分类号：H310.42　　单位代码：10335

密　级：无　　学　　号：12005002

浙江大学

博士学位论文

中文论文题目： 测量学与涉考群体视阈下的英语水平测试公平性研究：构建测试公平性评估模型

英文论文题目： **Evaluating the fairness of an English proficiency test from the psychometric and stakeholders' perspectives: Toward a model of test fairness evaluation**

申请人姓名：张娟

指导教师：何莲珍 教授

合作导师：

专业名称：外国语言文学

研究方向：语言测试

所在学院：外国语学院

论文提交日期 2025 年 6 月 30 日

Evaluating the Fairness of an English Proficiency Test from the Psychometric and Stakeholders' Perspectives: Toward a Model of Test Fairness Evaluation

Juan Zhang

A dissertation submitted in conformity with the requirements

for the degree of Doctor of Philosophy

in Foreign Language and Literature

Supervised by Prof. He Lianzhen

School of International Studies, Zhejiang University

June 2025

测量学与涉考群体视阈下的英语水平测试公平性研究：构建测试公平性评估模型

论文作者签名： 张娟

指导教师签名： 何立荣

论文评阅人 1：

评阅人 2：

评阅人 3：

评阅人 4：

评阅人 5：

答辩委员会主席：洪　岗教授　浙江外国语学院

委员 1：洪　岗教授　浙江外国语学院

委员 2：钱毓芳教授　浙江工商大学外国语学院

委员 3：梁君英教授　浙江大学外国语学院

委员 4：王　敏教授　浙江大学外国语学院

委员 5：闵尚超教授　浙江大学外国语学院

答辩日期：2025 年 5 月 16 日

Evaluating the Fairness of an English Proficiency Test from the Psychometric and Stakeholders' Perspectives: Toward a Model of Test Fairness Evaluation

Author's signature: 张勇

Supervisor's signature: 何江

Thesis reviewer 1: _____

Thesis reviewer 2: _____

Thesis reviewer 3: _____

Thesis reviewer 4: _____

Thesis reviewer 5: _____

Chair: (Committee of oral defence) 洪　岗教授　浙江外国语学院

Committeeman 1: 洪　岗教授　浙江外国语学院

Committeeman 2: 钱毓芳教授　浙江工商大学外国语学院

Committeeman 3: 梁君英教授　浙江大学外国语学院

Committeeman 4: 王　敏教授　浙江大学外国语学院

Committeeman 5: 闵尚超教授　浙江大学外国语学院

Date of oral defence: May 16th, 2025

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**浙江大学**或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：张如明　　签字日期：2015 年 5 月 12 日

学位论文版权使用授权书

本学位论文作者完全了解**浙江大学**有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权**浙江大学**可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：张如明　　导师签名：何正斌

签字日期：2015 年 5 月 12 日　　签字日期：2015 年 5 月 12 日

学位论文作者毕业后去向：

工作单位：　　电话：

通讯地址：　　邮编：

摘要

“公平性”已经成为评价语言测试的重要维度，同时也是语言测试领域追求的核心价值观之一。测试公平性这一概念具有双重属性，即测量学属性和价值判断属性。已有研究对一些高风险语言水平测试的公平性进行了有益探索。其中，部分研究运用定量研究方法评估语言测试的公平性，另有一些研究运用定性研究方法探究涉考群体对测试公平性的看法。然而，由于测试公平性这一概念具有多维性、社会建构性以及情境依赖性等特征，全面、系统的语言测试公平性评估工作往往因缺乏理论指导而难以顺利开展。

以中国某高校的一项高风险英语水平测试为例,本研究从测量学与涉考群体两个视角着手，评估该测试的公平性，并构建测试公平性评估模型。基于学界对测试公平性各维度的界定和现有的理论、实证研究成果，本研究先提出了一个测试公平性评估的概念模型。该模型以同心圆的形式呈现测试公平性的双重属性，内圈由测试公平性的四个重要维度构成，即可比性、可及性、一致性和问责制。这些维度既是测试公平性研究所关注的重要方面，同时也是语言测试项目公平与否的核心判别标准。模型外圈由四个核心涉考群体构成，即考生、考试开发者、施考人员和考试使用者。

应用上述概念模型，本研究采用混合研究方法评估了该高风险英语水平测试的公平性。定量层面：第一，本研究采用题项、题组和测试功能差异检验方法探究了测试分数在文、理科考生群体之间的可比性。第二，本研究运用描述性与推断性统计方法探究了听力、阅读输入材料在多套试卷之间的可比性。第三，参照概念模型涵盖的四个测试公平性维度，本研究自主设计了一份问卷，对 1,646 名考生进行了问卷调查，并采用探索性和验证性因子分析法对问卷数据的因子结构进行了探索和验证。定性层面，本研究采用半结构化访谈法探究了不同涉考群体对该语言水平测试公平性在四个测试公平性维度方面的看法。20 名考生、6 名教师、2 名施考人员和 3 名考试使用者参与了一对一访谈。本研究采用主题分析法对转写后的访谈文本进行了系统分析。

