spanish placement test. Building upon and expanding Bernhardt et al.'s research, the study analyzed testing data from 2,111 students enrolled in a college-level Spanish program, with 1,622 of them also taking the paper-based test. Besides providing evidence of high

test reliability, the researchers evaluated the functionality and use of the test scores by examining content relevance (the alignment between the test content and course materials) and score invariance across different modes of test delivery (the alignment between the test results obtained from the web-based test and the original paper-based test). The results suggested that the test was valid in terms of content relevance and placement decision appropriateness.

The need for the current study

The studies discussed so far have undoubtedly contributed valuable insights into the evaluation of measurement validation and the appropriateness of placement testing practices, advancing researchers' understanding of factors that contribute to placement test effectiveness. However, these studies are subject to two main methodological limitations.

Firstly, most validation studies on foreign language placement tests tend to overlook the importance of item-level data analysis. While the research endeavors mentioned previously emphasize the significance of examining total test scores in validation research, defensible score interpretations and uses should not be assumed without further evidence demonstrating that all test items function as intended when eliciting examinees' responses (e.g., items are free of bias; examinees at lower ability levels are less likely to correctly respond to difficult items compared to their peers with higher abilities). This point is evident in the following example: evidence suggesting that a test as a whole demonstrates good discriminating power and high reliability does not necessarily guarantee that all test items are problem-free and equally effective and appropriate in assessing and discriminating examinees' target abilities.

Secondly, the validity evidence reported in these studies is somewhat narrow in scope, as it can be observed that evidence supporting specific aspects of score interpretations and uses is often missing in building and supporting validity arguments for foreign language placement tests

(see Table 1). A closer examination of the validity evidence reported in these validation research studies reveals that, primarily for practical reasons, the validity evidence mainly focuses on generalization (reporting test reliability), explanation (group performance comparisons), and extrapolation (correlational studies on the relationship between test scores and other criteria) inferences. In contrast, the documentation of empirical results supporting domain description (content representation and relevance), evaluation (examination of item quality), utilization (stakeholders' perception of score usefulness), and consequence (washback effect) is comparatively limited. Research addressing this gap is necessary, as the seven inferences together form a complete, logical structure for validity evaluation.

Table 1. Summary of validity evidence in previous foreign language placement test validation studies

Study	Validity evidence
Bernhardt, Rivera, & Kamil (2004)	Domain description inference: ● Interviews with instructors about perceptions of the placement testing in relation to their teaching would reveal that the content assessed in the test is critical in successful course completion; Generalization inference: ● The test would yield scores with high reliability; Explanation inference: ● Students would perform significantly better on the second administration of the test.
Eda, Itomitsu, & Noda (2008)	Generalization inference: ● The test would yield scores with high reliability; Explanation inference: ● The test would effectively differentiate students at different course levels; Extrapolation inference: ● The scores on the test would be positively correlated with the scores on other tests (an in-house placement test and the OPI); ● Placement decisions made based on scores of the test would be in agreement with those based on scores of the in-house placement test and the OPI.
Heilenman (1983)	Explanation inference: ● Students who are enrolled in the progressively higher course levels would perform with higher scores on the cloze test than students at the preceding curricular levels; ● Extrapolation inference: The scores on the cloze test would be positively correlated with the scores on the Reading and Writing parts of the placement test.

Table 1 (cont'd)

Long, Shin, Geeslin, & Willis (2018)	Domain description inference: ● Assessment items would be matched with corresponding course content; Generalization inference: ● The test would yield scores with high reliability; Explanation inference: ● There would be a strong relationship between the scores on the web-based test and the paper-based test; Placement decisions made based on scores of the web-based test would be in agreement with those based on scores of the paper-based test.
Mozgalina & Ryshina-Pankova (2015)	Generalization inferences: ● The C-test would yield scores with high reliability; Explanation inferences: ● The C-test would elicit a wide distribution of scores from examinees of differing abilities ● The scores on the new C-test would be positively correlated with scores on the old C-test; ● Average C-test scores would increase between the beginning and the end of the semester; ● Students who are enrolled in the progressively higher curricular levels would perform with higher scores on all five texts than students at the preceding curricular levels; Extrapolation inferences: ● The new C-test scores would be positively correlated with the scores on the Reading and Listening comprehension parts of the placement test.

Table 1 (cont'd)

Norris (2004)	Generalization inferences: ● The C-test would yield scores with high reliability; Explanation inferences: ● The C-test would elicit a wide distribution of scores from examinees of differing abilities; ● Average C-test scores would increase between the beginning and the end of the semester; ● Students who are enrolled in the progressively higher curricular levels would perform with higher scores on all five texts than students at the preceding curricular levels; ● There would be positive relationships between the three placement exam sub-tests; Utilization inference: ● The errors associated with specific cut-scores on the tests would be small enough for the scores to be useful for making placement decisions; ● Teachers would perceive the test as a useful and effective tool for making accurate placement decisions;

Note: Validity evidence reported in these studies was organized and categorized by inference in the argument-based validity framework

Recognizing the need to address the methodological limitations identified in previous studies, with the current research I aim to provide a more comprehensive evaluation of foreign language placement tests. Therefore, my goal with my present study is to expand upon the existing validation research on foreign language placement testing by gathering and presenting validity evidence (*backing*) in an argument-based validation framework (following the seven-step interferences) that can be utilized to comprehensively evaluate the intended interpretation and use of test scores in the context of Chinese placement testing for a college-level language program. In addition, the study provides insight into how the validity evidence collected through the validation process can inform test revisions. Guided by the main purposes, I formulated the research questions as shown below focusing on obtaining backing for the domain description, evaluation, generalization, explanation, extrapolation, and utilization inferences:

RQ1: Do observations of performance on the MSU Chinese placement test reveal relevant Chinese knowledge, skills, and abilities required for the successful completion of language courses offered by the MSU Chinese program (warrant 1, see more information in Table 4)?

RQ2: Do tasks on the MSU Chinese placement test exhibit desired statistical characteristics (warrant 2)?

- Do test items yield item difficulty estimates that are appropriate for making placement decisions?
- Do test items show no evidence of item bias?
- Are correct options unambiguous and accurately keyed?

RQ3: Are score-based results generalizable to various measurement contexts (warrant 3)?

- Does the MSU Chinese placement test produce scores that are internally consistent?

- Are there adequate items to reliably differentiate students' abilities into three levels as intended?

RQ4: Can students' test scores be attributed to the construct of interest (warrant 4)?

- Does students' test performance vary according to the amount and quality of prior Chinese learning experience?
- Do students' test scores support the internal structure of the intended construct?

RQ5: Do students' test scores support the relationship between their performance on the test and other indicators of Chinese language proficiency (warrant 5)?

RQ6: Does the test produce results that are useful for test users (warrant 6)?

- From the perspective of course instructors, are students placed into appropriate course levels?
- From the perspective of students, are they placed into appropriate course levels?
- Are cut-off scores set appropriately?

RQ7: Does the test have positive effects on Chinese teaching and learning (warrant 7)?

CHAPTER 3: METHODOLOGY

Participants

Examinees of the MSU Chinese placement test (before test revisions)

The testing data for this study are pre-existing and come from 305 examinees (152 females, 153 males) who took the Chinese placement test and were planning to take Chinese language courses at Michigan State University (MSU). The examinees took the test between 2016 and 2020, and they were between the ages of 14 and 65 (Median = 18, Mean = 19.1, SD = 4.3) when they were taking the test. The anonymized data were provided to me as a loan after obtaining IRB approval. The data were collected over multiple years (2016 to 2020).

Examinees of the ACTFL language proficiency tests (before test revisions)

Among the 305 examinees whose placement test data were included in the analysis, 55 examinees' Chinese language skills were also measured using the ACTFL language proficiency tests from Language Testing International (LTI, https://www.languagetesting.com/) in speaking (the computerized oral proficiency test, or OPIc), reading (Reading Proficiency Test, or RPT), and listening (Listening Proficiency Test, or LPT). Students in the Chinese program at MSU completed the ACTFL tests during 2014-2018 as a curriculum requirement in conjunction with the federal grant, known as the Language Proficiency Flagship Initiative (see Winke et al., 2020). I borrowed the ACTFL test data as well, anonymized, but with codes to match them with the placement test data. Each student was offered three tests, but not all students took all three tests. Some students took the same test more than once as they were studying Chinese for more than one academic year at MSU. In such a case, I only included one test score for analysis purposes. The score considered was taken closest in time to the placement test for students who had taken the same test multiple times.

Examinees of the MSU Chinese placement test (after test revisions)

As will be explained in the procedures section, I collected new data from a separate cohort of students to gather data on students' performance in the revised placement test as well as their perceptions of the placement test post-item-analysis and test revisions. In the first week of Spring 2022, I sent out an email and invited all students who were enrolled in the Chinese language courses at MSU to participate in a three-phase research project (see more information in the procedure section). Thirty-seven students completed the first phase; thirty-two students completed the first two phases, and twenty-eight students completed all three phases. Table 2 presents the demographic information and Chinese learning background of the 28 students who completed all three phases of the study. After students completed the third phase, I reached out to 7 students (100-level: n = 2; 200-level: n = 3; 300-level: n = 2) and invited them to participate in semi-structured interviews.

Table 2. Demographic information of examinees of the placement test post revision

Course level	n	Gender (n)			Age	Years of Chinese instruction before college
		Female	Male	Other	Mean (SD)	Mean (SD)
100	13	9	3	1	18.8 (0.7)	4.15 (4.4)
200	9	4	4	1	19.9 (1.5)	6.5 (3.9)
300	6	2	4	-	20.3 (0.8)	4.67 (1.8)
Total	28	15	11	2	19.4 (1.2)	5 (3.8)

Chinese course instructors

The research project involved the participation of all three instructors from the Chinese program at MSU. These instructors were selected based on their qualifications, which included having at least 10 years of Chinese teaching experience and a minimum of 5 years teaching Chinese at MSU. As the entirety of the program's faculty, they provided a complete representation of the instructors involved. According to the results of the instructor questionnaire, each of these instructors has taught Chinese language courses at the 100-, 200-, and 300-levels at MSU.

Instruments

Michigan State University Chinese placement test

The MSU Chinese placement test is designed to help students register for the appropriate level of Chinese course by determining their starting level for college language study at MSU, based on their proficiency in Chinese. The test begins with a background questionnaire related to examinees' Chinese learning experiences, consisting of eight questions about their language learning history, such as the number of years spent studying Chinese, any standardized Chinese test scores, family connections to the Chinese language, and time spent in a Chinese-speaking country.

The test comprises four language assessment sections: listening, reading, speaking, and writing. The test was implemented in Qualtrics in 2016 (see the link to the test: https://msu.co1.qualtrics.com/jfe/form/SV_a2C5uBOWKlTCdoN). The placement test is not timed, but examinees usually finish the test within 25 minutes to an hour. The current study only includes students' responses to the questions in the listening and reading section, as only those who score above a certain level on these sections have their speaking and writing sections scored by instructors in the Chinese program.

The listening section features 14 multiple-choice questions, while the reading section contains 18 multiple-choice questions, each offering three or four choices. The test questions align with the course curriculum, as they were initially drafted by the program's instructors and based on materials or content provided to students during the various semester-level courses. Consequently, the placement test is a compilation of the language program's content, organized from start to finish according to instructional levels.

Examinees' total scores on the receptive part of the test are the sum of their scores on the 32 multiple-choice questions in the listening and reading sections. Students who receive scores of 19 or below are recommended for placement in 100-level (CHS101 or CHS102) courses, while those scoring between 20 and 29 are recommended for 200-level courses. Students with scores of 30 and above are tentatively approved to enroll in a 300-level course (CHS301) after an evaluation of their writing and speaking skills. They are also required to complete an in-person language assessment interview with an instructor during the first week of class to verify their language-level placement.

The cut-off scores were determined in a pilot study prior to the test's official launch in 2016. During the piloting stage, the test was administered to students enrolled in 100-, 200-, and 300-level Chinese courses offered by the program. The cut-off for each level was set to the score one standard deviation above the mean score obtained by students enrolled in the corresponding course level. The rationale behind this decision was to place students in the highest level course in which they have a good chance of success. While students' placement decisions are largely determined by their total scores from the listening and reading sections, their responses in the speaking and writing sections assist teachers in evaluating the accuracy of upper-level placement decisions.

Questionnaire for instructors

I developed a questionnaire to gather content experts' (i.e., Chinese language course instructors) perceptions of item difficulty, content representation, and relevance, in order to assess the extent to which the test captures the target domain. The questionnaire, implemented in Qualtrics (see the link: https://msu.co1.qualtrics.com/jfe/form/SV_9zVfsONQfsgPnp4), consists of three parts. The first part focuses on basic information about the course the instructor is

teaching, such as course level, class size, and evaluation criteria. This section concludes with a question that probes instructors' perceptions of the essential skills and knowledge required for successful completion of Chinese language courses at each level. The second part features a survey with the 32 items from the placement test. Instructors are asked to rate these items on a 6-point Likert scale in terms of overall item difficulty (1: very easy; 6: very difficult). Furthermore, instructors are presented with a series of checkboxes for each item to assess its relevance and appropriateness to the content of 100-level, 200-level, and 300-level courses. A separate checkbox is provided for instructors to indicate if the item is not relevant to any of the three course levels. Instructors are instructed to mark the appropriate checkbox(es) and leave the others unchecked. The final section collects feedback on the placement test and solicits suggestions for improvement.

Interviews with instructors

I conducted semi-structured interviews with all three Chinese course instructors. Each interview lasted approximately an hour and followed a set of six predetermined questions, which can be found in Appendix 1. Subsequently, the instructors were presented with their responses from the questionnaire during the interview and asked to provide further elaboration or clarification on their answers. I chose semi-structured interviews because they allowed for flexibility in exploring the instructors' experiences and perspectives, while still maintaining a consistent framework for comparing their responses. This approach facilitated rapport building with the instructors (DiCicco-Bloom & Crabtree, 2006) and enabled the collection of richer, more nuanced data to better understand their teaching strategies and challenges in mixed-proficiency classrooms (Galletta, 2013). While the primary focus was on the pre-drafted questions, I also explored new areas of discussion as they emerged during the interview. The

interviews were conducted in Mandarin Chinese and translated into English in two stages. The initial translation was done using machine translation software, Xunfeitingjian (https://www.iflyrec.com/zhuanwenzi.html). I then invited another researcher, who is a highly proficient L2 Mandarin speaker, to review the translation with me. We identified and discussed any translation errors or ambiguities, and made the necessary adjustments.

Questionnaire for students

I created a questionnaire to assess students' perceptions of item difficulty, content representation, and relevance. The questionnaire is implemented in Qualtrics (see the link: https://msu.co1.qualtrics.com/jfe/form/SV_734ClDDR4uChQIm) and consists of three parts. The first part gathers students' personal information and inquires whether they took the MSU placement test prior to their enrollment to the first Chinese language course at MSU. If so, they were asked about a few questions related to their perception of the accuracy of the placement test. The first part concludes with a question that taps into students' perceptions of the essential skills and knowledge that are required for successful completion of Chinese language courses that they were placed into. The second part is a survey with 32 items in the placement test. Students were asked to rate these items on a 6-point Likert scale in terms of the overall item difficulty (1:very easy; 6: very difficult). In addition, students were asked to rate on the relevance and appropriateness to the content of the course they were taking (e.g., 1 = *the item is NOT relevant to the course that I am taking;* 6 = *the item is highly relevant to the course that I am taking*). Note that the reason for using different ways to assess the relevance and appropriateness of test items to course content for students and instructors is that students are enrolled in different levels of Chinese language courses, and therefore some test items may be more appropriate and relevant to higher-level courses, while others may be more suitable for lower-level courses.

Instructors are experts in their field and are better equipped to evaluate the difficulty level and appropriateness of test items across different levels of courses. On the other hand, students are the ones better able to judge the relevance of the test items to the specific course material they are studying. By using different instruments, I can obtain more accurate and informative data on the difficulty level, appropriateness, and relevance of test items for different levels of Chinese language courses. The final section gathers feedback on the placement test and suggestions for improvement. The complete questionnaire is available in Appendix 2.

Interview with students

As noted earlier, I reached out to seven students and invited them for 45-minute semi-structured interviews. The interviews were guided by six pre-determined questions (see Appendix 3) and were conducted in English. Similar to the interviews with the instructors, I relied on the pre-drafted questions, but I went off-script and pursued other lines of inquiry when necessary. Subsequently, the students were presented with the responses they provided in the questionnaire and were asked to expand upon their responses with clarifying comments.

Procedures

Obtaining the pre-existing data

The pre-existing data for this study consists of (1) the testing data for the MSU Chinese placement test collected from 2016 to 2020 (hereinafter referred to as the placement test) and (2) the proficiency data from the American Council on the Teaching of Foreign Languages (ACTFL) Chinese language proficiency tests. I directly obtained the anonymized, pre-existing placement test data from the professor who designed and maintains the test for the MSU Chinese program. The professor downloaded the data from the MSU Qualtrics site where the test data is stored without names, and with codes instead. Additionally, I obtained the anonymized, pre-

existing ACTFL Chinese language proficiency test data from the Principal Investigator of the Language Proficiency Flagship Initiative, also with codes, not names. It is important to note that the ACTFL testing data was included only for those who also had available placement test data.

Revising the MSU Chinese placement test

Utilizing the pre-existing placement testing data and the employment of Rasch analysis and item-level analysis, I identified issues with a number of items in terms of their item characteristics. These psychometrically problematic items were flagged and revised according to the literature on Chinese grammar rules as well as the feedback from two L1 Chinese speakers with PhD in applied linguistics, a former Chinese course instructor at MSU and a language testing researcher. More information about the test revisions is provided in the results section.

Collecting data from instructors

In the Spring of 2022, I approached three instructors who were teaching Chinese language courses at the 100-level, 200-level, and 300-level. All three instructors agreed to participate in the research project, and were asked to complete a questionnaire evaluating their perceptions of the difficulty of the items in the placement test, content representation, relevance, and the accuracy of the placement test results. Upon completing the questionnaire, the instructors were invited to participate in a one-hour one-on-one, semi-structured interview.

Collecting testing data using the revised placement test from students

In Spring 2022, I obtained the consent of Chinese language course instructors to invite their enrolled students to participate in my research project. I emailed invitations to all eligible students, and out of the 37 students who expressed interest, 28 successfully completed all three phases of the study (100-level: n = 13; 200-level: n = 9; 300-level: n = 6). In the first phase, students were asked to take the revised MSU Chinese placement test during the first week of

Spring 2022. Similar to the original test, the revised version began with a background questionnaire about the examinee's Chinese language learning experiences, followed by 32 multiple-choice items (14 listening and 18 reading). In the second phase, students took the revised placement test again in the final week of Spring 2022. After completing the test, I sent a questionnaire to students to assess their perceptions of item difficulty, content representation, and relevance. Participants were compensated with either $30 or extra credit for their participation in the research project. As mentioned earlier, seven students were invited to participate in 45-minute one-on-one semi-structured interviews. These seven students received an additional $10 for their participation in the interviews.

Data analysis

Table 3 presents the methods for analysis that are used to answer each research question. I employed both quantitative and qualitative approaches to analyzing the data. For interview data, I used the iterative qualitative data coding procedures (open coding, theme development, and coding for patterns) described in Baralt (2012) to examine instructors' and students' perceptions of the accuracy of the placement test results (utilization inference), the effects of the test on teaching and learning (consequence implication inference), as well as the test content relevance and representativeness (domain description inference). For quantitative testing data, I conducted Rasch analysis, Differential Item Functioning (DIF) analysis, item analysis to address the evaluation inference. I reported Rasch-based reliability estimates to evaluate the generalization inference. To investigate the explanation inference, I compared students' placement test performance at the beginning and the end of the semester and across different course levels. In addition, I conducted an exploratory factor analysis to examine whether the results support the hypothesized internal structure of the measured construct of the placement

test. To examine the extrapolation inference, I calculated the *polyserial correlation* coefficients to assess the relationships between students' placement performance and their scores on ACTFL language proficiency tests. The polyserial correlation coefficient is better suited for calculating the correlation between a continuous variable and an ordinal variable in comparison to other commonly known correlation coefficients utilized by applied linguists, including Pearson's *r*, Spearman's *rho*, and Kendall's *tau* (Winke, Zhang, & Pierce, 2022). To evaluate whether the cut-off scores for the placement test are set appropriately (utilization inference), I analyzed teacher ratings on item relevance and assessed item distribution across course levels. To examine what consequence implications the test has on Chinese teaching and learning, I reported results yielded from questionnaire and interview data from teachers and students.

For quantitative data preparation, I binary-scored students' placement test responses as 1 (correct) or 0 (incorrect) for multiple-choice items, with unanswered questions coded as missing data (X). Test scores on the ACTFL language proficiency tests are linked directly to the *ACTFL Proficiency Guidelines*, a framework for language proficiency on "functional ability" (ACTFL, 2012, p.3), describing what individuals can do with the target language with each skill (i.e., speaking, reading, listening, writing). For each skill, the guidelines feature five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The major levels Advanced, Intermediate, and Novice are subdivided into High, Mid, and Low sublevels to distinguish the language learners at these levels more clearly. For ACTFL language proficiency data, I assigned a numeric value to each ACTFL proficiency level on a scale of 1 (Novice Low) to 10 (Superior), a practice following prior research (e.g., Isbell et al., 2018; Kenyon & Malabonga, 2001; Ma & Winke, 2019; Tigchelaar et al., 2017; Zhang et al., 2020).

Table 3. Summary of the warrant, assumptions, and associated backing in the MSU Chinese placement test interpretive argument

Inferences	Warrants	Assumptions Underlying Warrant	Sources for Backing
Domain description	Warrant 1	• The relevance of the test items and test criteria to the instructional domain and the appropriateness of the item difficulties are supported by test stakeholders.	• Questionnaire and interview data about test content relevance from course instructors* • Questionnaire and interview data about test content relevance from students*
Evaluation	Warrant 2	• Item difficulty estimates are appropriate for making placement decision.	• Rasch analysis • Teachers' perceptions of item difficulties • Students' perceptions of item difficulties • Item difficulties computed from students' actual test performance
		• Test items exhibit no evidence of item bias	• DIF analysis
		• Correct responses are unambiguous and accurately keyed.	• Item-level analysis
Generalization	Warrant 3	• The test produces scores that are internally consistent. • The test yields satisfactory item reliability	• Cronbach's alpha • Rasch-based item reliability estimates
		• The test yields satisfactory person reliability	• Rasch-based person reliability estimates

Table 3 (cont'd)

Explanation	Warrant 4	• Test performance varies according to the amount and quality of experience in learning Chinese • Test scores support the internal structure of the construct	• Comparison of test performance in first and second admissions* • Comparison of test performance between students at different course levels* • Exploratory factor analysis
Extrapolation	Warrant 5	• Scores on the MSU Chinese placement test are positively correlated with scores on RPT, LPT, and OPIc.	• Polyserial correlation analysis
Utilization	Warrant 6	• Test users (i.e., instructors) judge the scores to be useful. • Cut-off scores are set appropriately	• Questionnaire and interview data from course instructors about the accuracy of placement decisions • Interview data from students about the accuracy of placement decisions • Analysis of instructor ratings on item relevance and assessment of item distribution across course levels
Consequence implication	Warrant 7	• The test has positive effects on Chinese instruction and learning	• Questionnaire and interview data from teachers and students

Note: *Analysis conducted using the testing data from the revised MSU Chinese placement test

Criteria to determine strong, weak, or counter-evidence for proposed validity argument

The empirical results from the analyses were evaluated based on the following criteria to determine if they provide strong, weak, or counter-evidence for the proposed validity argument. Establishing these criteria is vital within the argument-based validity framework, as it leverages the framework's benefits. This framework emphasizes the systematic organization of validity evidence and urges researchers to articulate explicit inferences and assumptions that underlie the validity argument. By offering clear criteria for assessing evidence strength, the framework's advantages are enhanced, ensuring that the study's conclusions are robust, reliable, and well-founded. Additionally, this approach improves the research's transparency and comprehensibility, enabling readers and stakeholders to better grasp the results' implications and their impact on the placement test's validity.

1. Domain description

The relevance of the test items and test criteria to the instructional domain and the appropriateness of the item difficulties are supported by test stakeholders.

- Strong evidence: 5% or less of the items are considered not relevant to class content/evaluations; the appropriateness of the item difficulties is supported by test stakeholders
- Weak evidence: 10% or less of the items are considered not relevant to class content/evaluations.
- Counter evidence: More than 10% of the items are considered not relevant to class content/evaluations; the appropriateness of the item difficulties is not supported by test stakeholders

2. Evaluation

Item difficulty estimates are appropriate for making placement decisions.