从测量学视角与涉考群体视角观之，定量与定性研究结果表明该高风险语言水平测试整体而言较为公平。此外，研究结果为概念模型中的四个测试公平性维度提供了实证数据支持。定量层面：第一，关于考试分数在文、理科考生

群体之间的可比性，4 套听力和阅读题中，5%的听力题存在题项功能差异，40%的听力题组、22.22%的阅读题组存在题组功能差异，2 套听力和阅读题存在试卷功能差异。内容分析显示，题项、题组、试卷功能差异在听力与阅读测试中不构成测试偏颇。研究发现表明，测试分数在文、理科考生群体之间具有可比性。第二，听力输入材料的词汇、句法、语篇特征及语速在 4 套听力试卷之间具有可比性，阅读输入材料的词汇特征在 4 套阅读试卷之间具有可比性，句法、语篇特征及可读性指标上略有差异。第三，对测试公平性问卷数据的探索性和验证性因子分析结果表明，问卷因子与测试公平性各维度之间基本吻合，考生认为测试相关信息公开透明、施测环境及测试流程一致、测试过程中能够充分展现自己的英语水平、测试结束后可以申请成绩复议或对测试服务进行投诉。

定性层面：第一，关于涉考群体对该英语水平测试公平性的看法，主题分析结果表明，基于访谈数据得出的四个主题与概念模型中测试公平性的四个维度基本吻合，涉考群体认为该英语水平测试在可比性、可及性、一致性和问责制维度均实现了测试公平。第二，关于影响涉考群体对该英语水平测试公平性看法的因素，主题分析结果进一步揭示了影响涉考群体对该英语水平测试公平性看法的四个因素，即社会文化、教育机会、测试机构和考生个体因素。社会文化因素主要包括：英语在中国的重要地位、中国各高校的语言测评实践。教育机会因素主要包括：大学入学前考生接受的英语教育质量差异、大学就读期间英语学习资源的丰富性和有效性。测试机构因素主要包括：校方政策、测试开发人员在语言测试与评估方面的专业素养、该英语水平测试的施考设施。考生个体因素主要包括：考生的英语水平、考生对该英语水平测试的个人看法。

基于以上量性、质性研究发现，本研究在概念模型的基础上提出了一个测试公平性评估模型。与研究伊始提出的概念模型相比，这一模型在原有同心圆的基础上，新增了一个外围圈层，用于呈现影响涉考群体对语言测试公平性看法的四大因素（即社会文化、教育机会、测试机构和考生个体因素）。

本研究在理论、实践及方法层面均具有重要启示。理论层面，本研究提出了一个测试公平性评估模型。该模型不仅凸显了测试公平性的双重属性，还通过呈现测试公平性的四个重要维度、核心涉考群体以及影响核心涉考群体对测试公平性看法的四大因素，彰显了测试公平性的多维性、社会建构性以及情境依赖性等重要特征。该模型为高风险语言测试公平性评估工作的顺利开展提供

了不可或缺的理论参考。实践层面，针对本研究聚焦的某高风险英语水平测试，实证研究发现能够为该测试的质量提升和相关测评实践提供重要参考。此外，研究结果还将助力校本语言测试公平性标准或指南的起草工作。方法层面，本研究通过实践探索，初步验证了将混合研究方法应用于语言水平测试公平性评估工作的可行性。混合研究方法有助于开展兼具全面性与系统性的语言测试公平性评估工作。

关键词：测试公平性评估；英语水平测试；测量学；涉考群体

Abstract

Fairness has emerged as an important dimension of test evaluation and one of the core values pursued in the field of language testing. The concept of test fairness possesses a dual nature, encompassing both measurement attribute and value judgment attribute. To date, research efforts have been dedicated to evaluating the fairness of high-stakes language tests in the two aspects. One stream of research has employed quantitative approaches to evaluating test fairness. The other has examined stakeholders' perceptions of test fairness using qualitative approaches. However, due to the multifaceted, socially-constructed, and contextually-situated nature of test fairness, a comprehensive and systematic evaluation of test fairness has proven to be a challenging endeavor, particularly in the absence of theoretical guidance.