- Strong evidence:
    - Based on the Wright map, the item difficulty estimates target the students given their ability estimates and thus are appropriate for assessing and differentiating the students.
    - The item difficulty estimates that are computed from students' actual test performance are consistent with the expected difficulty level for the intended audience (students' and/or teachers' perceptions of item difficulties).
- Counter evidence:
    - Based on the Wright map, the items are too difficult or easy for the students, and thus are not appropriate for assessing and differentiating the students.
    - The item difficulty estimates derived from students' actual test performance do not align with the expected difficulty level for the intended audience (students' and/or teachers' perceptions of item difficulties).

Test items exhibit no evidence of item bias

- Strong evidence: 5% or less of the items exhibited DIF across gender
- Weak evidence: 10% or less of the items exhibited DIF across gender
- Counter evidence: More than 10% of the items exhibited DIF across gender

Correct responses are unambiguous and accurately keyed.

- Strong evidence: 95% or more of the items are shown to be unambiguous and accurately keyed

- Weak evidence: 90% or more of the items are shown to be unambiguous and accurately keyed
- Counter evidence: Less than 90% of the items are shown to be unambiguous and accurately keyed

3. Generalization

The test produces scores that are internally consistent

- Strong evidence: Cronbach's alpha is above .8
- Weak evidence: Cronbach's alpha is above .7
- Counter evidence: Cronbach's alpha is below .7

There are adequate items to reliably differentiate students' abilities into three levels as intended

- Strong evidence: Person reliability is above .9; person separation index is above 2
- Weak evidence: Person reliability is above .8; person separation index is above 1.5
- Counter evidence: Person reliability is below .8; person separation index is below 1.5

4. Explanation

Test performance varies according to the amount and quality of experience in learning Chinese

- Strong evidence:
    - There is a significant and meaningful change (medium to large effect) in students' test performance from the beginning of the semester to the end of the semester.
    - There is a significant and meaningful (medium to large effect) difference in test performance among students at different course levels.

- Weak evidence:
    - There is a statistically significant but less meaningful change (small effect) in students' test performance from the beginning of the semester to the end of the semester.
    - There is a statistically significant but less meaningful (small effect) difference in test performance among students at different course levels.
- Counter evidence:
    - There is no statistically significant change in students' test performance from the beginning of the semester to the end of the semester.
    - There is no statistically significant difference in test performance among students at different course levels.

Test scores support the internal structure of the construct

- Strong evidence: Compared to alternative models, there is strong evidence supporting a single factor (overall Chinese language ability) or a two-factor (reading and listening Chinese ability) model. The listening and reading items are loaded onto the corresponding factor.
- Weak evidence: Compared to alternative models, there is NO strong evidence against a single factor (overall Chinese language ability) or a two-factor (reading and listening Chinese ability) model. The listening and reading items are loaded onto the corresponding factor.
- Counter evidence: Compared to alternative models, there is strong evidence against a single-factor (overall Chinese language ability) or a two-factor (reading and listening

Chinese ability) model. The listening and reading items are not loaded onto the corresponding factor.

5. Extrapolation

Scores on the MSU Chinese placement test are positively correlated with scores on ACTFL proficiency tests.

- Strong evidence
    - There is a moderate to strong correlation ($r \geq 0.40$) between students' scores on the Chinese placement test and on ACTFL proficiency tests for corresponding skills (e.g., listening and listening).
    - There is a positive correlation between students' scores on the Chinese placement test and on ACTFL proficiency tests for non-corresponding skills (e.g., speaking and listening); however, the strength of this correlation is expected to be weaker compared to the correlation between corresponding skills.
- Weak evidence:
    - There is a weak positive correlation ($0.20 \leq r < 0.40$) between students' scores on the Chinese placement test and on ACTFL proficiency tests for corresponding skills (e.g., listening and listening).
    - The correlation between students' scores on the Chinese placement test and on ACTFL proficiency tests for non-corresponding skills (e.g., speaking and listening) is stronger, which is not consistent with the expectation of a weaker correlation compared to corresponding skills.

- Counter evidence:
    - There is no or a negative correlation ($-1.00 \leq r < 0.20$) between students' scores on the Chinese placement test and on ACTFL proficiency tests for corresponding skills (e.g., listening and listening).
    - The correlation between students' scores on the Chinese placement test and on ACTFL proficiency tests for non-corresponding skills (e.g., speaking and listening) is stronger, which is not consistent with the expectation of a weaker correlation compared to corresponding skills.

6. Utilization

Test users (i.e., instructors and students) judge the scores to be useful.

- Strong evidence: From test users' perspective, the test scores place most students, given their language ability levels, in appropriate Chinese language classes.
- Counter evidence: From test users' perspective, the test scores do not place most students, given their language ability levels, in appropriate Chinese language classes.

Cut-off scores are set appropriately

- Strong evidence:
    - There is an even distribution of items across all course levels.
    - The cut-off scores align well with the instructors' perceptions and the distribution of items, resulting in accurate placement of students in appropriate course levels.
- Weak evidence:
    - There is an imbalanced distribution of items across the course levels, which may lead to less accurate measurement of students' language proficiency at certain levels

    - The cut-off scores partially align with the instructors' perceptions and the distribution of items, but there is room for improvement in the accuracy of student placement.
- Counter evidence:
    - There is a highly imbalanced distribution of items across the course levels, leading to an inaccurate measurement of students' language proficiency at certain levels.
    - The cut-off scores do not align with the instructors' perceptions or the distribution of items, resulting in inaccurate placement of students in appropriate course levels.

7. Consequence implication

The test has positive effects on Chinese instruction and learning

- Strong evidence: Stakeholders indicate positive effects of the test on what teachers teach and how students learn.
- Counter evidence: Stakeholders indicate negative effects of the test on what teachers teach and how students learn.

CHAPTER 4: RESULTS

The descriptive statistics for the test scores of the 305 test takers who took the test between 2016 and 2020 (collapsed across years to further preserve anonymity) on the placement test are shown in Table 4. The table indicates that most students were advised to take 100- or 200-level courses, while only a small group of students with scores above 30 were recommended to take the 300-level course. As for the score comparison between female and male examinees, the descriptive statistics suggest that female and male examinees performed comparably on the placement test. In addition, examinees who were self-identified as heritage speakers (speaking the language at home) performed better than their peers who did not have family connections to the language. Not surprisingly, students who are self-identified as L1 speakers of Chinese (born and raised in a Chinese speaking country, having graduated from a Chinese-speaking high-school in that country, and at MSU as an international student to obtain a degree) perform exceptionally well on the test, with their scores being in close proximity to the maximum achievable score, as one would expect.

As described in the section on data analysis, I utilized Rasch measurement methods to identify potential problems and determine if revisions to the placement test should be made in order for appropriate uses and interpretations of test scores. The Rasch analysis used the item responses collected from 305 examinees who took the placement between 2016-2020. The analysis was conducted using the computer program, WINSTEPS Version 4.7.1 (Linacre, 2016).

Table 4. Descriptive statistics for test scores

	N	Mean (SD)	95% CI
Total score	305	21.4 (6.5)	[20.7, 22.1]
0 - 19	126	14.8 (2.90)	[14.3, 15.3]
20 - 29	137	24.5 (3.01)	[24.0, 25.1]
30 - 32	42	30.9 (.68)	[30.7, 31.1]
Gender			
Female	152	21.7 (6.8)	[20.6, 22.8]
Male	153	21.2 (6.3)	[20.2, 22.2]
Learner type			
Native speakers	15	30.2 (2.4)	[28.9, 31.5]
Heritage speaker	73	25.1 (5.91)	[23.7, 26.4]
Others	217	19.6 (5.87)	[18.8, 20.3]

Prior to conducting the main analysis, I first performed principal components analysis of the residuals, a standard Rasch approach to examine whether all items in the test could be considered unidimensional within the Rasch framework (Linacre, 1998). This is important because evidence of unidimensionality is required for Rasch analysis to yield accurate and reliable measurements of the underlying construct being assessed (Eckes, 2015). Evidence of multidimensionality is indicated by an eigenvalue greater than 2.0 for the first factor in the PCA and by a disattenuated correlation less than 1. The PCA of the total 32 items revealed evidence of multidimensionality which was indicated by an eigenvalue greater than 2 for the unexplained variance in the first factor. A close examination of the test showed that five items (see Appendix 4 for detailed information about these items) that were related to a common stimulus loaded heavily on the same dimension (see Table 5). In Rasch modeling, the group of items or the questions related to the same topic or prompt in a test is known as a *testlet* (Wang et al., 2005), which often results in locally dependent items (a form of violation of the assumption of

unidimensionality). To address the issue, I bundled the five items into a polytomous super-item and re-conducted the PCA. After the revision, the results of the PCA provided no evidence of multidimensionality as indicated by an eigenvalue of 1.92 and by a disattenuated correlation equal to 1.00 between the theta measures (i.e., test-takers' ability levels) on items clusters in contrasts.

Table 5. PCA results for the five items that loaded on the same dimension

Item No.	Loading	Measure	Infit MNSQ	Outfit MNSQ
Reading 10	.59	-.99	.84	.71
Reading 7	.49	-1.99	.87	.43
Reading 9	.49	-1.55	.91	.93
Reading 8	.45	-1.50	.89	.74
Reading 11	.28	-.95	1.00	.90

I then considered item fit and evaluated whether all items in the test measured the construct, Chinese language proficiency, as intended. An item is considered to have a good fit when it generates item responses that align with what the model predicts. Fit to the model associated with each item is assessed by two Rasch-based statistics, mean square outfit and mean square infit. Infit and outfit mean squares have an expected value of 1.0, which suggests that the item generates responses that align with what the model predicts (e.g., a more difficult item would likely elicit a correct response from a proficient test-taker but would likely elicit an incorrect response from a less proficient test-taker). Outfit mean squares are more sensitive to unexpected responses by persons on items that are very easy or very hard for them, whereas infit mean squares weigh the observations by their statistical information and are consequently more sensitive to unexpected responses by persons on items roughly targeted on them (Linacre, 2016). I employed the cut-off values of 0.6 and 1.4 as acceptable infit and outfit mean squares, as

suggested by Wright and Linacre (1994). A higher or lower value of the fit statistics indicates that the responses generated by the associated item were either too predictable (less than 0.6, overfit the model) or too unexpected (larger than 1.4, underfit the model). Finally, the items with the extreme outfit or infit mean square values were flagged as misfitting items and were closely examined by me to inspect reasons for the misfit.

The infit and outfit statistics suggested that out of 32 items, three items (see Table 6 for fit statistics), reading items #2, #3, and #12 displayed outfit mean square values outside the cut-off range. These misfitting items all had large outfit values, suggesting that the items elicited a few unexpected responses from test takers given their ability levels and the item difficulties. More specifically, the responses were considered unexpected when a high-ability test-taker failed to answer an easy item correctly or a low-ability test-taker answered a difficult item correctly. The finding was further confirmed by the low item discriminations for reading items #2 and #12. In other words, the items had a limited capacity to discriminate between the test-takers with higher proficiency and those with lower proficiency. I returned to these misfitting items and proposed strategies for addressing them in the section of [RQ 2c: Evaluation inference].

Table 6. Misfitting items from the MSU placement test

Item	Difficulty estimate (SE)	Outfit MNSQ (z-std)	Infit MNSQ (z-std)	Estimated discrimination
Reading #2	-.21 (.14)	1.52 (3.03)	1.29 (4.59)	.32
Reading #3	-1.50 (.18)	2.51 (3.67)	1.09 (.86)	.80
Reading #12	1.16 (.14)	1.51 (4.69)	1.31 (4.37)	.38

[RQ 1: Domain description inference]:

Are the relevance of the test items and test criteria to the instructional domain and the appropriateness of the item difficulties supported by test stakeholders?

The research question related to domain description inference was addressed by examining the results of instructors' and students' questionnaire data about test content relevance and item difficulties. As noted earlier, to evaluate the relevance and appropriateness of test items to the course materials for 100-level, 200-level, and 300-level courses, instructors were provided with a set of checkboxes for each item. Furthermore, instructors were given an extra checkbox to indicate that the item was not relevant to the content of any of the three courses. The number of test items relevant to course content by instructor and course is presented in Table 7. The data presented in Table 7 indicates that the instructors demonstrated some divergence of opinion with respect to the course level to which an item was relevant. Nevertheless, a clear trend emerged, showing that most of the test items were considered to be more relevant to lower-level courses than to higher-level courses. More importantly, it should be noted that none of the items were judged by any of the instructors as irrelevant to the course material across all three levels of instruction.

Table 7. Number of test items relevant to course content by instructor and course level

	100-level course	200-level course	300-level course	Irrelevant to any course
Instructor 1	20	7	5	0
Instructor 2	19	8	5	0
Instructor 3	21	9	2	0

As for the students' questionnaire, students were asked to rate test items on two 6-point Likert scales in terms of the overall item difficulty (1: *very easy*; 6: *very difficult*) as well as the relevance and appropriateness to the course material they were studying (e.g., 1 = *the item is NOT relevant to the course that I am taking;* 6 = *the item is highly relevant to the course that I am taking*). Table 7 provides descriptive statistics that illustrate how students perceive the relevance and difficulty of test items, as reported by course level. Further details regarding individual items can be found in Appendix 5. Data are presented by course level to account for the potential influence of the specific level of the course in which students are enrolled on students' perceptions of item difficulties and relevance. Table 8 and Appendix 5 further support this notion, showing that students' perceptions of item relevance to course content and item difficulties vary by course level. Specifically, in terms of item difficulty, as expected, students in higher-level courses found test items easier compared to those in lower-level courses. As for item relevance to course content, students in 100- and 200-level courses gave higher mean relevance scores compared to those in 300-level courses. This observation aligns with instructor ratings that considered the majority of items to be relevant to lower-level courses. Despite variations in students' perceptions of item relevance across different course levels and test items, there is a general pattern indicating that students view test items as relevant to their course, as evidenced by their relatively high mean relevance scores.

Table 8. Descriptive statistics of students' perceptions of test item relevance and difficulties

	Relevance			Difficulties		
	Mean	SD	95% CI	Mean	SD	95% CI
100-level	4.7	0.8	[4.4, 5]	3	0.9	[2.7, 3.3]
200-level	5.1	0.5	[4.9, 5.2]	2.7	0.8	[2.4, 3]
300-level	4.1	0.4	[3.9, 4.2]	2.2	1	[1.9, 2.6]
Total	4.7	0.5	[4.5, 4.8]	2.7	0.8	[2.4, 3]

[RQ 2a: Evaluation inference]:

Do test items yield item difficulty estimates that are appropriate for making placement decisions?

I examined the research question related to evaluation inference using two distinct methods. First, I applied a Rasch-based approach. Rasch modeling shares an important feature with other item response theory-based models: items and examinees are estimated and compared on a single common scale that is interval-level. This allows for an accurate comparison of examinees' abilities based on equal distances on the scale. Rasch measurement aims to achieve the highest precision in estimating an examinee's ability when the ability estimate aligns with the item difficulty, a concept known as *targeting* (Bond & Fox, 2015, p. 69).

The relationship between a person's ability and item difficulty is of significant interest to researchers. A Wright map is commonly used to visualize this relationship, providing a meaningful representation of the data. To determine if the item difficulties accurately depict students' abilities and if the test is sensitive to variations in the measured construct, I considered three key aspects as outlined by Beglar (2010): (a) the adequacy of the number of items included in the test; (b) the presence of targeting for the sampled examinees; and (c) any potential gaps in the empirical item hierarchy. By following Beglar's guidelines, I employed the Wright map to

investigate these factors and evaluate the appropriateness of the item difficulty estimates for making placement decisions.

Figure 2 displays the Wright map, where the ruler on the left (MEASR) indicates the logit values corresponding to test takers' ability levels and item difficulties, both measured on the same scale. The item difficulties in Figure 1 range from -2.34 to +1.78 logits, while the test-takers' ability measures cover a broader range (from -1.78 to +4.96[1]). A thorough examination of the Wright map reveals that 75% of the examinees (N = 229) fall within the overlapping range, indicating reasonable item targeting along their ability level. However, the Wright map also reveals a noticeable ceiling effect, with approximately one-fourth of the examinees (N = 76) having ability measures above all item difficulties.

This finding aligns with the descriptive statistics in Table 3, which show that 42 examinees scored 30 or higher, suggesting a need for additional, more challenging items to accurately assess these high-ability examinees. Notably, the background questionnaire results reveal that among these 42 high-ability examinees, 15 are L1 Chinese speakers and 18 are heritage Chinese speakers. Heritage speakers either were born in the United States with at least one parent from China who spoke Chinese at home or were immigrants who moved to the United States from China at a young age. Given the language backgrounds of these 33 examinees (15 L1 and 18 heritage speakers), their high test performance is not surprising. Their linguistic exposure and experiences may have provided them with an advantage on the test, resulting in higher scores.

[1] Eight test takers received a perfect score on the test, and their ability measures (+ 4.96) were not plotted on the Wright map.

Figure 2. Wright map of the MSU Chinese placement test items.

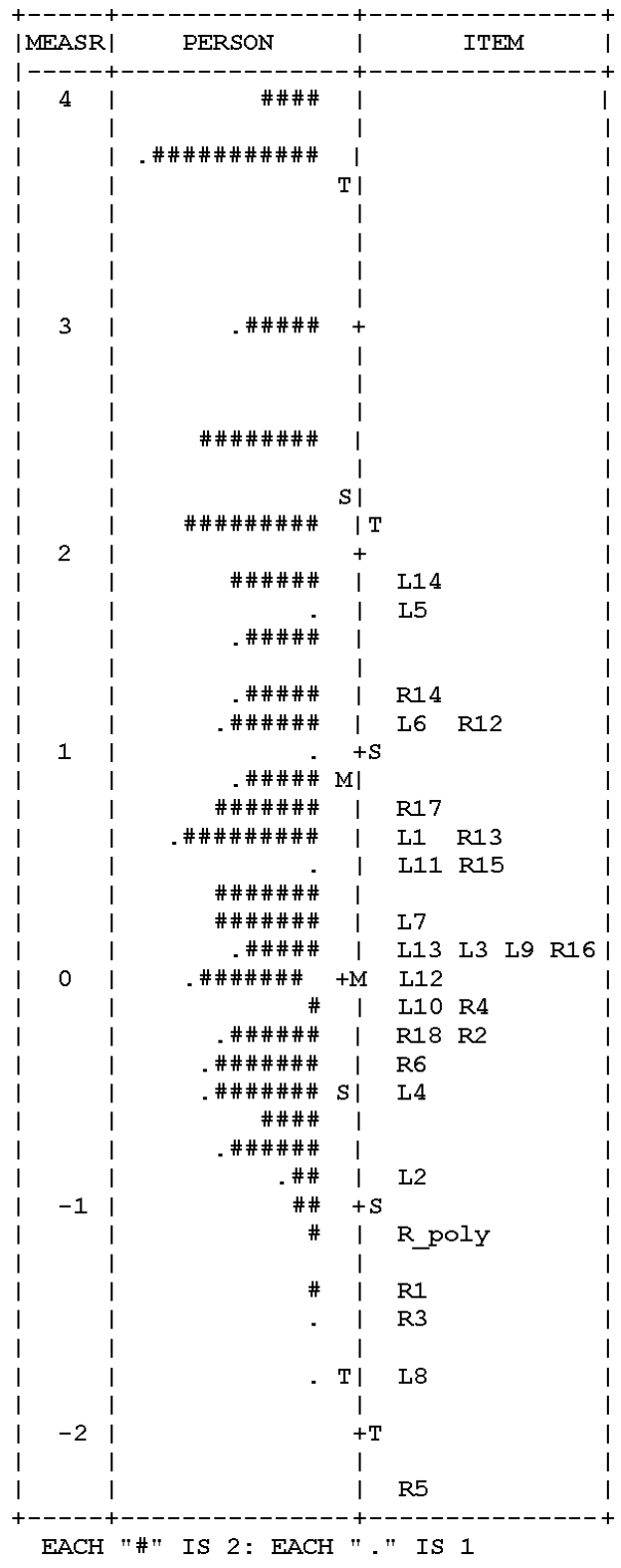
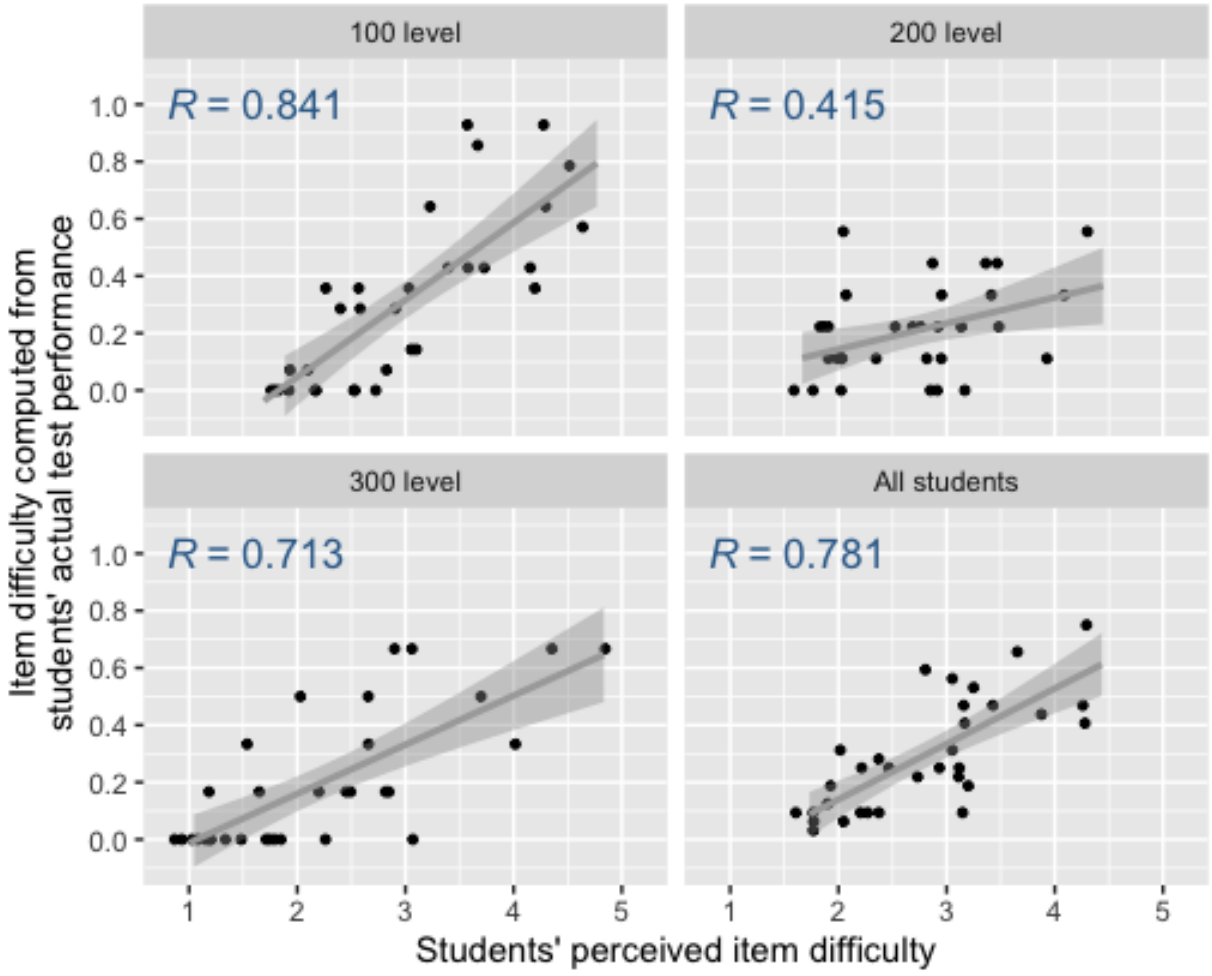


Figure 3. Relationship between students' perceived item difficulties and item difficulties computed from students' actual test performance.