This study, situated in the context of a high-stakes English Proficiency Test (EPT) administered in a Chinese university, aims to evaluate the fairness of the test from both psychometric and stakeholders' perspectives and propose a model of test fairness evaluation. A tentative conceptual model of test fairness evaluation is proposed based on the insights drawn from the dimensions of test fairness, existing fairness evaluation frameworks, and relevant empirical investigations. In response to the duality of test fairness, the tentative conceptual model features a two-layered structure arranged in concentric circles. The inner layer encompasses four dimensions—comparability, accessibility, consistency, and accountability—that serve as key focuses and essential criteria for test fairness evaluation. Surrounding these four dimensions is an outer layer comprising four key stakeholder groups: test takers, test developers, test administrators, and test users.

Informed by the tentative conceptual model, this study employed a convergent mixed-methods design to evaluate the fairness of the EPT. Quantitatively, (1) comparability of the test scores across test takers with humanities and science academic backgrounds was evaluated by performing differential item, bundle, and test functioning (DIF, DBF, and DTF) analyses; (2) comparability of the input materials across different test forms was evaluated by analyzing the characteristics of

the listening and reading input materials using both descriptive and inferential statistical methods; (3) test-takers' perceptions of the fairness of the EPT were evaluated using a self-designed questionnaire. The questionnaire, developed based on the four dimensions of test fairness outlined in the tentative conceptual model, was administered to 1,646 test takers. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed to explore and validate the factor structure of the questionnaire. Qualitatively, one-on-one semi-structured interviews were conducted with 20 test takers, six teachers, two administrators, and three test users to elicit their perceptions of the fairness of the EPT in terms of the four dimensions of test fairness. The interviews were transcribed and subjected to thematic analysis.

From the perspectives of psychometrics and stakeholders, quantitative and qualitative results indicate that the EPT is generally fair. The results also provide empirical support for the four dimensions of test fairness in the tentative conceptual model. Quantitatively: (1) Regarding the comparability of the test scores of test takers with humanities and science academic backgrounds: Among the four listening and reading test forms, 5% of the listening items exhibited DIF, 40% of the listening testlets and 22.22% of the reading testlets exhibited DBF, and two test forms exhibited DTF. Content analysis suggests that these statistically problematic DIF, DBF, and DTF do not introduce bias to the listening and reading subtests. The findings suggest that the test scores from listening and reading subtests are comparable across humanities and science test-taker groups. (2) In terms of the comparability of the input materials across different test forms: The listening input materials exhibited comparable levels of lexical complexity, syntactic complexity, discourse complexity, and speed of delivery. The reading materials showed slight variations in input characteristics (including syntactic complexity, discourse complexity, and readability) across the four test forms, with lexical complexity being the only exception. (3) As for the questionnaire survey: The EFA and CFA demonstrated a factor structure in the questionnaire that supported the dimensions of test fairness outlined in the tentative conceptual model. Test takers agreed that: (1) test-related information was transparent and accessible through various channels; (2)

the EPT was administered consistently in terms of test environment and procedures; (3) they had opportunities to fully demonstrate their English proficiency during the test; and (4) they could request a score review or voice any concerns about testing services after taking the EPT.

Qualitatively: (1) Regarding the stakeholders' perceptions of the fairness of the EPT: Thematic analysis identified four themes that aligned well with the four dimensions of test fairness in the tentative conceptual model. Interview data suggests that the four stakeholder groups believed that the EPT is fair in terms of comparability, accessibility, consistency, and accountability. (2) Regarding the factors influencing the stakeholders' perceived fairness of the EPT: Thematic analysis revealed four themes representing the sociocultural, educational, institutional, and personal factors that influence the stakeholders' perceptions of the fairness of the EPT. Sociocultural factors include the importance of English proficiency in China, and societal norms around English testing in China. Educational factors include disparities in pre-university English education, and adequacy and effectiveness of university learning resources. Institutional factors include institutional policy, test developers' expertise in language testing and assessment, and infrastructure for administering the EPT. Personal factors include test-takers' English proficiency, and test-takers' personal beliefs about the EPT.

Based on the tentative conceptual model and the aforementioned quantitative and qualitative results, a model of test fairness evaluation is proposed. Compared with the tentative conceptual model, the newly proposed model incorporates an additional layer of the identified factors that influence stakeholders' perceptions of the EPT's fairness (i.e., sociocultural, educational, institutional, and personal factors).

This study has theoretical, practical, and methodological implications. Theoretically, this study is one of the first attempts to propose a model of test fairness evaluation. The model takes the duality of test fairness into consideration. It also highlights the multifaceted, socially-constructed, and contextually-situated nature of test fairness by incorporating the dimensions of test fairness, key stakeholder groups, and factors influencing the stakeholders' fairness perceptions. The model offers

valuable theoretical guidance for the fairness evaluation of high-stakes language tests. Practically, the results from this study can be used to inform improvements to various aspects of the EPT and the associated testing practices, and facilitate the development of local fairness standards or guidelines. Methodologically, this study has demonstrated that a mixed-methods approach can help achieve a relatively comprehensive and systematic evaluation of the fairness of language tests.