The second approach I employed to address this research question involved using correlation analysis with Pearson's correlation coefficients. This analysis aimed to examine the agreement between students' and teachers' difficulty ratings, comparing these perceived difficulties with the empirical item difficulties obtained from the quantitative item analysis. Gaining insights from these different perspectives can be valuable in determining the test's suitability for making placement decisions (Embretson & Reise, 2000; Downing & Haladyna, 2006). Discrepancies between perceived and empirical item difficulties may point to issues with

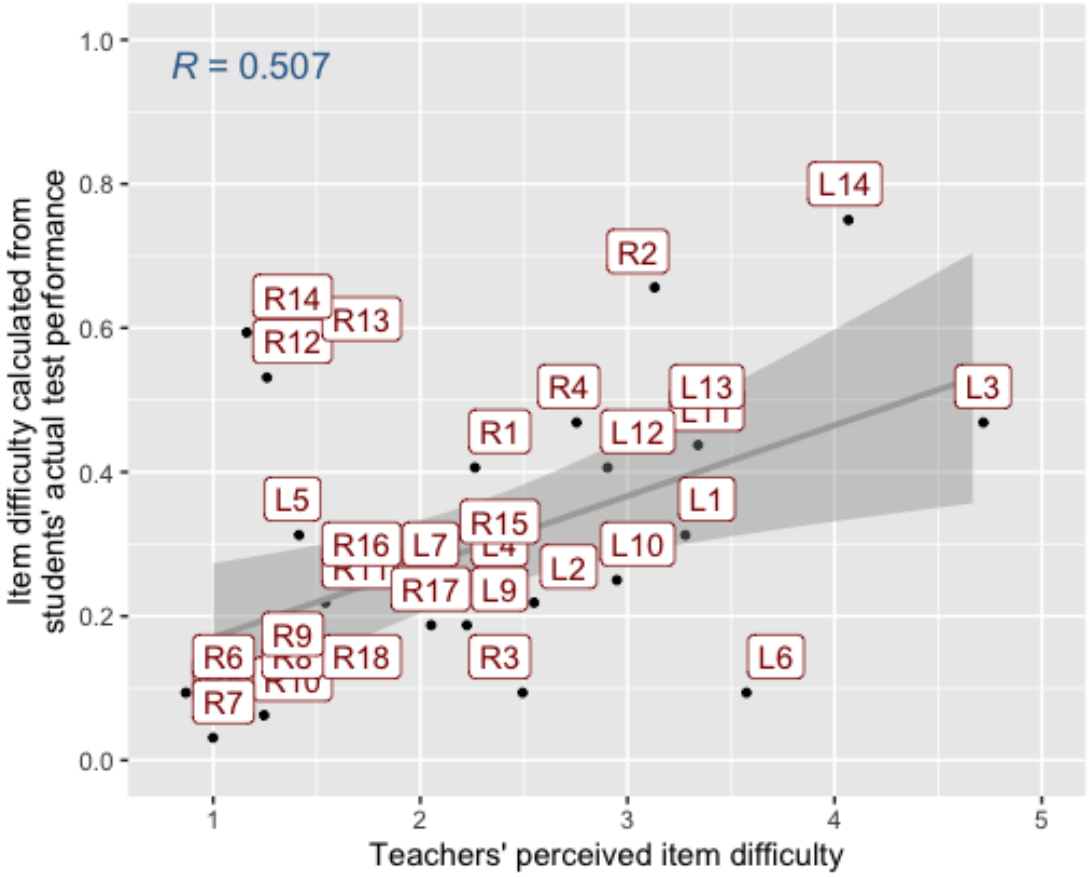
the test items. For instance, if students or teachers perceive certain items as more challenging than the empirical analysis suggests, this could indicate that the items contain ambiguous or unclear wording, causing confusion among examinees (Haladyna, Downing, & Rodriguez, 2002). Conversely, a high degree of agreement between students' and teachers' ratings and the empirical difficulties would support the test's appropriateness for placement purposes (DeMars, 2010). If perceived and empirical item difficulties align closely, this implies that the test items function as intended, accurately measuring the targeted construct and differentiating examinees based on their abilities (Hambleton, Swaminathan, & Rogers, 1991). Such close alignment would bolster confidence in using the test for placement decisions, as the items would provide a valid and reliable representation of examinees' abilities in the target domain (AERA, APA, & NCME, 2014).

To calculate Pearson's correlation coefficients, I computed the average ratings of item difficulties across students for each of the 32 items in the placement test, representing their perceived difficulties. Likewise, I calculated the average rating across teachers for each item. Figure 3 illustrates the relationship between students' perceived item difficulties and the empirical item difficulties derived from their actual test performance. The figure comprises four plots, each representing a different course level (100-level, 200-level, and 300-level) and one for the aggregated data. Each point in the scatterplots corresponds to an individual item in the test. As evident from the plots, the correlation between students' perceived item difficulties and empirical item difficulties varies across course levels. For the 100-level courses, the correlation coefficient is quite high at .841, suggesting a strong agreement between perceived and empirical item difficulties. For the 200-level courses, the correlation coefficient is lower at .415, indicating a weaker relationship between the two sets of item difficulties. In contrast, the correlation

coefficient for the 300-level courses reveals a moderate to strong relationship at .713. Examining the aggregated data, the overall correlation coefficient is .781, demonstrating a robust relationship between students' perceived item difficulties and the empirical item difficulties.

These findings imply that the relationship between students' perceptions of item difficulties and the empirical item difficulties depends on the course level. However, the strong overall correlation in the aggregated data suggests that the test items generally align well with students' perceptions, supporting the test's appropriateness for making placement decisions.

Figure 4. Relationship between teachers' perceived item difficulties and item difficulties computed from students' actual test performance.



Following the analysis of students' perceptions, Figure 4 presents a scatterplot depicting the relationship between teachers' perceived item difficulties and the empirical item difficulties computed from students' actual test performance. The correlation coefficient for this comparison is .507. While most items are situated along the regression line in the scatterplot, signifying agreement between teachers' perceptions and empirical item difficulties, a few items deviate from this trend.

For instance, reading items #12, #13, and #14 were perceived as very easy by the teachers (with a mean value around 1.3 on the Likert scale of 1 to 6), but the empirical item difficulty is around .6, suggesting that these items are among the most challenging in the test. Conversely, listening item 6 was regarded as an easy item by the teachers, while the empirical item difficulty is .1, indicating that most students did not encounter difficulty with this item. These findings reveal some discrepancies between teachers' perceptions of item difficulties and the actual item difficulties experienced by students. As a result, the relationship between teachers' perceptions and empirical item difficulties is not as strong as one might anticipate.

Figure 5. Relationship between teachers' perceived and students' perceived item difficulties.

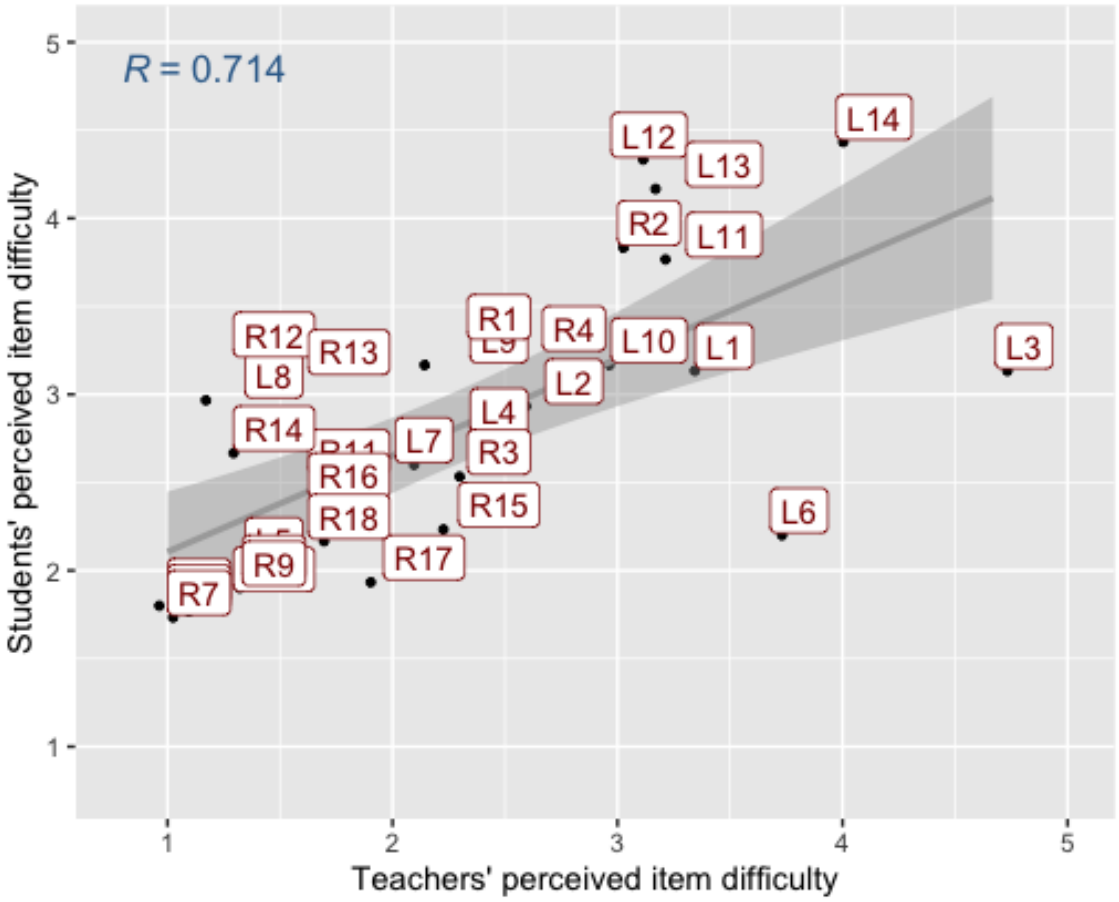


Figure 5 presents a scatterplot illustrating the relationship between teachers' perceived and students' perceived item difficulties, with a correlation coefficient of .714. While most items display a strong alignment between students' and teachers' perceptions of item difficulty, one item notably deviates from the regression line. Listening item #6, perceived as a relatively easy item by students (with a mean value of 2.2 on the Likert scale of 1 to 6), received higher difficulty ratings from teachers (with a mean value of 3.7 on the Likert scale of 1 to 6). Generally, the findings reveal a strong relationship between teachers' and students' perceptions of item difficulty, suggesting that both groups have similar views on the test items' difficulty. This alignment supports the notion that the test items are generally suitable for assessing students'

abilities. However, the observed discrepancy for listening item #6 implies that there might be a difference in how teachers and students interpret or understand this particular item. Factors such as differences in instructional focus, students' familiarity with the content, or other influences could impact their respective judgments.

In summary, the results suggest that the test items generally provide item difficulty estimates appropriate for making placement decisions, particularly for the 100-level and 300-level courses. However, some discrepancies and weaker relationships have been observed, especially for 200-level courses, necessitating further investigation and refinement of the test items to ensure their suitability across all course levels.

[RQ 2b: Evaluation inference]:

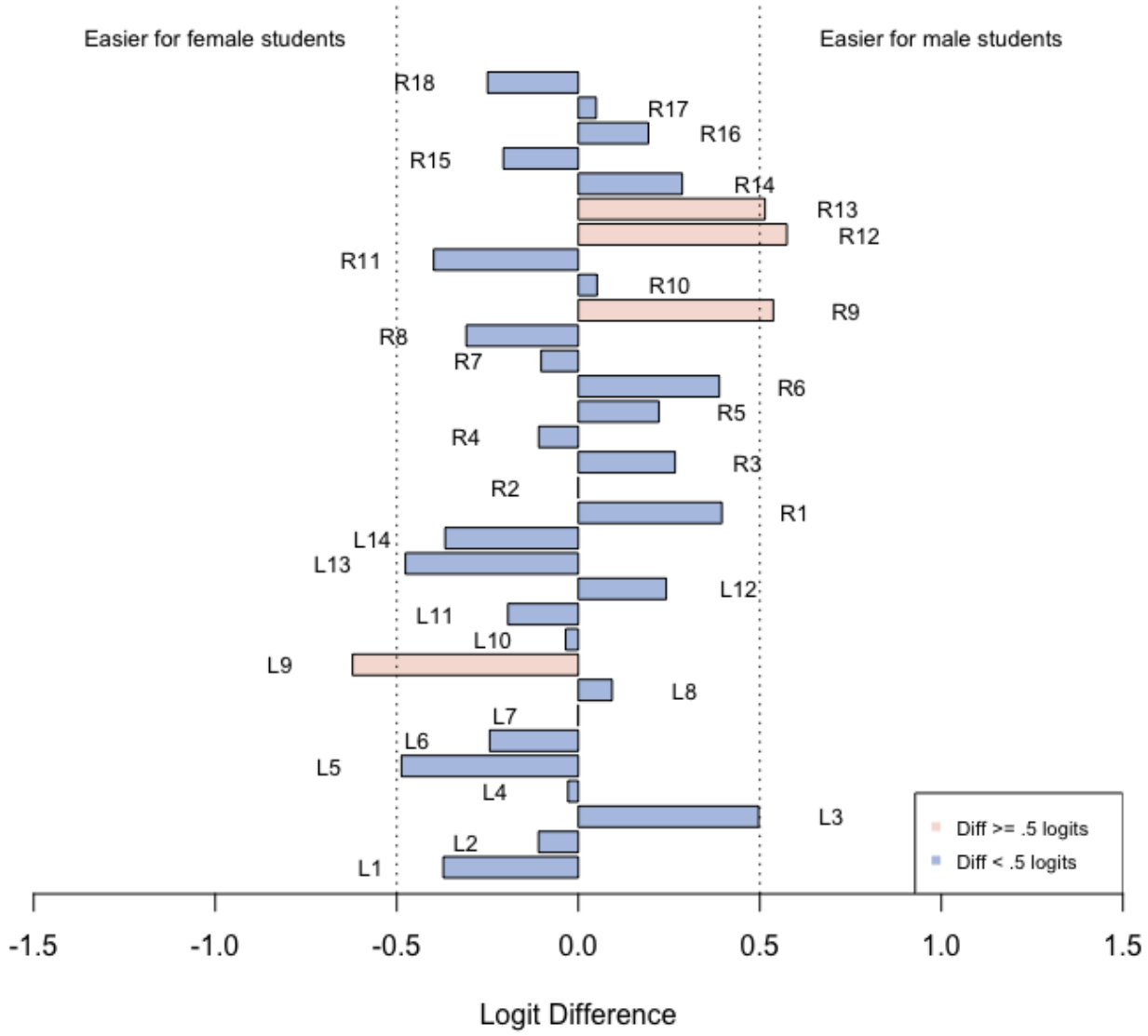
Do test items exhibit no evidence of item bias?

Building on the previous discussion, one key aspect to consider when examining the validity evidence for evaluation inference is the principle of invariance. According to this principle, item measures should remain invariant across different measurement contexts, meaning that item estimates (i.e., item difficulty estimates) should not depend on the subgroups of examinees responding to the item (Baker & Kim, 2017; Rasch, 1960). In this study, I investigated the extent to which item estimates are invariant across two examinee groups: female versus male examinees. Ensuring item invariance between female and male examinees is crucial to guarantee that the test items do not favor one gender over the other, thus providing evidence of fairness and impartiality in the assessment (Kunnan, 2000).

I analyzed the group invariance of item measures by examining differential item functioning (DIF). Specifically, DIF is detected for an item when two groups of examinees, matched on measures of the construct (Chinese language ability in this case), have different

probabilities of answering the item correctly (Ferne & Rupp, 2007; Harding, 2011). If DIF is found, it suggests that the item may be biased, challenging the validity evidence for the evaluation inference. I established two criteria for detecting potential DIF relative to the item difficulty estimates based on the responses of the two groups of examinees: a) statistical significance of the Mantel-Haenszel test at the .05 level after the Benjamini-Hochberg adjustment to correct for the inflation of Type I error due to multiple comparisons; b) a difference in item difficulty of at least .5 logit, considered large enough to impact ability estimates (Linacre, 2016). Items meeting both criteria were considered to show evidence of DIF. The results revealed that while four items (listening item #9, reading items #9, #12, and #13) exhibited a difference in item difficulty larger than .5 logit (see Figure 6), none of them were significant at the predetermined alpha level, suggesting no evidence of DIF across these two examinee subgroups (For more detailed information on the DIF analysis results, please refer to Appendix 6).

Figure 6. Bar plot of item difference (results of DIF).



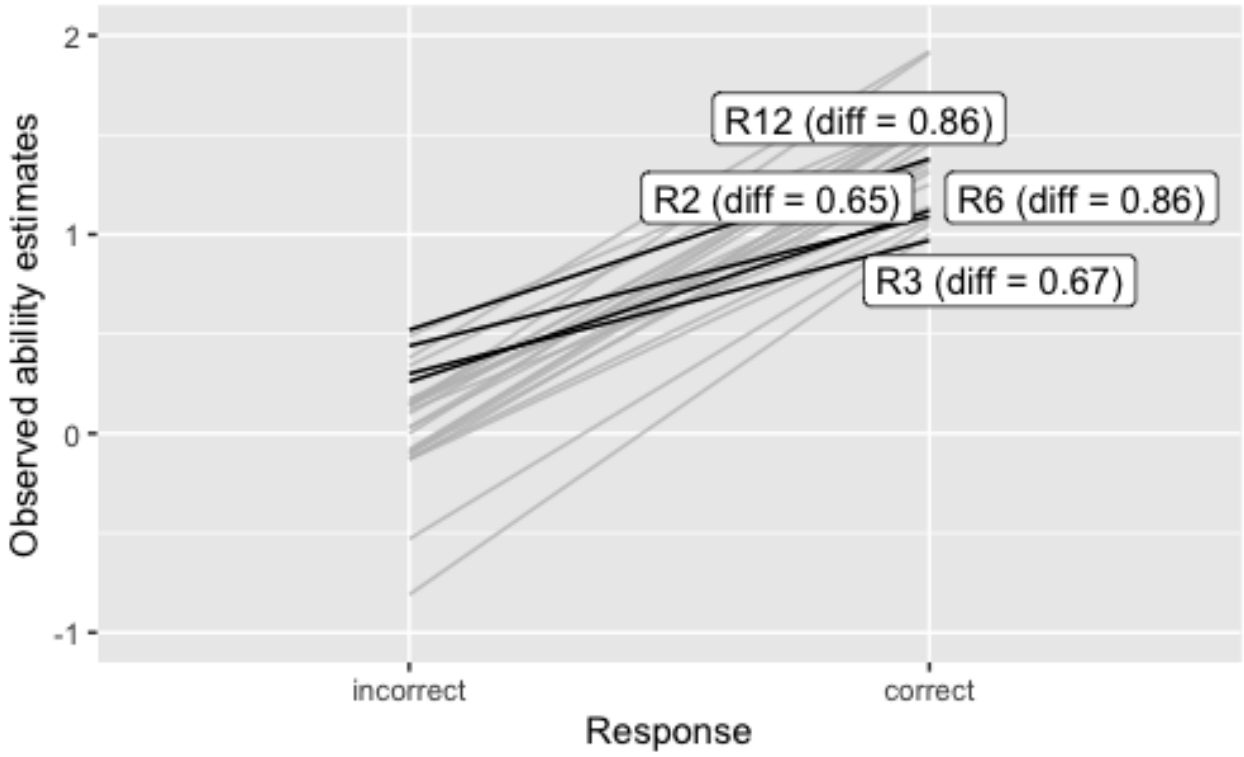
[RQ 2c: Evaluation inference]:

Are correct options unambiguous and accurately keyed?

Continuing from the previous section, it is important to emphasize that a well-constructed multiple-choice test should include effective distractors that challenge examinees, requiring them to demonstrate their language abilities to select the correct response among plausible alternatives. Therefore, examining distractors is an essential aspect of investigating the validity evidence for the evaluation inference, as it helps to determine whether the test items are functioning as intended and whether the correct options are unambiguous and accurately keyed. To address the research question, I conducted an analysis of distractors as an item quality indicator for all test items, aiming to assess the extent to which distractors for each item discriminated between examinees with different ability levels.

I compared the average ability estimates of examinees who selected the distractors and the keyed option. Theoretically, distractors should attract examinees with lower ability estimates on average, compared to those who select the keyed option (Wolfe & Smith, 2007; Osterlind, 1998). Figure 7 presents the mean ability estimates of examinees who chose the keyed options and those who did not for each item. As shown, the keyed options generally attracted higher-ability examinees compared to the distractors, as demonstrated by the upward lines. However, four items exhibited lower discriminating power, as indicated by their less steep lines. These items had mean ability estimate differences between the two groups of examinees of less than 1. Given that examinee ability estimates ranged from -1.78 to 4.96, these differences may not be practically or meaningfully significant in discriminating examinees' language abilities (Downing & Haladyna, 2006).

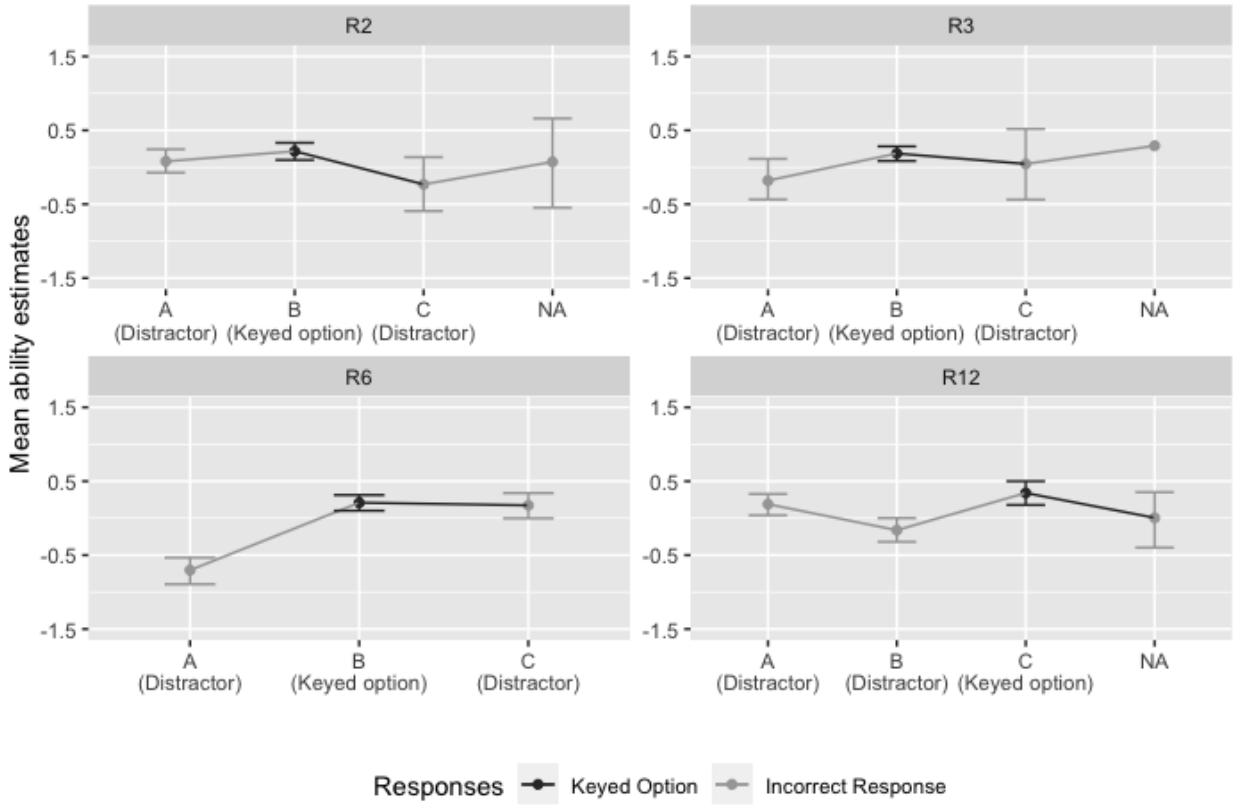
Figure 7. Items flagged as having potentially problematic distractors.



The functioning of the distractors in the four flagged items was further investigated. Figure 8 displays the mean ability estimates for each response option for these items. As illustrated in the graph, the mean estimates of examinees who selected the keyed option did not significantly differ from those who chose one of the distractors. This suggests that these distractors may be too similar or equally appealing to the keyed option, leading to examinees with similar ability levels choosing either option. The overlapping 95% confidence intervals of the mean ability estimates for the keyed option and the potentially problematic distractor (distractor A for Reading item #2; distractor C for Reading item #3; distractor C for Reading item #6; distractor A for Reading item #12) further confirm this observation. These findings indicate the need for a more detailed review of these items to ensure that the keyed options are clear and unambiguous and that the distractors are not too close in plausibility to the keyed

option, which could lead to inaccurate measurement of examinees' abilities. This point will be revisited later in this section to discuss possible improvements to the test items.

Figure 8. Mean ability estimates for distractor analysis.



Note: a) Error bars were 95% confidence intervals generated using non-parametric bootstrap; b) NA indicates missing data; c) there was only one participant that had missing data for Reading item #3 and no participant that had missing data for Reading item #6

Following the previous analysis, a closer examination of the four misfitting items revealed three potential reasons for the item misfit. These include: a) ambiguous phrasing in the item prompts, which may result in multiple correct responses; b) poorly selected distractors, leading to more than one response being justifiable as correct answers; and c) incongruent information provided in the prompt compared to the intended answer (Downing & Haladyna, 2006). These issues may have caused highly proficient test-takers to provide incorrect responses, negatively affecting the measurement of their Chinese language abilities. Given these concerns, it is essential to revise or remove these misleading items from the test, as they do not reliably assess test-takers' Chinese language abilities. Failing to address these issues could hinder meaningful score interpretations and compromise the test's appropriateness for making placement decisions.

Figure 9. Reading #3. If you want to order soup, how many choices do you have?

Reading item #3 (refer to Appendix 7) serves as an example of an ambiguous item. In this item, examinees were asked to count the number of soup dishes on a Chinese menu. The answer key indicated four dishes, which could be deduced by counting the dishes under the soup category (tāng lèi 汤类). However, several high-scoring examinees provided a different response, suggesting there were five choices if one were to order soup (see Figure 9). Upon closely examining each dish on the menu, I discovered one dish, seafood rice noodle soup (hǎixiān mǐfěn tāng), listed under the porridge and noodle soup category (zhōu, tāngmiàn lèi 粥，汤面类). This dish could arguably be considered a soup dish. The item's ambiguity led some high-performing examinees to include a dish that was not intended as part of the correct answer. However, one might argue that for these examinees, the reading processes involved (i.e., the ability to read and comprehend each dish on the menu) are indicative of higher ability. In light of these findings, it is recommended to either remove the item or revise it as suggested in Appendix 7 to ensure clarity and accurate assessment of examinees' abilities.

Figure 10. Reading item #6. Here is a message that Xiao Li sent to Lao Wang. Please answer the following questions after reading the note: At what time, should they meet?


老王：

后天晚上请你七点来我家吃晚饭，我六点三刻在三路车站等你。后天见。

小李 四月十八号 星期五

The issue of item ambiguity arose in another instance, specifically with reading item #6. The stem provided a note—an invitation to a friend—which was translated into English as follows: "I would like to invite you to come to my home for dinner at 7:00 PM the evening after tomorrow. I will be waiting for you at 6:45 PM at the Route 3 bus stop. See you the day after tomorrow." The item asked examinees to determine the time when the two individuals should

meet. Although the incorrect responses from several high-scoring examinees might be attributed to carelessness, varying interpretations of the invitation could also stem from cultural differences. In some cultures, guests may be expected to arrive at the scheduled dinner time. To address the ambiguity, I have provided suggested revisions in Appendix 7.

Figure 11. Reading item #2. Is this a sign for?

Regarding reading item #2, a thorough examination revealed that the information in the prompt did not align perfectly with the designated answer. The item assessed whether test takers could deduce the purpose of a sign based on accompanying pictures and Chinese descriptions. The answer key, 'shopping mall hours,' corresponded with the Chinese descriptions (yíngyè shíjiān, zǎo 6:00 - wǎn 10:00; 营业时间, 早 6:00-晚 11：00; operating hours, 6AM–10 PM). However, the sub-signs below (jìnzhǐ wàishí; 禁止外食; no outside food allowed) indicated that

the sign was more likely intended for a restaurant or movie theater, where outside food or beverages might be prohibited for health reasons, or to ensure the owner profits from food and beverage sales. In my proposed revision, I have amended the answer key to resolve this inconsistency.

The problem related to problematic item distractors occurred with another misfitting item, reading item #12. This item tested test takers' knowledge about the appropriate classifier used for chairs (yǐzi,椅子) in Chinese. The correction option is 把 (bǎ, classifier), while a few high-scoring test takers selected 张 (zhāng) as their response (see Figure 3). Research on Chinese classifiers indicates that multiple correct answers might exist for this question. For instance, Tai's (1994, p. 9) description of the classifier, zhāng, aligns with the responses of these high-scoring test takers:

'[i]t is a well-known fact that Mandarin Chinese use the classifier *zhāng* (张, classifier) to categorize *zhǐ* 'paper', *zhuōzi* 'table', and *chuáng* 'bed'. For many native speakers of Mandarin, the category of *zhāng* extends to cover *yǐzi* 'chair' and *dèngzi* 'bench', since they all have a flat surface like tables, the central member among the class of furniture categorized by *zhāng*.'

Furthermore, I investigated the association between yǐzi (chair) and its corresponding classifier in the Modern Chinese Corpus (现代汉语语料库, see the link: http://corpus.zhonghuayuwen.org). The findings indicated that the co-occurrence frequency of bǎ (classifier) and yǐzi was 45, while the frequency for zhāng (classifier) and yǐzi was 17. This suggests that both classifiers are actively and frequently used by native Chinese speakers, albeit to different extents. Consequently, the inclusion of zhāng as a distractor in the item was deemed inappropriate for placement purposes, as it does not effectively differentiate between test takers

with higher and lower proficiency levels. As a result, I recommend replacing this distractor with a new classifier, as illustrated in Appendix 7.

[RQ 3a: Generalization inference]:

Does the MSU Chinese placement test produce scores that are internally consistent?

Examining the internal consistency of test scores is crucial for supporting the generalizability inference, as it demonstrates the test items' reliable measurement of the intended construct—in this case, Chinese language proficiency. High internal consistency establishes confidence in the stability and accuracy of test scores across diverse settings and student groups. The MSU Chinese placement test's internal consistency was analyzed by calculating Cronbach's α, yielding a value of 0.88 (95% CI: [0.86, 0.90]), indicating strong internal consistency. An item analysis was also conducted to assess the impact of individual items on overall internal consistency (see Table 9). The results showed that removing certain items would lead to a slight decrease in Cronbach's α from 0.88 to 0.87, while for others, the α value would remain unchanged at 0.88. These findings suggest that the test items collectively measure the same underlying construct, with no individual item significantly affecting overall internal consistency. In summary, the results of this study provide strong evidence supporting the internal consistency of the MSU Chinese placement test scores.

Table 9. Cronbach's α if item dropped

Item	Cronbach's α	Item	Cronbach's α	Item	Cronbach's α	Item	Cronbach's α
L01	0.87	L09	0.88	R03	0.88	R11	0.88
L02	0.88	L10	0.87	R04	0.88	R12	0.88
L03	0.87	L11	0.87	R05	0.88	R13	0.88
L04	0.88	L12	0.88	R06	0.88	R14	0.87
L05	0.88	L13	0.87	R07	0.88	R15	0.87
L06	0.88	L14	0.88	R08	0.88	R16	0.87
L07	0.88	R01	0.88	R09	0.88	R17	0.88
L08	0.88	R02	0.88	R10	0.88	R18	0.88

[RQ 3b: Generalization inference]:

Are there adequate items to reliably differentiate students' abilities into three levels as intended?

Another key aspect of evaluating the generalization inference of a test involves assessing the test's ability to differentiate between students' abilities at different levels. Specifically, if a test is intended to place students into multiple ability levels (such as beginner, intermediate, and advanced), it is crucial to ensure that the test contains enough items that are appropriately calibrated to accurately differentiate between students' abilities and assign them to the correct level (Bachman & Palmer, 2010). Without a sufficient number of items that reliably differentiate students' abilities, there may be a risk that students are placed in incorrect levels, leading to inaccurate or inconsistent results. Such misplacement can significantly impact students' language learning experiences, as they may be placed in courses that are either too easy or too challenging for their actual abilities (Alderson, 2005). To investigate this inference, the Rasch measurement model is utilized, providing person reliability and person separation indices as reliability estimates. These indices determine if the test sufficiently discriminates the sample into the intended levels. Low person reliability or separation suggests that the instrument may not

effectively distinguish between high and low performers. Linacre (2012) offered guidelines for interpreting person reliability: 0.9 corresponds to 3 or 4 levels, 0.8 to 2 or 3 levels, and 0.5 to 1 or 2 levels. Regarding person separation indices, 1.50 represents an acceptable level of separation, 2.00 a good level, and 3.00 an excellent level (Wright & Masters, 1982; Fisher, 1992, as cited in Duncan et al., 2003).

The MSU Chinese placement test's reliability indices were as follows: person reliability = .84; person separation index = 2.32. These results indicate that there were enough examinees and items to precisely locate examinees' abilities on the underlying trait continuum (i.e., Chinese language proficiency) and confirm the hierarchy of examinees' abilities. The placement test effectively discriminated examinees into three levels, as intended by the test developers, given the person reliability value.

[RQ 4a: Explanation inference]:

Does students' test performance vary according to the amount and quality of prior Chinese learning experience?

To address the research question, I analyzed students' performance at the beginning (Time 1) and end (Time 2) of the semester, as well as across different course levels. This analysis enables a deeper understanding of the Chinese placement test's sensitivity to students' prior learning experiences, thereby providing crucial validity evidence for the test. Descriptive statistics were calculated for students' total, listening, and reading scores on the Chinese placement test at both time points during the Spring semester 2022. These results are presented in Table 10.

Table 10. Descriptive statistics of students' placement test scores

	Time 1			Time 2		
Section	Mean	SD	95% CI	Mean	SD	95% CI
Listening	8.5	3	[7.3, 9.7]	9.9	2.3	[9, 10.8]
Reading	11.8	3.2	[10.5, 13]	13.8	2.4	[12.8, 14.7]
Total	20.2	5.7	[18, 22.5]	23.6	4	[22.1, 25.2]

Table 11. Descriptive statistics of students' placement test scores by course level

	Time 1			Time 2		
100-level	Mean	SD	95% CI	Mean	SD	95% CI
Listening	6.9	2.7	[5.3, 8.4]	9.3	2.2	[8, 10.5]
Reading	9.4	2.1	[8.2, 10.6]	12.9	1.3	[12.1, 13.6]
Total	16.3	4.2	[13.9, 18.7]	22.1	2.7	[20.6, 23.7]
200-level	Mean	SD	95% CI	Mean	SD	95% CI
Listening	10	3	[7.5, 12.5]	10.6	2.4	[8.6, 12.6]
Reading	13.4	2.9	[11, 15.8]	14.1	3.6	[11.1, 17.2]
Total	23.4	5	[19.2, 27.6]	24.8	5.5	[20.2, 29.3]
300-level	Mean	SD	95% CI	Mean	SD	95% CI
Listening	10.3	2	[8.3, 12.4]	10.3	2.3	[7.9, 12.8]
Reading	15	0.9	[14.1, 15.9]	15.3	1.2	[14.1, 16.6]
Total	25.3	2.7	[22.5, 28.1]	25.7	3.2	[22.3, 29]

The data reveals that, on average, students' listening, reading, and total scores on the Chinese placement test improved from the beginning to the end of the semester. To further investigate the impact of course level on test scores, descriptive statistics for students' total, listening, and reading scores were calculated by course level and presented in Table 11. Additionally, boxplots and summary statistics were utilized to display the distribution of test

scores, mean scores, and 95% confidence intervals for the mean in error bars (Figure 12: total scores; Figure 13: listening scores; Figure 14: reading scores).

The results demonstrate that students at different course levels exhibited varying degrees of improvement in their listening, reading, and total scores on the Chinese placement test throughout the semester. Higher-level students generally achieved better scores than lower-level students. The most significant gains were observed among the 100-level students, followed by the 200-level students, while the least change in scores occurred for the 300-level students. This pattern was evident at both the beginning and the end of the semester.

Figure 12. Boxplots for total test scores by course level and test time.

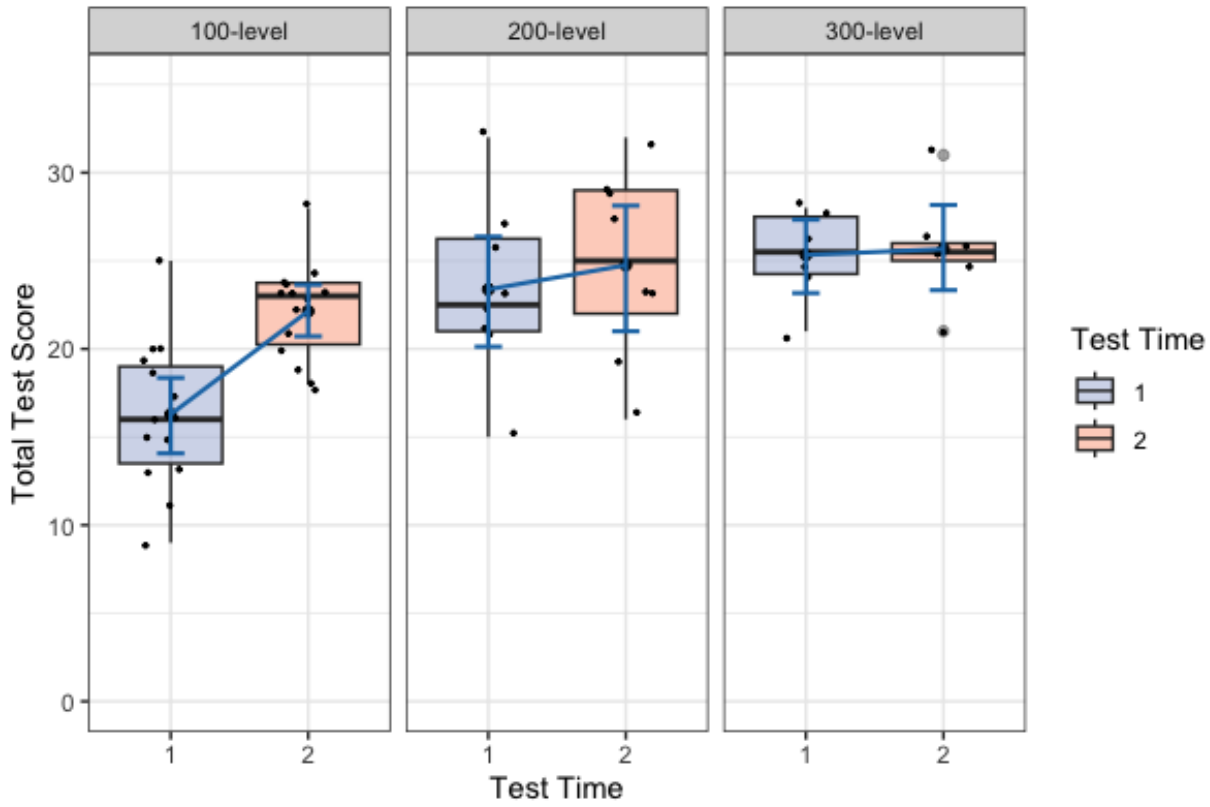


Figure 13. Boxplots for listening test scores by course level and test time.

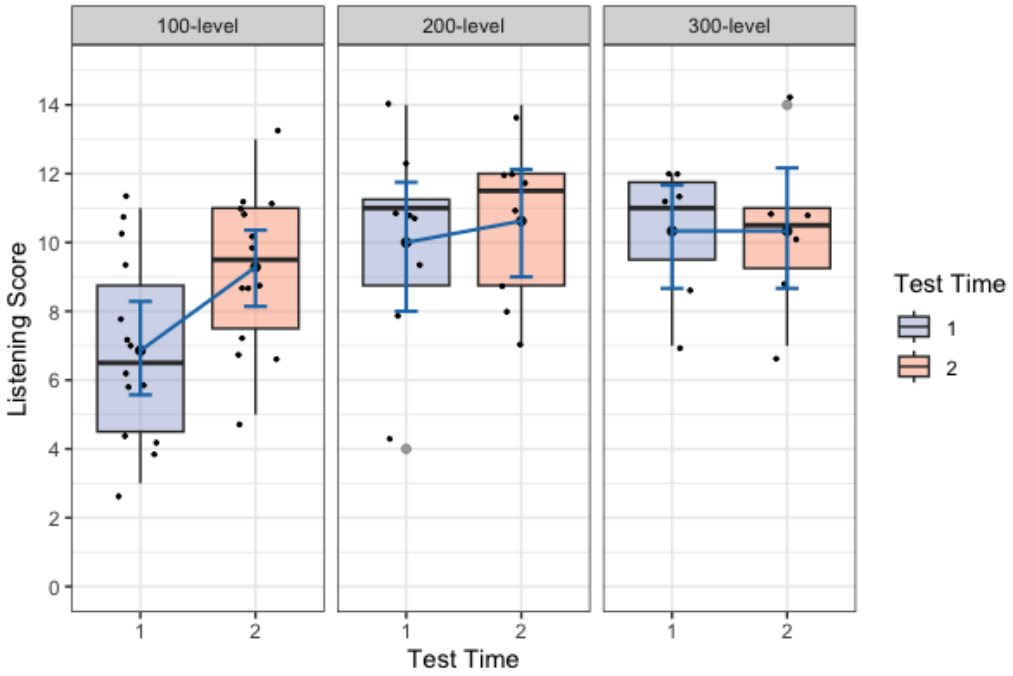
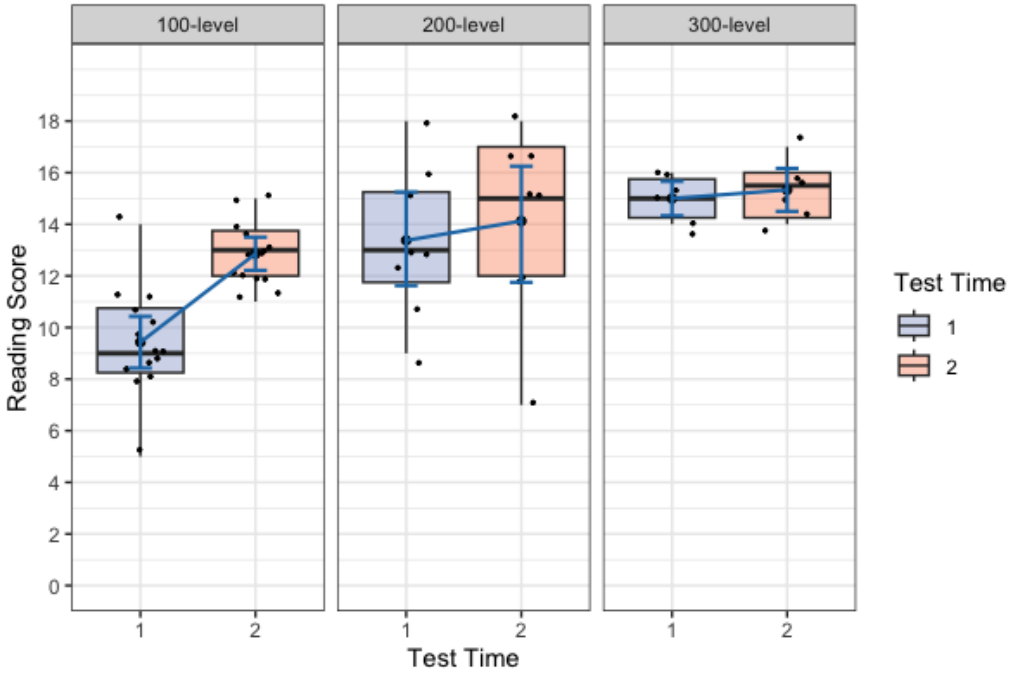


Figure 14. Boxplots reading test scores by course level and test time.



To determine if the observed total score changes across the semester and course levels were statistically significant, I conducted a repeated-measures 2 x 3 Analysis of Variance

(ANOVA). Students' total scores served as the dependent variable, time as the within-subject independent variable, and course level as the between-subject variable. I reported the Wald-type test statistics (WTS) and ANOVA-type statistics (ATS) calculated by the R package MANOVA.RM (Friedrich, Konietschke, & Pauly, 2019a) to address the small sample issue and the violation of the homogeneity of variance. These two statistics were determined using nonparametric methods that employ resampling techniques for approximating the sampling distribution, allowing for their application even in small sample sizes. These methods are suitable in the Behrens-Fisher situation, where equal covariance matrices across groups are not assumed (Friedrich, Konietschke, & Pauly, 2019b).

Table 12 summarizes the results of the ANOVA. The results show significant effects of course level and time, as well as a significant interaction between course level and time, on students' test scores. I, therefore, performed a series of post-hoc tests to identify where the significant score differences lie among the different course levels. I used the Bonferroni test to adjust for the Type I error rate. Table 13 presents the significant results of pairwise post-hoc comparisons between different time points and course levels, along with their corresponding Cohen's *d* effect sizes (please see Appendix 8 for the results of all comparisons). Here's a brief summary of the key findings:

1. For 100-level students, there was a significant increase in the total test scores from Time 1 to Time 2 ($t(25) = 5.9$, $p < .001$), with a large effect size (Cohen's $d = 1.67$).
2. At Time 1, both 200-level and 300-level students had significantly higher scores compared to 100-level students, with large effect sizes (200-level: $t(35.7) = 7.1$, $p = .005$, Cohen's $d = 1.68$; 300-level: $t(35.7) = 9$, $p < .001$, Cohen's $d = 2.13$).

3. At Time 2, both 200-level and 300-level students also had significantly higher scores compared to 100-level students, with large effect sizes (200-level: $t(35.7) = 8.5$, $p < .001$, Cohen's $d = 2.01$; 300-level: $t(35.7) = 9.4$, $p < .001$, Cohen's $d = 2.22$).
4. There were no significant differences between the other comparisons, as indicated by *p*-values of 1. However, it is worth noting that the effect sizes for these non-significant comparisons were generally small to medium, ranging from 0.08 to 0.62.

From these results, I can conclude that there was a significant improvement in scores for 100-level students from Time 1 to Time 2, with a large effect size. Additionally, 200-level and 300-level students consistently demonstrated higher scores compared to 100-level students at both time points, with large effect sizes. However, there were no significant differences in scores between 200-level and 300-level students or within these course levels over time, and the effect sizes for these comparisons were generally small to medium.

Table 12. Summary of the results of ANOVA (WTS and ATS reported)

Wald-type test statistics (WTS)	χ^2 value	*df*		*p*-value resampling
Course level	28.04	2		.002
Time	12.42	1		.004
Course level x Time	14.49	2		.01
ANOVA-type statistics (ATS)	*F* value	*df1*	*df2*	*p*-value
Course level	7.02	1.57	19.52	.008
Time	12.42	1	INF	<.001
Course level x Time	5.61	1.76	INF	.005

Table 13. Significant post-hoc pair-wise comparisons results (total scores)

Contrast	*t*-value	*SE*	*df*	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	5.9	0.92	25	<.001	1.67
Time 1 200-level - Time 1 100-level	7.1	1.77	35.7	.005	1.68
Time 2 200-level - Time 1 100-level	8.5	1.77	35.7	<.001	2.01
Time 1 300-level - Time 1 100-level	9	1.95	35.7	<.001	2.13
Time 2 300-level - Time 1 100-level	9.4	1.95	35.7	<.001	2.22

Next, I performed the repeated-measures multivariate analysis of variance (MANOVA) to investigate the effects of time and course level on students' listening and reading scores. A repeated-measures MANOVA was chosen for this analysis because it allows for the examination of the within-subject effects of time on multiple dependent variables (listening and reading scores) while accounting for the correlation between repeated measurements on the same student. Additionally, by including course level as a between-subjects factor, this method can assess the influence of different course levels on the listening and reading scores and identify any interactions between time and course level that may exist. I reported the Wald-type test statistics (WTS) and modified ANOVA-type statistics (MATS) calculated by the R package MANOVA.RM for similar reasons noted above.

Table 14 summarizes the results of the MANOVA. The results showed significant main effects of time and course level, indicating that students' scores improved from the beginning to the end of the semester and that higher-level students generally performed better. Additionally, a significant interaction effect between time and course level was found, suggesting varying degrees of improvement among students from different course levels.

To further explore these effects, a series of post-hoc tests were conducted to identify significant listening and reading score differences among various course levels. I used the

Bonferroni test to adjust for the Type I error rate. Table 15 presents the significant results of pairwise post-hoc comparisons for listening scores between different time points and course levels, along with their corresponding Cohen's *d* effect sizes (please see Appendix 9 for the results of all comparisons). The key findings can be summarized as follows:

1. A significant improvement in scores ($p = 0.03$) from Time 1 to Time 2 was observed for 100-level students, with a moderate effect size (Cohen's $d = 0.68$).
2. A significant difference in scores ($p = 0.02$) was found between Time 2 for 200-level students and Time 1 for 100-level students, with a large effect size (Cohen's $d = 0.83$).
3. The remaining pairwise comparisons in the table were not significant and showed small effect sizes, indicating that the differences between those specific groups and time points were not significantly meaningful.

As for the reading scores, the post-hoc results are presented in Table 15 (see Appendix 10 for the results of all comparisons). The main results are summarized below:

1. Improvement over time: 100-level students demonstrated a significant improvement in scores from Time 1 to Time 2 ($t(25) = 3.4$, $p < .001$), with a large effect size (Cohen's $d = 0.96$).
2. Comparisons between course levels at Time 1 and Time 2: Both 200-level and 300-level students consistently scored significantly higher than 100-level students across Time 1 and Time 2, with large effect sizes (Cohen's *d* ranging from 0.83 to 1.25).
3. The remaining comparisons in the table were not statistically significant ($p > 0.05$).

Table 14. Summary of the results of MANOVA (WTS and MATS reported)

Wald-type test statistics (WTS)	χ^2 value	*df*	*p*-value resampling
Course level	104	4	< .001
Time	13.58	2	.011
Course level x Time	20.86	4	.01
Modified ANOVA-type statistics (MATS)	*F* value	-	p-value resampling
Course level	93.16	-	<.001
Time	8.02	-	.008
Course level x Time	15.17	-	.004

Note: For the MATS, degrees of freedom is not reported since here inference is only based on resample (Friedrich et al. 2019b, p.391)

Table 15. Significant post-hoc, pair-wise comparisons results (listening scores)

Contrast	*t*-value	*SE*	*df*	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	2.4	0.7	25	.03	0.68
Time 2 200-level - Time 1 100-level	3.8	1.09	42	.02	0.83

Table 16. Significant post-hoc, pair-wise comparisons results (reading scores)

Contrast	*t*-value	*SE*	*df*	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	3.4	0.66	25	<.001	0.96
Time 1 200-level - Time 1 100-level	3.9	0.97	44.4	.003	0.83
Time 2 200-level - Time 1 100-level	4.7	0.97	44.4	<.001	1
Time 1 300-level - Time 1 100-level	5.6	1.07	44.4	<.001	1.19
Time 2 300-level - Time 1 100-level	5.9	1.07	44.4	<.001	1.25

In addressing the research question, which seeks to determine whether students' test performance varies according to the amount and quality of prior Chinese learning experience, the analysis of students' performance (total scores, listening scores, and reading scores) at the

beginning (Time 1) and end (Time 2) of the semester and across different course levels yielded several key findings.

The results provide evidence supporting the explanation inference of the validity argument. First, there was a significant improvement in test scores for 100-level students from Time 1 to Time 2, indicating that students' test performance improved as they gained more experience in learning Chinese. However, it is important to note that for certain course levels, no significant difference was observed between Time 1 and Time 2 performance. Second, at both Time 1 and Time 2, 200-level and 300-level students consistently scored significantly higher than 100-level students, with large effect sizes. This suggests that students with more advanced learning experience in Chinese demonstrated better test performance than those with less experience. Interestingly, there was no significant difference observed between 200-level and 300-level students, implying that their test performance was relatively similar. Overall, the findings support the notion that students' test performance varies according to the amount and quality of their prior Chinese learning experience. The results lend validity evidence to the explanation inference, as the observed differences in test performance can be attributed to the variation in students' prior learning experiences

[RQ 4b: Explanation inference]:

Do students' test scores support the internal structure of the intended construct?

Investigating the internal structure of a language placement test is critical for establishing the test's effectiveness and identifying areas for improvement. By analyzing the factor structure of the test items, I can determine if the test scores align with the intended construct, in this case, Chinese language proficiency. A strong alignment between the test items and the underlying construct offers evidence that the test is a valid measure of language proficiency (In'nami &

Koizumi, 2016). Conversely, if the internal structure reveals inconsistencies or an inadequate representation of the construct, it can highlight areas that require revision to better assess the intended language skills.

I conducted an exploratory factor analysis (EFA) to examine the factor structure characterizing the 32 items in the placement test. The purpose of the analysis was to examine whether the internal structure of the scores collected via the placement test is consistent with a theoretical view of language proficiency. As noted earlier, I binary-scored students' responses as 1 (correct) or 0 (incorrect) for the multiple-choice items and treated the item responses as categorical. The pattern of eigenvalues supported a two-factor model, as shown in the scree plot in Figure 15. The fit statistics further supported the two-factor solution (TLI = .89, root-mean-square error of approximation [RMSEA] = .05), which was found to be superior to the one-factor solution (TLI = .73, RMSEA = .07). After a close examination of Promax rotated factor loadings indicated that the second factor was almost entirely driven by seven relatively easy items (at least 85% of students answered these items correctly). Additionally, five of the items were related to the same prompt (reading items #7-11), indicating that they may be measuring similar language skills. I bundled the five items into a polytomous super-item and re-conducted the EFA. Collapsing these five items led to a substantially smaller second eigenvalue (from 2.46 to 1.86, see Figure 16) and to a better fit for both the one-factor (TLI = .87, RMSEA = .05) and the two-factor solution (TLI = .92, RMSEA = .04). Although the fit statistics suggest stronger support for the two-factor model, it is important to note that the second factor seems to be mainly driven by relatively easy items and may not necessarily represent a separate dimension of language proficiency. Overall, the results of the exploratory factor analysis suggest that the 32-item placement test does appear to measure language proficiency as intended.

Figure 15. Scree plot for exploratory factor analysis (all 32 items).

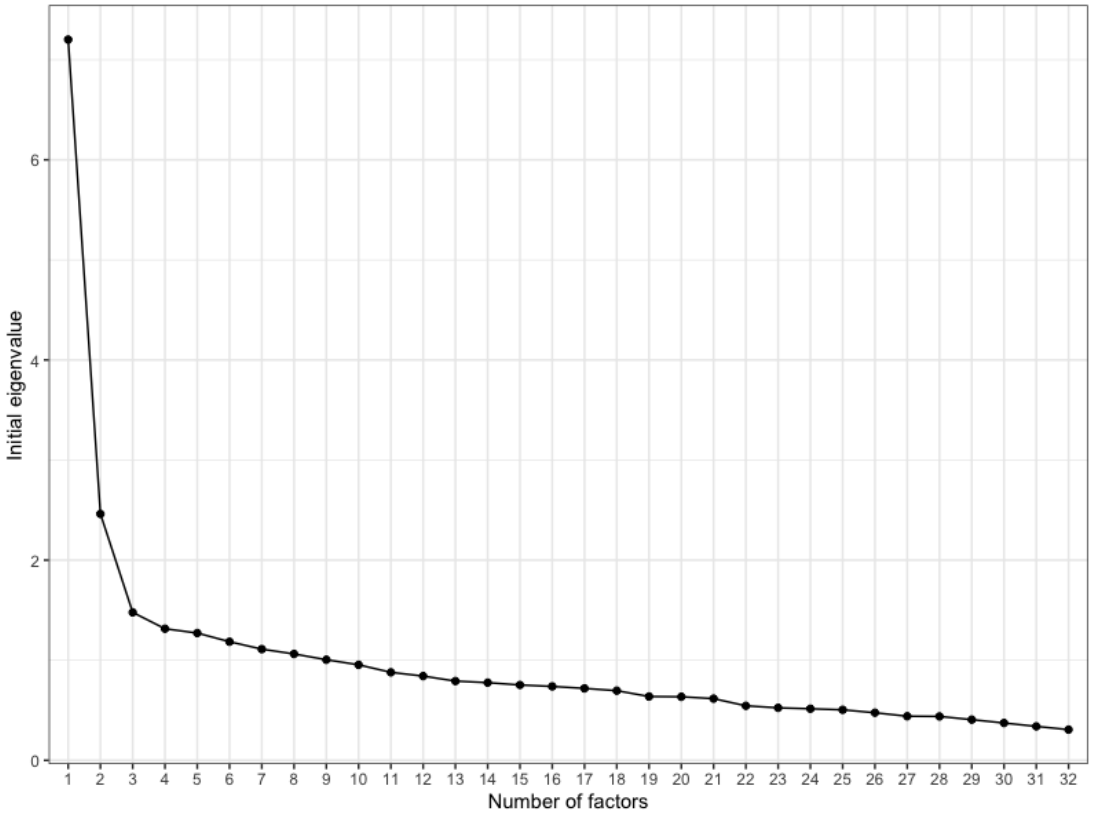
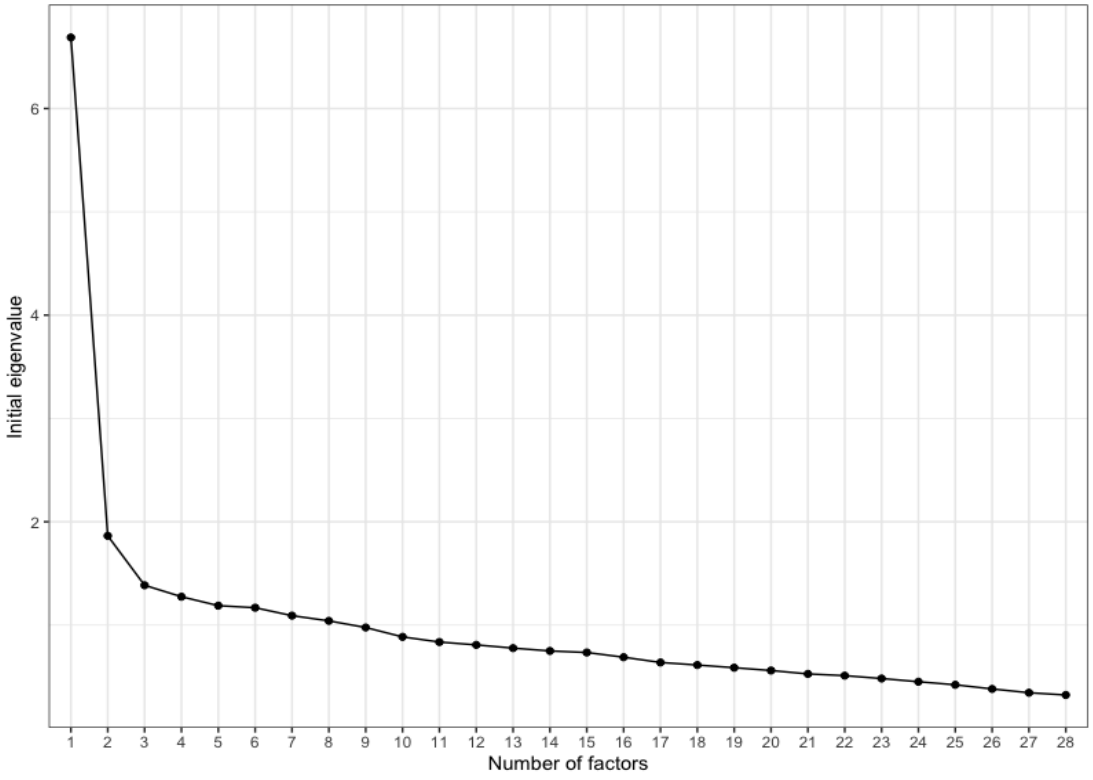


Figure 16. Scree plot for exploratory factor analysis (after collapsing the five items to the same prompt).



[RQ 5: Extrapolation inference]:

Do students' test scores support the relationship between their performance on the test and other indicators of Chinese language proficiency?

Establishing the extrapolation inference is a critical aspect of the validity argument, as it contributes to understanding the degree to which the Chinese placement test scores can be applied to other indicators of Chinese language proficiency. The investigation of the extrapolation inference seeks to demonstrate that the test scores not only reflect the students' performance on this particular test but also accurately represent their overall language proficiency.

Correlational analyses were employed to address the research question related to extrapolation inference. This analytical approach facilitates the assessment of the strength and direction of the relationship between two variables. In this study, these variables encompass the students' placement test scores and their performance on the ACTFL tests. By scrutinizing these relationships, the research seeks to examine the extent to which the placement test serves as a valid measure of Chinese language proficiency, consistent with other well-established indicators such as the ACTFL tests.

The rationale for expecting performances on similar skills (e.g., listening) to exhibit a strong correlation with one another, while performances on different skills (e.g., listening versus speaking) display a weaker correlation, lies in the assumption that proficiency in one language skill should be closely related to proficiency in a similar skill. Consequently, if the placement test accurately reflects Chinese language proficiency, it should demonstrate a stronger relationship with corresponding skills (listening and reading) on the ACTFL tests, while a weaker relationship should be observed with non-corresponding skills (e.g., speaking). In this case, the observed relationships can be considered supportive evidence for the validity of the Chinese placement test.

Table 17 displays the number of students with available ACTFL scores for each test, as well as the descriptive statistics for these ACTFL ratings and their corresponding placement test scores. It can be observed from the table that more students took the OPIc test than the RPT and LPT tests. This was expected as the speaking tests were conducted during class time with their teacher in attendance, making participation more convenient. In contrast, students had to visit the language lab independently to take the RPT and LPT tests, resulting in lower participation despite the incentive of extra credit (Winke & Ma, 2019).

Table 17. ACTFL scores and placement test results for students

	With available OPIc scores			With available LPT scores			With available RPT scores		
	N	M (SD)	95%CI	N	M (SD)	95%CI	N	M (SD)	95%CI
ACTFL Score	54	4.11 (1.53)	[3.69, 4.53]	46	2.85 (1.7)	[2.34, 3.35]	43	2.93(1.82)	[2.37,3.49]
Placement total		21.74 (5.84)	[20.15, 23.34]		22.41 (5.96)	[20.64, 24.18]		22.23(6.08)	[20.36,24.1]
Placement listening		8.96 (3.25)	[8.08, 9.85]		9.33 (3.24)	[8.36, 10.29]		9.19(3.28)	[8.18,10.19]
Placement reading		12.78 (3.42)	[11.84, 13.71]		13.09 (3.51)	[12.04, 14.13]		13.05(3.57)	[11.95,14.14]

Note: OPIc = the computerized oral proficiency test; LPT = Listening Proficiency Test; RPT = Reading Proficiency Test.

Figures 17-19 are the scatterplots displaying the relationship between students' ACTFL scores (listening, reading, and speaking) and their performance in the placement test, with separate plots for listening, reading, and total scores. The polyserial correlation coefficients indicate the strength of the relationship between the ACTFL speaking scores and each of the three placement test score components. Upon reviewing the figures and the polyserial correlation coefficients, the following observations can be made:

1. ACTFL Listening: The strongest correlation is observed between ACTFL listening scores and placement test listening scores (0.769), followed by total scores (0.783) and reading scores (0.623). This suggests that students with higher listening proficiency on the ACTFL test perform better on listening and overall components of the placement test, with a relatively weaker relationship observed for reading scores.
2. ACTFL Reading: The correlation coefficients for ACTFL reading scores reveal moderately strong relationships with placement test reading (0.706), total scores (0.756), and listening scores (0.637). This indicates that students with higher reading proficiency on the ACTFL test perform better on reading and overall components of the placement test, with a relatively weaker relationship observed for listening scores.
3. ACTFL Speaking: The correlation coefficients for ACTFL speaking scores show moderately strong relationships with placement test listening (0.63) and total scores (0.598), and a weaker relationship with reading scores (0.47). This suggests that students with higher speaking proficiency on the ACTFL test perform better on listening and overall components of the placement test, with the weakest relationship observed for reading scores.

The varying correlation coefficients between ACTFL scores and placement test scores may be attributed to the distinct nature of language skills being compared. Generally, stronger relationships are observed between corresponding skills (e.g., ACTFL listening vs. placement test listening), as these skills share many underlying linguistic competencies. In contrast, weaker relationships are observed as expected between non-corresponding skills (e.g., ACTFL speaking vs. placement test reading), which might be due to differences in the specific linguistic and cognitive processes involved in each skill.

These findings contribute to establishing the extrapolation inference by demonstrating that the Chinese placement test scores not only represent students' performance on the test itself but also show meaningful relationships with other indicators of Chinese language proficiency. This evidence supports the argument that placement test scores can be used to make inferences about students' broader language skills and proficiencies beyond the test context.

Figure 17. Scatterplots of ACTFL LPT scores and placement test scores.

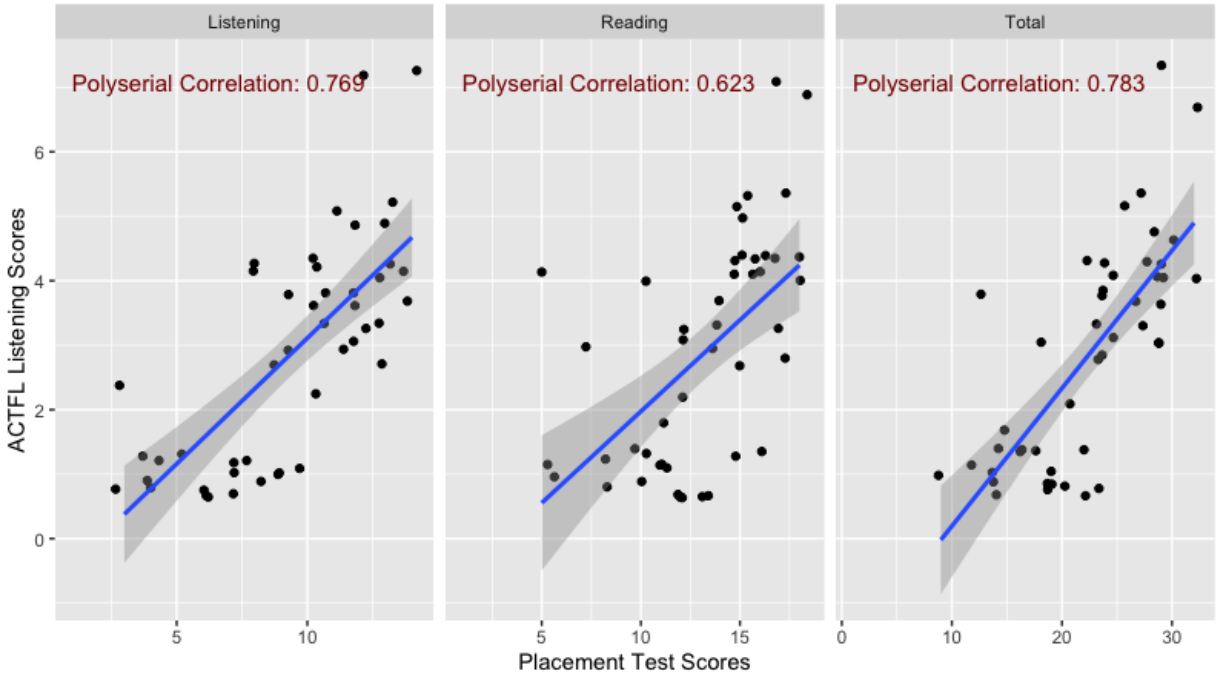


Figure 18. Scatterplots of ACTFL RPT scores and placement test scores.

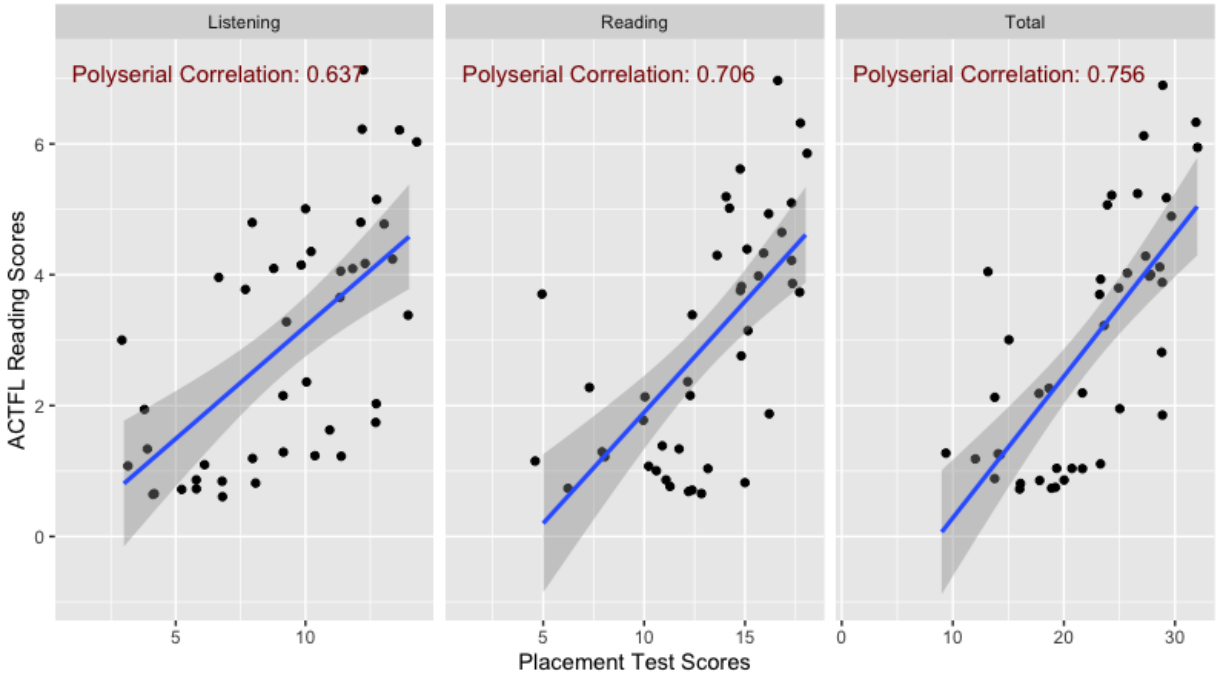
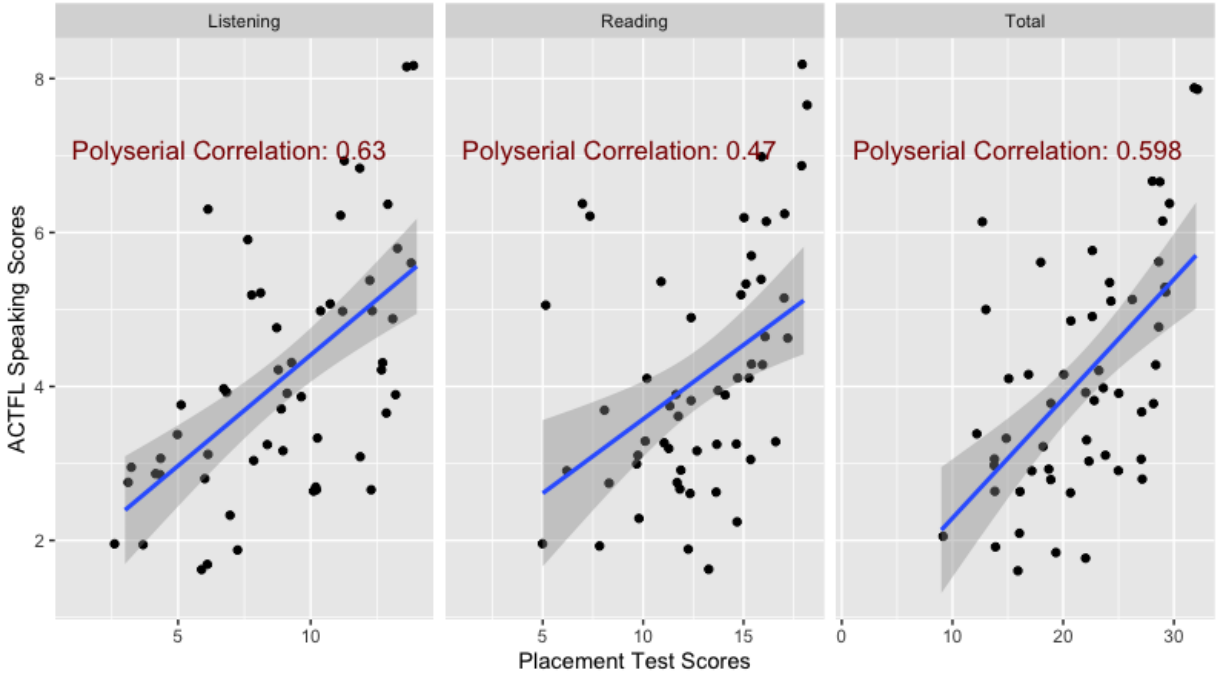


Figure 19. Scatterplots of ACTFL OPIc scores and placement test scores.



[RQ 6a: Utilization inference]:

From the perspective of course instructors, are students placed into appropriate course levels?

An important aspect of the validity argument is to examine the utilization inference, which provides insight into whether students are placed in course levels that align with the expectations of test stakeholders. This is crucial because appropriate placement ensures that students have the best learning experience and can achieve their full potential in Chinese language courses. By analyzing instructors' perspectives, I can gain valuable insights into the effectiveness of the placement test in reflecting the students' language proficiency levels.

To address this research question, both questionnaires and interviews were employed to collect data from three Chinese course instructors. The questionnaire consisted of several questions that required instructors to rate various aspects of the placement process on a scale of 6 (1 - never true; 2 - usually not true; 3 - rarely true; 4 - occasionally true; 5 - usually true; 6 - always true). The questions pertained to the accuracy of student placement and whether the course was too easy or too difficult for some students.

The results from the questionnaire are summarized in Table 18, which indicate that the instructors generally believed that students were accurately placed in the courses according to their prior Chinese knowledge and language proficiency, as they all responded with a rating of 5 for all the courses (CHS101/102, CHS201/202, and CHS301/302). However, when it came to the questions about the difficulty of the course for some students, there were mixed responses, with some instructors rating the course as being too easy or too difficult for certain students. These findings suggest that, overall, the Chinese placement test is effective in placing students into appropriate course levels, as reflected by the instructors' perspectives. However, there may still

be room for improvement in terms of better tailoring the course difficulty to the needs of individual students.

Table 18. Instructor ratings on student placement and course difficulty

Course Level	Question	Instructor A	Instructor B	Instructor C
CHS101/102	● Accurate placement given prior knowledge and language proficiency	5	5	5
	● Class too easy for some students	6	4	6
	● Class too difficult for some students	6	4	6
CHS201/202	● Accurate placement given prior knowledge and language proficiency	5	5	5
	● Class too easy for some students	6	4	6
	● Class too difficult for some students	6	4	6
CHS301/302	● Accurate placement given prior knowledge and language proficiency	5	5	5
	● Class too easy for some students	6	2	6
	● Class too difficult for some students	6	4	6

Note: The questions were rated using a Likert scale ranging from 1 to 6, where 1 = never true, 2 = usually not true, 3 = rarely true, 4 = occasionally true, 5 = usually true, and 6 = always true.

In the interviews, instructors further discussed the issue of mismatched course difficulty for some students. They explained that this problem is more prevalent in higher-level courses, primarily because of the diverse range of students' abilities. One instructor mentioned that students might reach higher-level courses without possessing the necessary skills, not solely due to the placement test, but also because of how students progress through the course levels:

> This issue [that the class may pose a challenge for certain students] is common, particularly in recent years, due to the diverse range of students' abilities. In the first grade, it's less of a problem as most students have a relatively low skill level. However, by the third grade, there's a wide range of abilities. In the fourth-grade classes I've seen, some students' skills are at the second or third-grade level, while others are at the fourth-grade level. It may seem surprising that students reach higher levels without having the necessary skills. This issue extends beyond the placement test. For instance, a student may have completed first, second, and third grades at our school, just barely passing with 60% marks, and then advanced to the next grade. Unfortunately, their language proficiency and actual understanding of the material remain quite unsatisfactory (instructor #3).

Another instructor provided similar comments, emphasizing that the challenges faced by certain students in Chinese classes are not necessarily due to errors in placement tests resulting in misplacement:

> “Often, the issue lies with the students themselves. For instance, a student may begin a course like 101, which should be relatively easy to master, but only achieve a 70% passing grade, indicating poor mastery. Despite this, MSU does not prevent students from advancing to 102 based solely on their 70% grade. This reveals the problem: students haven't mastered the content of 101, yet they are allowed to progress. Similarly, if they achieve only 60% or 65% in 102, even worse than their performance in 101, they can still advance to 201. Consequently, students' performance declines with each grade level, leaving the weakest students consistently lagging behind. These struggling students can be observed in courses 101, 102, 201, 202, and 301. Additionally, given the high number

of credits associated with Chinese language courses (101 to 202 are five credits each, and 301 and 302 are four credits each) and the expensive tuition fees, instructors may feel compassionate towards their students. They attempt to provide as much help as possible, such as awarding extra points to students who revise for each exam. Without these additional support measures, many students might be at risk of failing or barely passing (instructor #1)."

When asked about the effectiveness of the placement test, all instructors agreed that, in most cases, the test successfully placed students into appropriate levels, as evidenced by the following excerpts from the interview:

Researcher: Is there a difference in the language abilities of students who are placed into 201 through the placement test compared to those who progress from 101 and 102? Or is the placement test effective in identifying their levels, resulting in both groups of students being quite similar?

Instructor #3: In my opinion, both groups of students are quite similar. Interestingly, regardless of whether they have reached the 201 level at MSU or through other means before enrolling in the Chinese language courses at MSU, they tend to make the same mistakes. We need to constantly remind them of areas where they are prone to making errors, such as the use of "be" verbs in English and Chinese, which are actually different. Even in the fourth year, some students still make this mistake. Overall, I think the placement test generally has a good accuracy rate, particularly at the beginning and intermediate levels.

However, there were some cases where the placement test was considered less effective:

> "The placement test may not be as reliable for specific student groups, such as those who have learned through family connections or cultural exposure and therefore have not followed a traditional textbook-based curriculum. Heritage learners acquire language skills differently, which can lead to the placement test inaccurately reflecting their proficiency level. (instructor #2)"

When the placement test scores are not accurate enough, teachers may need to resort to in-person interviews as an additional means of assessment for the next step of evaluation. In other words, interviews can be used as an alternative method to evaluate the proficiency level of students if the placement test results are not sufficient.

Taken together, the findings offer support for the utilization inference of the validity argument, as most instructors agreed that the placement test effectively placed students into appropriate course levels. However, there are instances where the test may not have been as effective for specific student groups (heritage learners) or when considering the course difficulty for individual students. This point will be further explored and addressed in the discussion section.

[RQ 6b: Utilization inference]:

From the perspective of students, are they placed into appropriate course levels?

In addition to the insights provided by course instructors, the inclusion of students' voices allows for a more nuanced understanding of the placement process's effectiveness. Students may offer unique perspectives on the placement process, highlighting aspects that instructors may overlook. By sharing their experiences and providing feedback, students can offer valuable information on whether the course levels align with their language proficiency and learning needs. To address the research question, I analyzed students' interview responses and

questionnaire data. In the questionnaire, all students were asked about whether they took the Michigan State University Chinese placement test before enrolling in their first Chinese language course. However, only students who responded affirmatively were asked the following questions (see the questions in Appendix 2):

1. The advised Chinese course based on the test result.
2. Their GPA in the assigned course.
3. A rating of the student's preparedness for the course on a scale of 6 (1 - unsatisfactory; 2 - needs improvement; 3 - slightly below expectations; 4 - meets expectations; 5 - exceeds expectations; 6 - outstanding).
4. A rating of the student's overall course performance on a scale of 6 (1 - unsatisfactory; 2 - needs improvement; 3 - slightly below expectations; 4 - meets expectations; 5 - exceeds expectations; 6 - outstanding).

These questions are crucial for answering the research question as they provide insights into various aspects of students' experiences with the placement process and their subsequent performance in the assigned courses. The first question helps determine the alignment of students' placement test scores with the assigned course levels. The second question focuses on students' academic performance in the assigned courses, serving as an indicator of the appropriateness of the course levels. The third question captures students' perceptions of their preparedness for the course, revealing any potential gaps or mismatches between their prior knowledge and course expectations. Lastly, the fourth question allows students to evaluate their overall performance in the course, reflecting their engagement, effort, and success in the learning process. By analyzing the data collected from these four questions, the study can assess the

effectiveness of the placement test in assigning students to appropriate course levels, taking into account their language proficiency, preparedness, and academic performance.

Table 18 summarizes students' responses to the four questions, revealing several key findings. A majority of the students achieved high GPAs in their assigned courses, with 89% in the 100-level courses and 75% in the 200-level courses having a GPA of 3.5 or higher, indicating academic success and appropriate course placement. Additionally, students generally felt well-prepared for their courses, as shown by the high mean preparedness ratings across all course levels. Lastly, students also reported positive overall course performance, with high mean performance ratings in each course level.

Table 19. Summary of student placement outcomes and perceptions in Chinese language courses

	N	GPA - N (%)		Preparedness		Performance	
		3.5 +	3.0 +	Mean (SD)	Range	Mean (SD)	Range
100	9	8 (89%)	1 (11%)	4.89 (.93)	[4, 6]	5.44 (.88)	[4, 6]
200	8	6 (75%)	2 (25%)	4.25 (.89)	[3, 6]	54.62 (.92)	[3, 6]

Upon reviewing Table 19, it was evident that although most students felt well-prepared and performed well in their assigned courses, the range value of 3 in Table 18 in both preparedness and overall performance ratings indicated that some students rated these aspects as slightly below expectations. To gain a deeper understanding of the challenges faced by these students and to identify any potential limitations of the placement test, examining the interview data of a student who provided a lower rating was essential.

This particular student's experience highlights some issues that may not have been captured by the average ratings. Initially having learned traditional Chinese, she faced difficulties with the placement test due to the use of simplified Chinese, which led to stress and uncertainty

about her placement in Chinese 201. After attending a few classes, she found the workload and the transition to simplified Chinese challenging. This prompted her to consult the Chinese supervisor, who assessed her skills and recommended Chinese 102. However, since the class was only offered in the spring, she opted for Chinese 101, which she found more comfortable and better suited to her needs:

> "I first learned traditional Chinese... So coming in and taking the placement exam was a little bit difficult for me... I actually changed to Chinese 101 after attending two or three classes... I think it kind of like made me a little bit nervous because of like the amount of classwork there was and I still wasn't really comfortable with simplified Chinese... I met with the Chinese program supervisor... she's like, I think 102 would probably be the best, but because they only offer that in the spring. I decided to just like let's just do 101... I feel way more comfortable in Chinese 101 compared to 201." (student #1, CHS101)

This student's experience underlines the importance of considering individual learner backgrounds and the possible discrepancies between traditional and simplified Chinese when evaluating the placement test's effectiveness. It also emphasizes the value of communication and collaboration between students and program supervisors to ensure appropriate course placement, especially when students encounter challenges that the placement test may not fully capture.

In summary, the study results indicate that the Chinese placement test generally assigns students to suitable course levels, as evidenced by the high GPAs, preparedness ratings, and performance ratings. However, some students may find their placement below expectations. A student with a traditional Chinese background encountered difficulties due to the test's focus on simplified Chinese and its inability to fully capture individual learning needs. Although she found a more appropriate course after consulting the Chinese supervisor, her experience

highlights the need to address these limitations. In summary, from the students' perspective, the Chinese placement test effectively places them in appropriate course levels, but addressing potential limitations and considering individual learner backgrounds can further enhance the test's accuracy and its ability to meet students' needs.

[RQ 6c: Utilization inference]:

Are cut-off scores set appropriately?

The appropriateness of test cut-off scores plays a critical role in valid score utilization and interpretation, as it directly impacts the precision and effectiveness of the exam in assigning students to suitable course levels. To answer the research question, as noted earlier, I collected teacher ratings on their perceptions of the relevance of the items targeting each level of the course. Specifically, instructors assessed all 32 items in the placement test concerning their relevance and appropriateness to the course content of 100-level, 200-level, and 300-level courses. A series of checkboxes were provided for each item to enable instructors to make their assessments. For ease of analysis and interpretation, when instructors' ratings differed regarding the course an item was targeting, the course with the most instructors selecting it was determined as the item's target level.

Upon examining the results, it was found that for 7 out of 32 items, instructors' perceptions of item relevance and appropriateness to course content differed by one-course level. However, none of the items had instructor ratings that differed by two levels, indicating that the items were generally well-aligned with the intended course levels, albeit with some variation. This consistency in instructors' perceptions of item relevance is an important factor to consider when evaluating the appropriateness of the cut-off scores. For a summary of the number of items perceived as relevant to each course level, please see Table 20.

Table 20. Summary of items perceived by instructors as relevant to each course level

Course level	N of items	Cut-offs for placement decisions
100-level	20	< 20
200-level	9	< 30
300-level	3	>= 30

Table 20 shows that of 32 items in the test, 20 items were perceived as relevant to the 100-level course, 9 items to the 200-level course, and 3 items to the 300-level course. These numbers correspond to the placement cut-off scores of less than 20 for 100-level courses, less than 30 for 200-level courses, and 30 or greater for 300-level courses. Although the number of items matched to each course level seems to correspond with the cut-off scores, the distribution of items across the levels appears to be imbalanced. The 100-level courses have significantly more items (20) than the 200-level (9) and 300-level (3) courses. This imbalance may lead to a less accurate measurement of students' language proficiency in the upper-level courses, as there are fewer items to gauge their abilities. These findings suggest that the Chinese placement test might benefit from a more balanced distribution of items across the various course levels to better assess students' language proficiency at each level. By adjusting the number of items targeting the 200-level and 300-level courses and ensuring a more even distribution across all levels, the test's accuracy in placing students in appropriate course levels can be improved.

In summary, the analysis of the teacher ratings and the distribution of items across course levels provides mixed evidence regarding the appropriateness of the cut-off scores for the Chinese placement test. On one hand, the fact that the instructor ratings for item relevance did not differ by more than one-course level for any item suggests that the items are generally well-aligned with the intended course levels. This consistency in instructors' perceptions of item

relevance supports the validity evidence of the cut-off scores. On the other hand, the uneven distribution of items across the course levels, with the 100-level courses having significantly more items than the 200-level and 300-level courses, raises concerns about the accuracy of the test in measuring students' language proficiency in the upper-level courses. This finding suggests that the test might benefit from a more balanced distribution of items across the various course levels to better assess students' language proficiency at each level.

Therefore, while the evidence does not outright reject the appropriateness of the cut-off scores, it does indicate that improvements can be made to enhance the test's accuracy in placing students in appropriate course levels. By adjusting the number of items targeting the 200-level and 300-level courses and ensuring a more even distribution across all levels, the validity evidence for the utilization inference of the Chinese placement test can be further strengthened.

[RQ 7: Consequence implication inference]:

Does the test have positive effects on Chinese teaching and learning?

The consequence implication inference of the Chinese placement test delves into the real-world effects of the test on teaching and learning, specifically investigating its impact on Chinese language instruction and student learning experiences. Gaining insights into the implications of the Chinese placement test is crucial for understanding its validity, effectiveness in promoting learning experiences, and areas for potential enhancement. This knowledge is valuable for educators, administrators, and other stakeholders, enabling them to make well-informed decisions about adopting and implementing the test to fulfill the needs of both instructors and students.

To thoroughly address the research question and gain a comprehensive understanding of the Chinese placement test's consequence implication inference, I employed qualitative analysis

of the feedback from students and instructors collected from the interviews. The analysis enables a deep exploration of students' and instructors' subjective experiences and perceptions regarding the test's impact, capturing the intricacies of how the test influences teaching and learning processes. The collected data from student interviews suggest that the Chinese placement test has positive effects on Chinese teaching and learning. Students generally reported that the placement test results were helpful in guiding their decisions regarding course selection. For example, one student stated,

> "I think [the placement test results] were pretty accurate. (Student #2, CHS202)."

However, it should be noted that some students viewed the test results as a suggestion rather than a strict guideline, using it as a basis to make their own decisions about their level of comfort and willingness to engage with the course material. This student further elaborated on this point, saying,

> "I know some people who were placed higher, but they chose to go back a step or start over completely. I don't know too many people who pushed ahead, which is interesting. (Student #2, CHS202)."

Another student, who was initially placed in Chinese 201, decided to take Chinese 101 instead due to concerns about the time commitment involved in a five-credit course during her first semester. She explained,

> "I feel like I could do it (take 201) if I put in the time for it, but I didn't really want to do it for my first semester. I just felt with it being a five-credit course, I didn't want it to take up the majority of my time compared to my other classes. (Student #7, CHS102)."

Similarly, a student who was advised to take Chinese 201 based on the test results opted for Chinese 101 because of concerns that his Chinese language skills had become rusty. Upon reflection, the student admitted,

> "I think the placement test results were more accurate in describing what I would be able to do, given that Chinese 101 was very easy. I definitely would have taken 201 if I had gone back in time and said to myself, look, this is what it's gonna be. (Student #3, CHS102)."

In light of these individual decisions, another student drew attention to the fact that not all students in a given Chinese class were equally proficient. This observation underscores that factors other than the placement test, such as prior language learning experience, could contribute to differences in proficiency levels. The student observed,

> "I don't think that's necessarily the case [that all students were equally proficient in my class]. I believe there is a significant role played by how much Chinese has been taken before high school or before college. Some students (who are placed in) had 4 or 5 years of prior Chinese learning experience and are noticeably more proficient. (Student #4, CHS302)."

This point further emphasizes the importance of considering individual factors and preferences when making course placement decisions. Building on the insights gained from student interviews, the analysis of instructors' interview data further explores the impact of the Chinese placement test on teaching and learning. Instructors provided valuable perspectives on the positive effects of the test, as well as the challenges they face in accommodating students with varying levels of proficiency. One instructor mentioned the advantages of the placement test, stating,

"Students who enter our class through the placement test are generally easier to manage, as they are usually placed at the appropriate level. (instructor #1)"

This demonstrates the positive effects of the placement test in accurately assessing students' proficiency and ensuring they are enrolled in suitable courses. However, the instructor also noted that there is a range of proficiency levels among students who have advanced from the 101 and 102 classes, rather than being placed by the test. This variation in proficiency creates a stratified learning environment, where instructors must tailor course difficulty to accommodate the majority of students and follow their learning pace. For high-achieving students, the instructor encourages them to take on extra work, such as writing more in-depth essays or improving their presentations. On the other hand, supporting struggling students who have advanced from lower-level courses can be quite challenging for teachers. As one instructor noted,

"We used to have a teaching assistant (TA) or a Chinese language helper in our department. However, due to the pandemic, we haven't had such support for the past couple of years. Foreign language teaching assistants (FLTAs) can provide some help, but it is often insufficient. (instructor #3)"

The instructor also highlighted the difficulties in dealing with students who took a break from their studies and experienced a decline in their language proficiency. She shared,

"For example, I have encountered one or two students who took a year or two off, forgot what they learned, but still earned credits. This creates a difficult situation. Despite their enthusiasm, their proficiency has dropped to the 200 level, but they have already earned 300-level credits and need to graduate. (instructor #3)"

As a result, these students enroll in 400-level courses, posing a challenge for both the instructor and the students themselves. To mitigate this issue, the instructor allows these students to audit lower-level classes if their schedule permits, offering them additional support to catch up with their peers.

In conclusion, the data gathered from student interviews demonstrate that the Chinese placement test has positive effects on both Chinese language instruction and student learning experiences. Students reported that the test results were accurate and helpful for course selection, while instructors found it easier to manage students placed at appropriate levels. However, it is essential to consider individual factors and preferences when making course placement decisions to ensure the best possible learning outcomes for all students.

CHAPTER 5: DISCUSSION

In the current study, I aimed to provide a comprehensive examination and evaluation of the test score uses and interpretations for the listening and reading sections of an in-house, college-level Chinese placement test. Filling a gap in the literature on foreign language placement testing, the study focused on a language other than English and addressed methodological limitations commonly found in existing research. Using an argument-based validation framework conceptualized by Kane (2006) and expanded by Chapelle et al. (2008), for the study I collected and evaluated quantitative and qualitative validity evidence across seven inferences: domain description, evaluation, generalization, explanation, extrapolation, utilization, and consequence implication.

The primary goals of the study were to (1) investigate the functioning of test items by identifying and revising psychometrically problematic items, if any; (2) utilize the empirical results to inform test revisions, if needed; (3) demonstrate how the collected quantitative and qualitative results serve as strong or weak evidence or counter-evidence for the validity argument; and (4) provide an overall evaluation of the intended interpretation and use of the placement test scores. By employing mixed-methods, the study aimed to contribute to the larger discussion of foreign language assessment practices and argument-based test validation, while also offering insight into the ongoing development of validity research.

In this discussion section, I will summarize and evaluate the validity evidence for each research question using the criteria determined earlier. I will then discuss the results in relation to previous SLA literature when applicable. Then, the chapter closes with a brief discussion of some of the limitations of the current study.

[RQ 1: Domain description inference]:

Are the relevance of the test items and test criteria to the instructional domain and the appropriateness of the item difficulties supported by test stakeholders?

The results of the first research question, concerning the domain description inference, were evaluated by examining the instructors' and students' perceptions of the test items' relevance, appropriateness, and difficulty. Based on the results, strong evidence supports the relevance and appropriateness of the test items and criteria to the instructional domain. None of the items were considered irrelevant to the course material across all three levels of instruction by the instructors. In addition, students in all course levels generally perceived the test items as relevant to their course, as evidenced by their relatively high mean relevance scores. Furthermore, the appropriateness of the item difficulties is supported by the findings that students in higher-level courses found test items easier compared to those in lower-level courses. This trend is expected, as students in advanced courses should have a higher proficiency level in the language, allowing them to find the items less challenging.

The results align with the existing literature that emphasizes the importance of test content relevance and appropriateness for ensuring the validity of language assessments (Xi, 2010; Chalhoub-Deville & Deville, 2018). The alignment between the test items and the instructional domain contributes to the test's ability to accurately measure students' language proficiency and place them in suitable course levels.

[RQ 2a: Evaluation inference]:

Do test items yield item difficulty estimates that are appropriate for making placement decisions?

The research question focused on the evaluation inference, investigating whether the test items produced item difficulty estimates suitable for making placement decisions. To address

this question, two distinct methods were employed: a Rasch-based approach using a Wright map and correlation analysis with Pearson's correlation coefficients comparing students' and teachers' perceived difficulties with empirical item difficulties.

The Wright map analysis indicated that 75% of the examinees fall within the overlapping range of item difficulties and examinees' ability measures, suggesting reasonable item targeting along their ability level. However, a noticeable ceiling effect was observed, with approximately one-fourth of the examinees having ability measures above all item difficulties. This finding underscores the necessity of including more challenging items to address the ceiling effect for high-ability examinees, allowing for a more accurate assessment and facilitating appropriate placement decisions.

The correlation analysis revealed varying degrees of agreement between students' and teachers' perceived difficulties and empirical item difficulties. The overall correlation coefficient of .781 for the aggregated data, as well as data from 100- and 300-level students, demonstrates a robust relationship between students' perceived item difficulties and empirical item difficulties. This suggests that the test items generally function as intended and accurately measure the targeted construct.

However, the findings also uncovered some discrepancies and weaker relationships, particularly for the 200-level courses, where the correlation between students' perceptions of item difficulties and empirical item difficulties was lower (r = .415). One possible explanation for this observation is the increased heterogeneity in students' abilities and prior exposure to the content. The 200-level courses may consist of a more diverse group of students in terms of their abilities and prior exposure to the content, leading to greater variability in students' perceptions of item difficulties and a lower correlation with empirical item difficulties. This explanation

aligns with Ma and Winke's (2019) study, which found that intermediate students' self-assessed skills tend to be less accurate compared to those of beginner or advanced students.

Another observation from the results is the moderate correlation between teachers' perceptions of item difficulties and the empirical difficulties computed from students' performance (r = .507). Existing literature has revealed that there may be a mismatch between teachers' perception of item difficulties and students' actual performance on these items (e.g., (van de Watering & van der Rijt, 2006). This observation could be attributed to several factors. For instance, teachers might not be fully aware of the specific strategies students use when taking the test. Interviews with students and teachers revealed that students sometimes employed test-taking strategies when responding to items (e.g., listening for keywords instead of trying to understand every sentence in the listening items), whereas teachers often evaluated item difficulties based on the inclusion of difficult vocabulary or phrases. Another possible factor that may contribute to the moderate correlation between teachers' perceptions and empirical item difficulties is teachers' cognitive biases, such as overestimating the difficulty of items they themselves find challenging or underestimating the difficulty of items they consider easy (van de Watering & van der Rijt, 2006). Providing teachers with more insights into students' test-taking strategies and refining their understanding of item difficulty can help improve the alignment between teachers' perceptions and empirical item difficulties, ultimately leading to better test development and more accurate placement decisions.

[RQ 2b: Evaluation inference]:

Do test items exhibit no evidence of item bias?

The research question centered on determining whether the test items exhibit no evidence of item bias, specifically in terms of invariance across gender. Ensuring item invariance between

female and male examinees is vital for maintaining a fair and unbiased assessment (Kunnan, 2000). To address this question, the study analyzed the group invariance of item measures by examining differential item functioning (DIF) between female and male examinees. Based on the DIF analysis, the results revealed no evidence of DIF across the two examinee subgroups, as none of the items met both DIF criteria (i.e., statistical significance of the Mantel-Haenszel test at the .05 level after the Benjamini-Hochberg adjustment and a difference in item difficulty of at least .5 logit). This finding contributes to the strong validity evidence for evaluation inference and highlights the test's capacity to provide unbiased measures of language ability for both genders.

[RQ 2c: Evaluation inference]:

Are correct options unambiguous and accurately keyed?

For this research question, the investigation centered around the clarity and accuracy of the correct options in the test items. The aim was to determine whether the correct options were unambiguous and accurately keyed, as well-crafted multiple-choice items should include effective distractors that challenge examinees and require them to demonstrate their language abilities to select the correct response among plausible alternatives. To address this question, an analysis of distractors was conducted as an item quality indicator, aiming to assess the extent to which distractors for each item discriminated between examinees with different ability levels.

The analysis revealed that the keyed options generally attracted higher-ability examinees compared to the distractors. However, four items exhibited lower discriminating power, with mean ability estimate differences between the two groups of examinees of less than 1. Further investigation of these items revealed issues such as ambiguous phrasing in the item prompts, poorly selected distractors, and incongruent information in the prompt compared to the intended

answer (Downing & Haladyna, 2006). These findings indicate the need for a more detailed review of these items to ensure that the keyed options are clear and unambiguous, and the distractors are not too close in plausibility to the keyed option, which could lead to inaccurate measurement of examinees' abilities. Considering the criteria provided, the findings show weak evidence for the evaluation inference, as approximately 90% of the items are shown to be unambiguous and accurately keyed. This suggests that while the test generally provides accurate information on examinees' abilities, there is room for improvement, particularly in addressing the issues found in the four problematic items.

The importance of test revisions in test development cannot be overstated, as it is crucial to ensure that test items are reliable and valid measures of examinees' language abilities (Downing & Haladyna, 2006). Additionally, content experts can provide valuable feedback on items, identifying potential issues and suggesting improvements (Haladyna, Downing, & Rodriguez, 2002). Several guidelines can be followed to develop good items, such as ensuring that items are clear and concise, avoiding misleading or ambiguous language, and selecting distractors that are plausible but clearly incorrect (Haladyna et al., 2002). Using item-analysis and the functioning of distracters is an effective approach to examining the effectiveness of distractors and items (Wolfe & Smith, 2007; Osterlind, 1998). By understanding how examinees with different abilities respond to various distractors, test developers can make necessary revisions to ensure that the test items accurately assess language proficiency.

[RQ 3a: Generalization inference]:

Does the MSU Chinese placement test produce scores that are internally consistent?

The research question aimed to determine if the MSU Chinese placement test produces scores with internal consistency, which is critical for assessing the reliable measurement of

Chinese language proficiency across various contexts and student populations. High internal consistency allows for greater trust in the stability and precision of test scores. To analyze the internal consistency of the MSU Chinese placement test, Cronbach's α was computed, resulting in a value of 0.88 (95% CI: [0.86, 0.90]), demonstrating strong internal consistency.

Furthermore, an item analysis was carried out to evaluate the influence of each item on the overall internal consistency (refer to Table 8). The rationale behind examining Cronbach's α after removing each item is to identify any problematic items that could potentially lower the internal consistency of the test. Some factors that may contribute to a noticeable decrease in Cronbach's α include poor item quality, item difficulty (an item is significantly more difficult or easier compared to the rest of the test items), lack of content coverage (measuring a different aspect of the construct), item redundancy, and low item discrimination.

In this study, the analysis revealed that the removal of specific items would cause a minor decrease in Cronbach's α from 0.88 to 0.87, while for others, the α value would remain stable at 0.88. These outcomes indicate that the test items are consistently measuring the same underlying construct, and no single item significantly impacts the overall internal consistency. This item-level examination provides valuable information for test developers, enabling them to refine and improve the test's quality by identifying and addressing any problematic items. In conclusion, the findings offer robust evidence in support of the MSU Chinese placement test scores' internal consistency, thus reinforcing the generalizability inference in the context of language assessment.

[RQ 3b: Generalization inference]:

Are there adequate items to reliably differentiate students' abilities into three levels as intended?

In addressing the research question, the findings reveal that the MSU Chinese placement test exhibits a person reliability of .84 and a person separation index of 2.32. Based on the provided criteria, these results present weak evidence for the test's ability to reliably differentiate students' abilities into three levels. Despite not reaching the strong evidence threshold, the test still demonstrates satisfactory performance in assigning students to appropriate proficiency levels, as intended by the test developers.

Person reliability and person separation index have been commonly used in educational and psychological research to examine the psychometric properties of measurement instruments (e.g., Fan et al., 2021; Hu et al., 2022; Jefferies et al., 2021). For instance, Jefferies et al. (2021) applied the Rasch model to explore the psychometric properties of PLAYself, a tool designed for self-description of physical literacy in children and youth. They reported person reliability values ranging from .7 to .82, which indicates that PLAYself has good internal consistency and can reliably distinguish between different levels of physical literacy. In the study of Fan et al., the researchers examined the psychometric properties of the Norwegian Self-Efficacy for Therapeutic Use of Self questionnaire using Rasch analysis, excellent item and person separation were observed across all three parts (N-SETMU, N-SERIC, and N-SEMIE). The person separation index ranged from 2.8 to 4.6, indicating the successful differentiation of subjects into three to five distinct levels of self-efficacy.

Although not as commonly reported in SLA research, the use of person reliability and person separation index is common in educational and psychological research. The results highlight the utility of these indices in examining the validity evidence for placement tests, as the purpose of these tests matches well with the objectives of using these two measures: to effectively distinguish between different levels of language proficiency.

[RQ 4a: Explanation inference]:

Does students' test performance vary according to the amount and quality of prior Chinese learning experience?

In addressing the research question regarding whether students' test performance varies according to the amount and quality of prior Chinese learning experience, the study found significant improvements in listening, reading, and total scores on the Chinese placement test from the beginning to the end of the semester. The level of improvement varied across different course levels, with the most significant gains observed among the 100-level students, followed by the 200-level students. The least change in scores occurred for the 300-level students. The study's findings provide strong evidence for the validity of the Chinese placement test and contribute to the understanding of the relationship between prior learning experience and language test performance.

However, the observed differences in score improvements might be influenced by factors such as the ceiling effect and practice effect. The ceiling effect could be a potential explanation for the smaller improvements among higher-level students, as they already performed well in the first test administration, leaving less room for improvement. This observation suggests that adding more challenging items to the test might better differentiate the proficiency levels of higher-level students. However, this explanation remains speculative, and further research is needed to confirm the presence of the ceiling effect and its impact on the results. The practice effect, resulting from students taking the same test twice, could potentially inflate the observed improvements in test performance (Calamia et al., 2013). If the practice effect is substantial, it might lead to an overestimation of the actual gains in language proficiency. Although the current study cannot definitively establish the extent to which the practice effect impacted the findings, it

is essential to consider this potential limitation when interpreting the results and evaluating the validity argument.

It is worth noting that the sample size for the 300-level students in the study was small (n = 6). Small sample sizes can lead to low statistical power and increase the likelihood of Type II errors (Cohen, 1992). In the context of this study, the small sample size for the 300-level students may limit the ability to draw robust conclusions about the performance of this group and may not accurately represent the broader population of 300-level students.

[RQ 4b: Explanation inference]:

Do students' test scores support the internal structure of the intended construct?

The research question is whether students' test scores provide support for the internal structure of the intended construct. To answer this question, an exploratory factor analysis (EFA) was conducted to examine the factor structure of the 32 items in the placement test. The purpose of the analysis was to determine if the scores collected via the placement test reflect a theoretical view of language proficiency. If the test items align strongly with the underlying construct, this provides evidence that the test is a valid measure of language proficiency (In'nami & Koizumi, 2016). Conversely, inconsistencies or an inadequate representation of the construct in the internal structure can highlight areas that need improvement to better assess the intended language skills.

In the initial EFA, it was found that the second factor was mainly driven by relatively easy items, with five of these items being related to the same prompt. This finding prompted further investigation into the structure of the test items to better understand their impact on the construct validity. The five items were collapsed into a polytomous super-item, and the EFA was re-conducted. After this adjustment, both the one-factor and two-factor solutions showed improved fit. For the adjusted models, the fit statistics suggest stronger support for the two-factor

solution. However, it is important to note that the second factor still appears to be mainly driven by relatively easy items and may not necessarily represent a separate dimension of language proficiency.

These findings suggest that the placement test does appear to measure language proficiency as intended, but highlights areas that require improvement. One approach to address the issue is to consider removing the relatively easy items from the test. This solution could potentially increase the test's ability to differentiate between proficiency levels and reduce the influence of the identified issues on test validity. However, removing these items might also result in a loss of content coverage and may not fully address the underlying construct representation. Another approach is to revise the identified items or add more items to better represent the intended construct. This could involve introducing more challenging items or diversifying the prompts, which may help mitigate the potential impact of the relatively easy items and the clustering of items related to the same prompt on the internal structure of test items. When implementing this approach, it is important to ensure that the revised or added items align with the theoretical view of language proficiency and maintain content coverage (In'nami & Koizumi, 2016).

[RQ 5: Extrapolation inference]:

Do students' test scores support the relationship between their performance on the test and other indicators of Chinese language proficiency?

The extrapolation inference plays a pivotal role in the validity argument by determining if Chinese placement test scores can be effectively generalized to other indicators of Chinese language proficiency. This study explored the extrapolation inference by analyzing the relationship between students' placement test scores and their performance on the ACTFL tests.

In this study, strong evidence was found for the relationship between students' scores on the Chinese placement test and on ACTFL proficiency tests for corresponding skills (e.g., listening and listening). The correlation coefficients for corresponding skills were generally moderate to strong ($r \geq 0.40$), which supports the extrapolation inference. Furthermore, positive correlations between students' scores on the Chinese placement test and on ACTFL proficiency tests for non-corresponding skills (e.g., speaking and listening) were positive but weaker compared to the correlation between corresponding skills, providing additional evidence for the extrapolation inference.

The findings of this study not only contribute to the body of knowledge on validity evidence for the extrapolation inference, but also highlight the importance of considering correlations between corresponding and non-corresponding skills in language assessment research. In this context, a study by Eda, Itomitsu, and Noda (2008) serves as a valuable example. They investigated the validity evidence for JSKIT, a Japanese skills test, used as a placement tool in a summer intensive language program. The researchers examined the correlations between the subcomponents of the JSKIT and the corresponding subcomponents of the in-house placement test. Their findings indicated that the structure section of the JSKIT was most strongly correlated with the corresponding grammar section of the placement test ($r = .806$), and the reading section of the JSKIT was most strongly correlated with the corresponding reading section of the placement test ($r = .766$).

Comparing correlations between corresponding and non-corresponding skills is an essential aspect of examining validity evidence for language assessments. Although this crucial aspect has been addressed in other fields such as psychological and educational measurement research, it has often been overlooked in foreign language placement testing. Many studies in

this area have primarily focused on establishing positive correlations for corresponding skills, without considering the comparison of these correlations with those of non-corresponding skills.

[RQ 6a: Utilization inference]:

From the perspective of course instructors, are students placed into appropriate course levels?

To answer the research question related to utilization inference, data was collected from three Chinese course instructors using questionnaires and interviews. The findings offer strong evidence in support of the utilization inference, as instructors generally agreed that the placement test effectively placed students into appropriate course levels, aligning with the expectations of test stakeholders.

However, the current study also revealed challenges and difficulties in using placement tests for placing heritage learners, who typically have unique language learning experiences due to exposure to the target language through family or cultural experiences (Li & Duff, 2008). As a result, traditional placement tests may not adequately capture the abilities of heritage learners, leading to potential inaccuracies in their placement and, ultimately, an inappropriate course level that does not align with their needs. This issue is particularly relevant for heritage learners, as they often possess a strong foundation in oral and listening skills but may struggle with more formal aspects of the language, such as grammar and writing (Campbell, 2000; Kondo-Brown, 2005). Consequently, placement tests that heavily weigh formal language skills may not provide an accurate representation of heritage learners' true proficiency levels. Recognizing this limitation, teachers may need to resort to alternative assessment methods when the placement test scores are insufficient.

Oral interviews and background questionnaires, for instance, are the most widely used alternatives due to the lack of standardized placement tests specifically designed for heritage

learners (Li & Duff, 2008). These methods can offer valuable insights into a learner's proficiency, as the amount of schooling received in the target language is considered the most reliable indicator of heritage language proficiency. Nevertheless, individual differences still may exist even among students from the same class, which presents a significant challenge in placing Chinese heritage language learners from diverse educational systems into appropriate classes.

One reason why heritage learners may struggle with placement tests is their often uneven grasp of the heritage language. As noted earlier, some learners possess strong receptive or conversational skills while lacking in literacy, grammar, and vocabulary (Sohn, 2004). Additionally, their sociolinguistic and pragmatic competence may be limited, further complicating the placement process. To address these issues, some college or university Chinese language programs have adopted separate tracks for language learners, including heritage tracks for students who have had previous exposure to the target language, regardless of their ethnicity (Li & Duff, 2008). However, not all Chinese programs can afford to implement these dual tracks due to low enrollment, especially in recent years. This limitation may result in mixed classrooms with varying degrees of proficiency among heritage and non-heritage learners, leading to challenges in meeting the diverse needs of all students.

In response to these challenges, instructors from the current study have developed several strategies to cope with mixed proficiency levels in the classroom. For struggling students, there were teaching assistants (TAs), Chinese language helpers in the department, and foreign language teaching assistants (FLTAs) to provide after-class assistance. For high-achieving students, the instructor encourages them to take on extra work, such as writing more in-depth essays or improving their presentations. These strategies aim to better support learners with diverse proficiency levels and help them make the most of their language learning experience.

While these approaches have proven beneficial, further enhancement could be achieved through the implementation of a heritage track at MSU, specifically designed to address the unique needs of these students.

In order to open a heritage track at MSU, it would likely be most effective to focus on the upper levels, particularly if that is where most heritage learners are placed. Scheduling these classes at a time that accommodates the majority of heritage learners would be crucial to ensure adequate enrollment. However, enrollment size may still be an issue as MSU requires at least 15 students in an undergraduate course for it to run. In some cases, this requirement can be overridden, as has been done in certain departments for language classes, but careful consideration of class scheduling and enrollment size will be essential to successfully implement a heritage track at MSU. This addition could complement the strategies already employed by instructors in mixed proficiency classrooms, ultimately providing a more tailored and effective learning experience for heritage learners.

[RQ 6b: Utilization inference]:

From the perspective of students, are they placed into appropriate course levels?

Building upon the insights gained from the instructors' perspective on the effectiveness of the Chinese placement test, this section focuses on the students' perspective to provide a more comprehensive understanding of the placement process. While instructors have provided valuable information on the challenges and strategies associated with placing heritage learners into appropriate courses, it is essential to consider the students' experiences and feedback to assess the test's effectiveness fully. By doing so, I can gain a deeper understanding of how the test outcomes impact students' learning experiences and identify any potential gaps or mismatches between their prior knowledge, course expectations, and assigned course levels.

To address the research question, I analyzed students' interview responses and questionnaire data, revealing several key findings that suggest strong evidence in support of the utilization inference. Students generally achieved high GPAs in their assigned courses and reported feeling well-prepared and performing well in these courses, indicating appropriate course placement. However, similar to the instructors' perspective, there were instances where students faced difficulties in their assigned courses. In one particular case, a student with a background in traditional Chinese experienced challenges with the placement test due to its focus on simplified Chinese. It is important to note that traditional and simplified Chinese are different orthographic systems, with traditional characters being more complex and used in regions such as Taiwan and Hong Kong, while simplified characters are used in Mainland China.

Additionally, different phonetic systems are used in these regions, with Pinyin being used in Mainland China and Zhuyin in Taiwan. These differences highlight the importance of considering individual learner backgrounds when evaluating the test's effectiveness. This example emphasizes the need for open communication and collaboration between students and program supervisors to ensure appropriate course placement, especially when students encounter challenges that the placement test may not fully capture.

In examining the students' experiences with the Chinese placement test, this study highlights a gap in the literature regarding how students' simplified or traditional Chinese learning experiences are influenced by their prior traditional or simplified Chinese background. There has been limited research investigating the challenges and difficulties students may face when transitioning between these different orthographic and phonetic systems. This lack of research could be partly attributed to the growing preference for Hanyu Pinyin and simplified characters in recent years. The majority of teachers and students prefer Hanyu Pinyin, and even

Chinese heritage schools that traditionally teach Zhuyin have started to teach Hanyu Pinyin in the higher grades (Kwoh, 2007). The College Board's decision to use a computer-based AP Chinese test has further driven the adoption of Hanyu Pinyin, as it simplifies the input and typing process for students. Additionally, more schools have begun teaching simplified characters and Pinyin due to the increasing political and economic influence of Mainland China's official language, Putonghua (Wei & Hua, 2010). While this trend has led to a shift in focus away from traditional characters and Zhuyin, it is important to consider the diverse backgrounds and experiences of students when evaluating the effectiveness of language placement tests.

In light of these factors, it is crucial to further explore the challenges and difficulties faced by students with different backgrounds in traditional or simplified Chinese when transitioning between these orthographic and phonetic systems. By doing so, researchers and educators can better understand the unique needs of these students and develop more effective placement tests and language programs that cater to their diverse learning experiences.

[RQ 6c: Utilization inference]:

Are cut-off scores set appropriately?

The research question related to utilization inference focuses on the appropriateness of cut-off scores in the Chinese placement test. Cut-off scores are critical for valid score interpretations and uses, as they directly impact the test's precision and effectiveness in assigning students to suitable course levels. Accurate cut-off scores ensure that students are placed in appropriate courses, leading to better learning outcomes, engagement, and satisfaction with the language program.

To address the research question, I examined teacher ratings on their perceptions of item relevance and appropriateness to the course content of 100-level, 200-level, and 300-level

courses. The results revealed that the number of items matched to each course level corresponds with the cut-off scores, providing supportive validity evidence. However, there were variations in instructors' perceptions of item relevance and appropriateness for 7 out of 32 items, with these items' ratings differing by one-course level. Despite the overall consistency in cut-off scores and item relevance, the distribution of items across the course levels appears to be imbalanced, with the 100-level courses having significantly more items than the 200-level and 300-level courses. This imbalance may lead to a less accurate measurement of students' language proficiency in the upper-level courses, as there are fewer items to gauge their abilities.

Considering the analysis of the teacher ratings and the distribution of items across course levels, this study provides weak evidence concerning the appropriateness of the cut-off scores for the Chinese placement test. Although the instructors' perceptions of item relevance align with the established cut-off scores for the intended course levels, the imbalanced distribution of items across course levels raises concerns regarding the test's accuracy in assessing students' language proficiency, particularly for upper-level courses. In light of the findings, several suggestions can be made to improve the Chinese placement test and enhance the evidence for appropriate test score uses and interpretations:

1. Reevaluate and revise the test items to ensure a more balanced distribution across all course levels. A more even distribution of items will help in accurately assessing students' language proficiency for upper-level courses and lead to more precise course placements.
2. Consider conducting a comprehensive review of the test items, taking into account the variations in instructors' perceptions of item relevance and appropriateness. This process

may involve revising, removing, or adding items to better align with the course levels and reduce variations in instructors' perceptions.

3. Regularly update and review the test content to ensure that it reflects the evolving course content and student profiles, which will contribute to maintaining the accuracy and relevance of the test over time.

By implementing these suggestions, the Chinese placement test can be improved, leading to a better assessment of students' language proficiency and more accurate course placements. Furthermore, these improvements can contribute to the overall quality of Chinese language programs, as accurate placement of students promotes more effective teaching and learning experiences.

[RQ 7: Consequence implication inference]:

Does the test have positive effects on Chinese teaching and learning?

The consequence implication inference of the Chinese placement test is explored, focusing on its effects on Chinese teaching and learning. To address this research question, qualitative analysis of feedback from students and instructors collected through interviews was conducted. The findings provide strong evidence that the test has positive effects on Chinese language instruction and student learning experiences. Students generally reported that the test results were accurate and helpful for course selection, while instructors found it easier to manage students placed at appropriate levels.

However, the results also revealed that some students viewed the test results as a suggestion rather than a strict guideline and made their own decisions about their level of comfort and willingness to engage with the course material. It is crucial to acknowledge that instances of students not being placed at the most suitable level are not always due to

misplacement, but rather the choices these students make. Program administrators and course instructors can provide recommendations to the students regarding placement decisions based on the match between course difficulty and students' current proficiency level. However, there are other considerations such as whether the course credits will be recognized by their major or program, as this can have a direct impact on their graduation timeline. For instance, from the interview, an instructor shared that an engineering student was advised to take a 200-level class based on the placement test results and the Chinese program coordinator's recommendation, but he insisted on taking a 300-level course because the 200-level course credits would not be useful for his major, and he needed to graduate. In the end, he did not take the Chinese class.

Moreover, students' motivation to learn a foreign language may also influence their final course selection, which can indirectly impact the effectiveness of the Chinese placement test. Some students are not highly motivated to learn a foreign language and only take the course to fulfill the university's foreign language requirement. As a result, even if their proficiency is higher, they may choose to enroll in lower-level courses to minimize effort and secure easy credits. Consequently, classes may consist of students with mixed proficiency levels, which presents challenges for course instructors.

These challenges, revealed through the interviews, underscore the limitations of the Chinese placement test in accommodating students with varying levels of proficiency. Instructors reported having to tailor course difficulty to accommodate the majority of students and follow their learning pace, which can be demanding. Moreover, the lack of teaching assistant support due to the pandemic has further exacerbated these challenges, highlighting the need for additional resources and strategies to support both instructors and students in diverse classroom settings.

Building on the identified challenges, several recommendations can be considered to enhance the effectiveness of the Chinese placement test and better support instructors and students. While increasing the number of teaching assistants may not always be feasible, exploring alternative support resources, such as peer tutoring or online materials, could help manage classes with diverse proficiency levels more effectively. For students who have taken a break from their studies or experienced a decline in their language proficiency, alternative support strategies could be offered. These may include providing access to supplementary learning resources, creating customized study plans, or allowing students to audit lower-level classes alongside their current courses to bridge proficiency gaps. Ultimately, fostering clear communication between students, program coordinators, and instructors through in-person interviews becomes essential to ensure the best possible learning outcomes for all students. By maintaining open dialogue and considering individual factors, a more nuanced and effective approach to course placement can be achieved, maximizing the benefits of the Chinese placement test for both teaching and learning.

Limitations and future research directions

The current study has several limitations and areas for future research to address. First, validation should be an ongoing endeavor, as this study has highlighted specific psychometric issues within the Chinese placement test. These issues might stem from factors such as ambiguous item phrasing, poorly selected distractors, and incongruent information in the prompts compared to the intended answers. Although suggested revisions were proposed, it is not clear to what extent the issues will be resolved. Ideally, the study would have collected more data using a revised version of the test, incorporating the suggested revisions, and re-run the analysis to investigate whether the revisions effectively addressed the identified issues. However,

this limitation is closely related to the small sample size and the tight timeline, which hinders the robustness of Rasch analysis and underscores the need for future studies to address a range of factors, including sample size and time constraints, to strengthen the validation process.

Second, the relatively small sample size of 28 students also limits the generalizability of the results The participants in the current study were drawn from three course levels and three instructors at Michigan State University. This specific context may limit the generalizability of the identified issues and the results yielded in the study to other Chinese language programs in different institutions. Thus, when interpreting the results, readers should consider the specific context in which the study was conducted, such as the unique characteristics of the Chinese language program at Michigan State University, the specific course materials used, and the teaching approaches employed by the course instructors. Future research should not only aim to include larger sample sizes but also address other factors, such as diversity in participants' backgrounds and institutional contexts, to enhance the generalizability and reliability of the findings.

Third, another related limitation is the inability to carry out the originally planned analysis related to the utilization inference, which was intended to examine whether cut-off scores are set appropriately. This analysis involved comparing the class performance of students placed by the placement test to those who did not take the test. Due to the small sample size, this analysis could not be carried out. Future research should examine the performance of students placed by the test in comparison to those who did not take the test, as this would provide additional insights into the effectiveness of the placement test and the appropriateness of the cut-off scores.

A fourth limitation is the potential impact of the COVID-19 pandemic on data collection and student experiences. While data collection took place in Spring 2022 when most classes were in-person, it remains unclear how the pandemic-induced shift to online instruction may have affected students' learning motivation, course enrollment, and language performance. Participants in this study might have experienced different language learning trajectories compared to students who did not face the challenges posed by the pandemic. Future research should investigate the potential impact of the COVID-19 pandemic on language learning, placement test accuracy, and students' language learning experiences. Some students indicated in interviews that they felt foreign language courses were significantly impacted by the shift to online instruction, as it disrupted the interactive nature of in-person classes.

In light of these limitations, future research should also expand the scope of the study to include different contexts and a broader range of students, which would enhance the generalizability of the findings. By addressing these points in future research, a clearer understanding of the validity evidence for the score uses and interpretations of the Chinese placement test can be achieved, ultimately benefiting students and educators in the field of Chinese language learning.

CHAPTER 6: CONCLUSIONS

In conclusion, with this study I aimed to provide a comprehensive examination and evaluation of the test score uses and interpretations for the listening and reading sections of an in-house, college-level Chinese placement test. The primary goal was to address the existing gaps in the literature, particularly the limited discussion on tests in languages other than English and the methodological constraints of previous research. To achieve this, I utilized an argument-based validation framework, collecting and evaluating both quantitative and qualitative validity evidence. By employing mixed-methods, I sought to thoroughly assess the functioning of test items, identifying any problematic items and proposing revisions as necessary. Additionally, it aimed to utilize the empirical findings to inform improvements to the placement test, evaluate the strength of the validity argument based on collected evidence, and offer a comprehensive analysis of the test score interpretations and uses.

The findings of this study contribute to the larger discussion on the practices of foreign language assessment and argument-based test validation. Furthermore, the research offers valuable insights into the ongoing development of validity research in the field of second language testing. By providing a comprehensive examination of the Chinese placement test, this study helps to enhance the understanding of test score uses and interpretations, supporting more effective and reliable language placement decisions for students in higher education settings.

REFERENCES

ACTFL. (2012). ACTFL proficiency guidelines. *ACTFL*. https://www.actfl.org/uploads/files/general/ACTFLProficiencyGuidelines2012.pdf

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. *American Educational Research Association.*

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* Continuum.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Baralt, M. (2012). Coding qualitative data. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition* (pp. 222-244). Wiley-Blackwell.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assessing Writing, 37*, 1–12. https://doi.org/10.1016/j.asw.2018.01.001

Bernhardt, E. B., Rivera, R. J., & Kamil, M. L. (2004). The practicality and efficiency of web-based placement testing for college-level language programs. *Foreign Language Annals, 37*(3), 356–365. https://doi.org/10.1111/j.1944-9720.2004.tb02694.x

Bond, T., & Fox, C. (2007). Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.). Lawrence Erlbaum Associates.

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist, 27*(7), 1077-1105. https://doi.org/10.1080/13854046.2013.809795

Chalhoub-Deville, M., & Deville, C. (2018). Revisiting language testing validation: Empirical, analytical, and theoretical considerations. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment* (pp. 33-48). Springer.

Campbell, R. (2000). Heritage language. In J. W. Rosenthal (Ed.), *Handbook of undergraduate second language education* (pp. 165–184). Lawrence Erlbaum Associates.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C. A. (2020). Argument-based validation in testing and assessment. SAGE Publications.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language.* Routledge. Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.

DeMars, C. (2010). *Item response theory.* Oxford University Press.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). *The qualitative research interview. Medical education, 40*(4), 314-321. https://doi.org/10.1111/j.1365-2929.2006.02418.x

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Lawrence Erlbaum Associates.

Duncan, P. W., Bode, R., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*(7), 953. https://doi.org/10.1016/S0003-9993(03)00035-2

Eckes, T. (2015). Introduction to many facet Rasch measurement: Analyzing and evaluating rater mediated assessment (2nd ed.). Peter Lang.

Eda, S., Itomitsu, M., & Noda, M. (2008). The Japanese skills test as an on-demand placement test: Validity comparisons and reliability. *Foreign Language Annals, 41*(2), 218–236. https://doi.org/10.1111/j.1944-9720.2008.tb03290.x

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Lawrence Erlbaum Associates.

Fan, C. W., Yazdani, F., Carstensen, T., & Bonsaksen, T. (2021). Rasch analysis of the self-efficacy for therapeutic use of self questionnaire in Norwegian occupational therapy students. *Scandinavian Journal of Occupational Therapy, 28*(4), 274-284. https://doi.org/10.1080/11038128.2020.1726453

Ferne T., Rupp A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges and recommendations. *Language Assessment Quarterly, 4*(2), 113–148. https://doi.org/10.1080/15434300701375923

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

Friedrich, F., Konietschke, F., & Pauly, M. (2019a). MANOVA.RM: Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs [Computer software]. Version 0.4.1. http://github.com/smn74/MANOVA.RM

Friedrich, S., Konietschke, F. & Pauly, M. (2019b). Resampling-based analysis of multivariate data and repeated measures designs with the R package MANOVA.RM. *The R Journal, 11*(2), 380–400. https://doi.org/10.32614/RJ-2019-051

Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication.* New York University Press. https://doi.org/10.18574/nyu/9780814732939.001.0001

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* SAGE Publications.

Heilenman, L. (1983). The use of a cloze procedure in foreign language placement. *The Modern Language Journal, 67*(2), 121–126. https://doi.org/10.1111/j.1540-4781.1983.tb01482.x

Hu, X., Jiang, Y., & Bi, H. (2022). Measuring science self-efficacy with a focus on the perceived competence dimension: using mixed methods to develop an instrument and explore changes through cross-sectional and longitudinal analyses in high school. International *Journal of STEM Education, 9*(1), 47. https://doi.org/10.1186/s40594-022-00363-x

In'nami, Y., & Koizumi, R. (2016). Factor *structure of the revised TOEIC test: A multiple-sample analysis. Language* Testing, *33*(1), 99-119. https://doi.org/10.1177/0265532211413444

Isbell, D. R., Winke, P. M., & Gass, S. M. (2019). Using the ACTFL OPIc to assess and monitor progress in a tertiary foreign languages program. *Language Testing, 36*(3), 439–465. https://doi.org/10.1177/0265532218798139

Jefferies, P., Bremer, E., Kozera, T., Cairney, J., & Kriellaars, D. (2020). Psychometric properties and construct validity of PLAYself: a self-reported measure of physical literacy for children and youth. *Applied Physiology, Nutrition, and Metabolism, 46*(6), 579-588. https://doi.org/10.1139/apnm-2020-0410

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Greenwood.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 74–83. https://doi.org/10.1111/jedm.12001

Kane, M., Crooks, T., & Cohen, A. (2005). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology, 5*(2), 60–83. https://www.lltjournal.org/item/2357

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Kondo–Brown, K. (2005). Differences in language skills: Heritage language learner subgroups and foreign language learners. *The Modern Language Journal, 89*(4), 563-581. https://doi.org/10.1111/j.1540-4781.2005.00330.x

Kunnan, A. J. (2000). Fairness and validation in language assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium* (pp. 1-16). Cambridge University Press.

Kwoh, S. (2007). Mainstreaming and professionalizing Chinese-language education: A new mission for a new century. *Chinese America: History and Perspectives*, 261-265.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing, 34*(4), 451–475. https://doi.org/10.1177/0265532217713951

Li, D., & Duff, P. (2008). Issues in Chinese heritage language education and research at the postsecondary level. In He, A. W., & Xiao. Y. (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp.13–32). National Foreign Language Resource Center.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions, 12*, 636. https://www.rasch.org/rmt/rmt122m.htm

Linacre, J. M. (2012). A user's guide to Winsteps Ministeps Rasch-model computer programs [version 4.7.1]. Retrieved from http://www.winsteps.com/index.htm

Linacre, J. M. (2016). WINSTEPS (Version 4.7.1) [Computer program]. Chicago: MESA Press.

Long, A. Y., Shin, S.-Y., Geeslin, K., & Willis, E. W. (2018). Does the test work? Evaluating a web-based language placement test. *Language Learning & Technology, 22*(1), 137–156. https://dx.doi.org/10125/44585

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency overtime? *Foreign Language Annals, 52*(1), 66-86. https://doi.org/10.1111/flan.12379

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Mozgalina, A. & Ryshina–Pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *The Modern Language Journal, 99*(2), 346-370. https://doi.org/10.1111/modl.12217

Norris, J. M. (2004). Validity evaluation in foreign language assessment [Unpublished doctoral dissertation]. Georgetown University.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal, 93*, 836-847. https://doi.org/10.1111/j.1540-4781.2009.00976.x

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117. https://doi.org/10.1023/A:1006985528729

Sohn, S. (2004). Placement of Korean heritage speakers: Challenges and strategies. Unpublished invited lecture, *Center for Korean Research*, University of British Columbia.

Tai, J. H. (1994). Chinese classifier systems and human categorization. In M. Chen (Ed), *In honor of William S. Y. Wang: Interdisciplinary studies on language and language change* (pp. 479-494). Taipei: Pyramid Press.

Tigchelaar, M., Bowles, R., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency. *Foreign Language Annals, 50*(3), 584–600. https://doi.org/10.1111/flan.12286

Toulmin, S. ([1958] 2012). The uses of argument. Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Toulmin, S. (2001). Return to reason. Harvard University Press.

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review, 1*(2), 133-147. https://doi.org/10.1016/j.edurev.2006.05.001

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Wei, L., & Hua, Z. (2010). Voices from the diaspora: Changing hierarchies and dynamics of Chinese multilingualism. *International Journal of the Sociology of Language, 205*, 155–171. https://doi.org/10.1515/ijsl.2010.043

Winke, P., Zhang, X., & Pierce, S. J. (2022). A closer look at a marginalized test method: Self-assessment as a measure of speaking proficiency. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263122000079

Winke, P., Zhang, X., Rubio, F., Gass, S., Soneson, D., & Hacking, J. (2018). The proficiency profile of language students: Implications for programs. *Second Language Research & Practice, 1*(1). https://doi.org/10125/69840

Wolfe, E. W & Smith E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II–Validation activities. *Journal of Applied Measurement, 8,* 204–234.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Mesa Press.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170. https://doi.org/10.1177/0265532209349465

Yan, X., & Staples, S. (2020). Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE. *Language Testing, 37*(2), 189–214. https://doi.org/10.1177/0265532219876226

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199–225. https://doi.org/10.1177/0265532214557113

Zhang, X., Winke, P., & Clark, S. (2020). Background characteristics and oral proficiency development over time in lower-division college foreign language programs. *Language Learning, 70*(3), 807-847. https://doi.org/10.1111/lang.1239

APPENDIX 1: INSTRUCTOR INTERVIEW QUESTIONS

1. In your opinion what were the essential skills that students need to have for successful performance in the courses you teach

2. Do you feel the difficulty of the course was appropriate given the language proficiency levels of all students?

3. Do you feel some students were misplaced to your class? What did you do to accommodate these students?

4. Are you aware of the placement procedures for the Chinese language courses?

5. After reviewing the placement test, do you agree what is assessed in the test is targeting and representative of the content covered in your class?

6. Is there anything you feel important in your language class but is missing from the test?

APPENDIX 2: STUDENT QUESTIONNAIRE

Personal information survey

1. Your first name: _____; Your last name: _____

2. Your MSU PID (the number found under your name on your student ID, starting with 'A' and then 8 digits): _____

3. Your email address: _____

4. Your age: _____

5. Your gender: _____

6. What is your major: _____

7. Year of graduation: _____

Questions on students' perception of the placement test results

1. Have you taken the Michigan State University Chinese placement test (https://msu.co1.qualtrics.com/jfe/form/SV_a2C5uBOWKlTCdoN)?

Yes/No

2. Which Chinese course are you taking this semester?

CHS101/CHS102/CHS201/CHS202/CHS301/CHS302/CHS401/CHS402

3. According to the test result, which Chinese course were you placed to?

CHS101/CHS102/CHS201/CHS202/CHS301/CHS302

4. Your GPA of the course to which you were placed to?

5. For the course to which you were placed, please rate your performance in the following categories on a scale of 5 (1- unacceptable; 2- needs improvement; 3- meets expectations; 4 - exceeds expectations; 5 - outstanding):

- Your overall preparedness for the course before the course started
- Your overall course performance
- Your overall Chinese proficiency
- Please provide any comments if you have.

6. For the course to which you were placed, please rate the overall difficulty of the course on a scale of 5 (1 - very easy; 2 - easy; 3 - medium; 4 - difficult; 5 - very difficult).

Could you please elaborate on your selection? Which aspects of the course you find easy, medium, or difficult?

7. For the course to which you were placed, in your opinion what were the essential skills for successful performance?

APPENDIX 3: STUDENT INTERVIEW QUESTIONS

1. Have you taken the Michigan State University Chinese placement test (https://msu.co1.qualtrics.com/jfe/form/SV_a2C5uBOWKlTCdoN)?

2. According to the test result, which Chinese course(s) are you taking right now?

3. If you have completed one or more Chinese language courses. What are your GPAs for the course(s)?

4. For the course to which you were placed, could you please comment on:

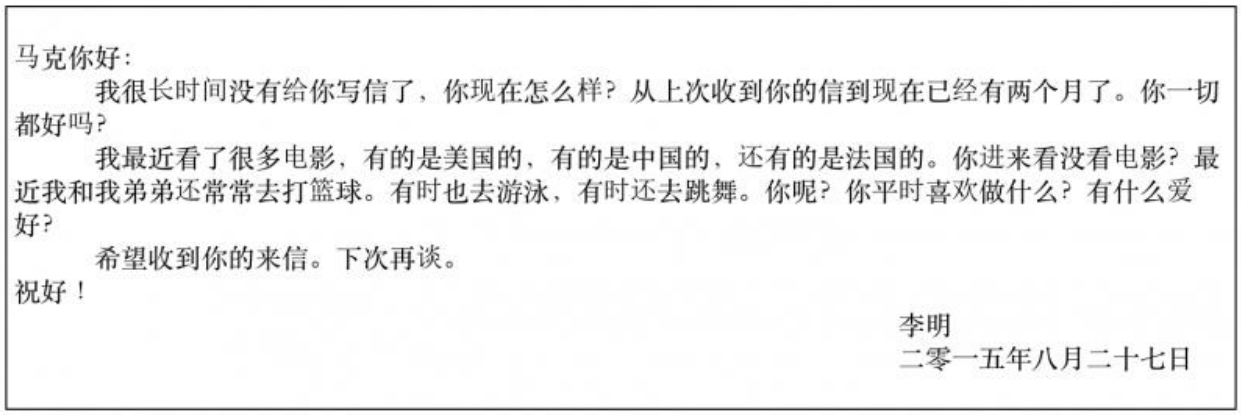
- Your overall course performance
- Your overall Chinese proficiency
- Your overall preparedness for the course before the course started

5. For the course to which you were placed, do you think the overall difficulty of the course was appropriate given your Chinese proficiency level? In other words, do you feel the course was too easy or too difficult given your Chines proficiency level?

6. For the course to which you were placed, in your opinion what were the essential skills for successful performance?

APPENDIX 4: ITEMS LOADED ON THE SAME DIMENSION

Figure 20. Read the email from Li Ming to Ma Ke. Then answer Questions 7 to 11 below.



马克你好：

我很长时间没有给你写信了，你现在怎么样？从上次收到你的信到现在已经有两个月了。你一切都好吗？

我最近看了很多电影，有的是美国的，有的是中国的，还有的是法国的。你进来看没看电影？最近我和我弟弟还常常去打篮球。有时也去游泳，有时还去跳舞。你呢？你平时喜欢做什么？有什么爱好？

希望收到你的来信。下次再谈。

祝好！

李明

二零一五年八月二十七日

Reading #7. On what date, does the letter written?

a. September 8, 2015

b. February 28, 2014

c. August 27, 2015

Reading #8. How long has it been since their last correspondence?

a. Three weeks

b. Two months

c. Four months

Reading #9. What kinds of movies has Li Ming seen recently?

a. French, American and Chinese

b. American, British and Chinese

c. Italian, Russian and Chinese

Reading #10. What kinds of sports has the writer done lately?

a. Soccer and jogging

b. Basketball and swimming

c. Football and swimming

Reading #11. Based on this letter, how well do you think the two know each other?

a. They are good friends who see each other very often.

b. They have never met each other, but they are relatives.

c. They are pen pals who are not familiar with each other.

APPENDIX 5: ITEM RELEVANCE AND DIFFICULTIES

Table 21. Descriptive statistics for students' ratings for item relevance and difficulties

Item ID	Relevance				Difficulties			
	Total	100	200	300	Total	100	200	300
L01	4.64	4.23	5.33	4.5	3.11	3.54	2.78	2.67
L02	4.79	4.38	5.33	4.83	2.89	3.08	3	2.33
L03	4.46	4.15	5	4.33	3.11	3.54	2.78	2.67
L04	4.39	3.92	4.89	4.67	2.71	3.15	2.67	1.83
L05	4.5	5.15	4.22	3.5	1.93	2.15	2	1.33
L06	4.39	4.38	4.78	3.83	2.11	2.38	2	1.67
L07	4.79	4.77	4.67	5	2.54	2.38	2.89	2.33
L08	4.71	4.92	5	3.83	2.93	2.69	3.22	3
L09	4.68	4.85	5	3.83	3.14	3	3.44	3
L10	4.71	4.92	5	3.83	3.14	2.92	3.56	3
L11	4.5	4.54	4.78	4	3.79	4	3.56	3.67
L12	4.04	3.69	4.44	4.17	4.39	4.77	4.11	4
L13	4.11	3.92	4.22	4.33	4.21	4.15	4.11	4.5
L14	3.79	3.15	4.56	4	4.5	4.38	4.44	4.83
R01	3.25	2.38	4	4	3.29	4.31	3	1.5
R02	3.86	3.38	4.67	3.67	3.86	4.46	3.67	2.83
R03	4.93	5.46	5.33	3.17	2.46	2.54	2.67	2
R04	4.61	5.15	4.78	3.17	3.21	3.46	3	3
R05	5.18	5.54	5.56	3.83	1.68	1.85	1.78	1.17
R06	5.07	5.46	5.33	3.83	1.64	1.85	1.67	1.17
R07	5.04	5.54	5	4	1.61	1.69	1.89	1
R08	5.11	5.54	5.22	4	1.82	2	2.11	1
R09	5.07	5.54	5.11	4	1.79	2.08	1.78	1.17
R10	5.11	5.46	5.33	4	1.75	2	1.89	1
R11	4.79	5	5	4	2.46	2.46	2.89	1.83
R12	4.5	4	5.44	4.17	3.18	3.85	2.67	2.5
R13	4.64	4.23	5.44	4.33	3.07	3.46	3	2.33

Table 21 (cont'd)

R14	4.93	4.77	5.56	4.33	2.61	3.08	2.22	2.17
R15	5.11	4.92	5.78	4.5	2.14	2.62	1.89	1.5
R16	5.25	5.38	5.67	4.33	2.32	2.85	2.11	1.5
R17	5.29	5.46	5.56	4.5	1.82	2	1.89	1.33
R18	5.29	5.38	5.67	4.5	2.07	2.31	2.22	1.33

APPENDIX 6: RESULTS OF DIF

Table 22. Results of DIF: The Mantel-Haenszel test results by item

Item ID	Mantel-Haenszel χ^2	*p*-value	Adj. *p*-value
L01	0.71	.4	.9
L02	0.39	.53	.9
L03	4.55	.03	.66
L04	0.02	.9	.95
L05	2.56	.11	.66
L06	0.29	.59	.9
L07	0.04	.84	.95
L08	0.25	.62	.9
L09	3.27	.07	.66
L10	0.01	.93	.95
L11	0.08	.77	.95
L12	0.25	.62	.9
L13	3.07	.08	.66
L14	0.52	.47	.9
R01	2.38	.12	.66
R02	< .01	.94	.95
R03	0.21	.65	.9
R04	< .01	.95	.95
R05	0.03	.86	.95
R06	0.97	.32	.9
R07	0.69	.4	.9
R08	0.27	.6	.9
R09	0.18	.67	.9
R10	0.02	.89	.95
R11	1.62	.2	.84
R12	1.41	.24	.84
R13	2.55	.11	.66
R14	1.52	.22	.84
R15	0.42	.52	.9

Table 22 (cont'd)

R16	1.23	.27	.86
R17	0.55	.46	.9
R18	0.41	.52	.9

Note: Multiple comparisons made with Benjamini-Hochberg adjustment of *p*-values

APPENDIX 7: MISFITTING ITEMS AND PROPOSED REVISIONS

Figure 21. Reading item #3. If you want to order soup, how many choices do you have?

a. 3 **b. 4** c. 5

Suggested revisions:

Possible revision 1:

Replacing soup with fried rice in the stem:

If you want to order fried rice, how many choices do you have?

a. 3 b. 4 c. 5

Possible revision 2:

Revise the dish on the menu that causes ambiguity and confusion

Figure 22. Reading item #3. If you want to order soup, how many choices do you have?

Figure 23. Reading #6. Here is a message that Xiao Li sent to Lao Wang. Please answer the following questions after reading the note:

老王：

后天晚上请你七点来我家吃晚饭，我六点三刻在三路车站等你。后天见。

小李 四月十八号 星期五

At what time, should they meet?

a. 5:30 PM **b. 6:45 PM** c. 7:00 PM

Suggested revisions:

a. 6:03 PM b. 6:30 PM **c. 6:45 PM**

Figure 24. Reading #2. Is this a sign for?

a. 公车时刻表	**b.商场上班时间**	c.飞行时间
a. a bus schedule	**b. shopping mall hours**	c. Flight hours

Suggested revisions:

a. 公车时刻表	**b.餐厅开放时间**	c.飞行时间
a. a bus schedule	**b. restaurant hours**	c. Flight hours

Reading #12. 教室有几 （ ） 椅子？

How many (classifier needed) chairs are there in the classroom?

a. 张	b. 条	**c. 把**
a. *zhāng*	b. tiáo	**c. bǎ**

Suggested revisions:

a. 只	b. 条	**c. 把**
a. *zhī*	b. tiáo	**c. bǎ**

APPENDIX 8: POST-HOC TEST RESULTS (TOTAL)

Table 23. All post-hoc pair-wise comparisons results (total scores)

Contrast	*t*-value	SE	df	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	5.9	0.92	25	<.001	1.67
Time 1 200-level - Time 1 100-level	7.1	1.77	35.7	.005	1.68
Time 2 200-level - Time 1 100-level	8.5	1.77	35.7	<.001	2.01
Time 1 300-level - Time 1 100-level	9	1.95	35.7	<.001	2.13
Time 2 300-level - Time 1 100-level	9.4	1.95	35.7	<.001	2.22
Time 1 200-level - Time 2 100-level	1.2	1.77	35.7	1	0.28
Time 2 200-level - Time 2 100-level	2.6	1.77	35.7	1	0.62
Time 1 300-level - Time 2 100-level	3.2	1.95	35.7	1	0.76
Time 2 300-level - Time 2 100-level	3.5	1.95	35.7	1	0.83
Time 2 200-level - Time 1 200-level	1.4	1.21	25	1	0.4
Time 1 300-level - Time 1 200-level	2	2.16	35.7	1	0.47
Time 2 300-level - Time 1 200-level	2.3	2.16	35.7	1	0.54
Time 1 300-level - Time 2 200-level	0.6	2.16	35.7	1	0.14
Time 2 300-level - Time 2 200-level	0.9	2.16	35.7	1	0.21
Time 2 300-level - Time 1 300-level	0.3	1.4	25	1	0.08

APPENDIX 9: POST-HOC TEST RESULTS (LISTENING)

Table 24. All post-hoc pair-wise comparisons results (listening scores)

Contrast	*t*-value	SE	df	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	2.4	0.7	25	.03	0.68
Time 1 200-level - Time 1 100-level	3.1	1.09	42	.09	0.68
Time 2 200-level - Time 1 100-level	3.8	1.09	42	.02	0.83
Time 1 300-level - Time 1 100-level	3.5	1.2	42	.09	0.76
Time 2 300-level - Time 1 100-level	3.5	1.2	42	.09	0.76
Time 1 200-level - Time 2 100-level	0.7	1.09	42	1	0.15
Time 2 200-level - Time 2 100-level	1.3	1.09	42	1	0.28
Time 1 300-level - Time 2 100-level	1	1.2	42	1	0.22
Time 2 300-level - Time 2 100-level	1	1.2	42	1	0.22
Time 2 200-level - Time 1 200-level	0.6	0.93	25	1	0.17
Time 1 300-level - Time 1 200-level	0.3	1.33	42	1	0.07
Time 2 300-level - Time 1 200-level	0.3	1.33	42	1	0.07
Time 1 300-level - Time 2 200-level	-0.3	1.33	42	1	-0.07
Time 2 300-level - Time 2 200-level	-0.3	1.33	42	1	-0.07
Time 2 300-level - Time 1 300-level	0	1.07	25	1	0

APPENDIX 10: POST-HOC TEST RESULTS (READING)

Table 25. All post-hoc pair-wise comparisons results (readings scores)

Contrast	*t*-value	*SE*	*df*	*p*-value	Cohen's *d*
Time 2 100-level - Time 1 100-level	3.4	0.66	25	<.001	0.96
Time 1 200-level - Time 1 100-level	3.9	0.97	44.4	.003	0.83
Time 2 200-level - Time 1 100-level	4.7	0.97	44.4	<.001	1
Time 1 300-level - Time 1 100-level	5.6	1.07	44.4	<.001	1.19
Time 2 300-level - Time 1 100-level	5.9	1.07	44.4	<.001	1.25
Time 1 200-level - Time 2 100-level	0.5	0.97	44.4	1	0.11
Time 2 200-level - Time 2 100-level	1.3	0.97	44.4	1	0.28
Time 1 300-level - Time 2 100-level	2.1	1.07	44.4	.76	0.45
Time 2 300-level - Time 2 100-level	2.5	1.07	44.4	.38	0.53
Time 2 200-level - Time 1 200-level	0.8	0.88	25	1	0.23
Time 1 300-level - Time 1 200-level	1.6	1.18	44.4	1	0.34
Time 2 300-level - Time 1 200-level	2	1.18	44.4	1	0.42
Time 1 300-level - Time 2 200-level	0.9	1.18	44.4	1	0.19
Time 2 300-level - Time 2 200-level	1.2	1.18	44.4	1	0.25
Time 2 300-level - Time 1 300-level	0.3	1.01	25	1	0.08