Key words: test fairness evaluation; English proficiency tests; psychometrics; stakeholders

Acknowledgements

As I reflect on the journey that has morphed me into a problem solver, creative thinker, and a resilient researcher, my heart overflows with gratefulness toward those whose presence and companionship have illuminated my doctoral odyssey.

My sincerest appreciation goes to Prof. He Lianzhen, my supervisor, for her unwavering support at the nadir of my doctoral journey. Six years ago, the course *Language Testing*, taught by Prof. He, influenced me profoundly and ignited my passion to delve deeper into this field. As a connoisseur of language, Prof. He guided me through the intricacies of academic writing and provided constructive feedback on the early drafts of my dissertation. Without her perspective, perseverance, and patience, navigating the challenges of doctoral research would have been far more difficult. Prof. He helped me reach *a point of no return*, where my previous fears proved unfounded, and I became fully committed to my academic goals thereupon. The guidance she provided was an unwavering source of strength, inspiring me to push beyond boundaries and achieve new heights in my academic journey.

I am equally grateful to Prof. Min Shangchao, an inspiring figure at the Research Center for Language Development and Assessment, Zhejiang University. Her feedback on the theoretical model and quantitative analysis for this doctoral study was invaluable. I also want to thank Prof. Min for her uplifting words and encouraging smiles during our hotpot gatherings. Her academic and emotional support meant a great deal to me.

My heartfelt thanks also go to Prof. Yu Guoxing, my co-supervisor during my study at the School of Education, University of Bristol, for his guidance in the research design and writing of my doctoral dissertation. His critiques of my early ideas in the research proposal encouraged me to think outside the box and enhanced the originality of my work.

I extend my sincerest thanks to the Social Science Foundation of China, the China Scholarship Council, and the Xinmiao Talents Foundation for their generous support, which helped me access invaluable academic resources and greatly

facilitated participant recruitment and data collection. It is their belief in the significance of this project that made this dissertation possible.

My deepest gratitude goes to the faculty members of the Foreign Languages Teaching Center, Zhejiang University, for their tremendous help in the collection of questionnaire and interview data.

I am deeply indebted to the participants in this study, including the test takers, teachers, test administrators, and test users. Their openness in sharing their experiences and insights has greatly enriched the data presented in this dissertation.

I would also like to extend my sincere gratitude to my fellow colleagues in the language testing and assessment research team at the School of International Studies, Zhejiang University, including Dr. Xiong Lidi, Dr. Jiang Ziyun, Li Yue, Wang Ziyang, and Wang Zhaori for unstoppable laughs and uplifting talks. My sincere gratitude also goes to my colleagues at the School of Education, University of Bristol, especially Dr. Liu Sha, Dr. Sun Zhuoren, Dr. Qian Zhao, Dr. Wu Tongyu, Pu Pu, and Daniel Yu-Sheng Chang, for coffees and delightful post-prandial conversations.

I owe a world of gratitude to my high-energy bestie, Dr. Zhang Yi, who helped me realize that the pursuit of knowledge is not a herculean task, but rather an odyssey through the rich landscape of my life.

I am deeply grateful to my close friends in Bristol, especially Chen Leran, Lu Xinyue, and Yang Yilong, for our shared enjoyment of delicacies, Ping Pong, and beautiful guitar melodies. Our cooking and hiking adventures spiced up my life while contributing to my mental and physical well-being during my stay in the UK.

My heartfelt gratitude also goes to my parents and husband. Their unconditional love and unswerving belief in me sustained me through this particularly challenging period. Despite missing many cherished moments with them due to exhaustive efforts demanded by my research, their faith in me never wavered. Their unyielding support has been a beacon of hope amid many uncertainties along this academic journey.

Finally, I want to thank myself for picking myself up and dusting myself off every time I encountered difficulties and discouragements along this doctoral journey.

Contents

Figures

Tables

Abbreviations

AERA	American Educational Research Association
AUA	Assessment Use Argument
AWE	Automated Writing Evaluation
AWL	Academic Word List
BC	Banked cloze
C/AS	Clauses per AS-unit
C/S	Clauses per sentence
C/T	Clauses per T-unit
CET-4	College English Test Band 4
CET-6	College English Test Band 6
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CN/AS	Complex nominals per AS-unit
CN/C	Complex nominals per clause
CN/T	Complex nominals per T-unit
CP/C	Coordinate phrases per clause
CP/T	Coordinate phrases per T-unit
CSE	*China's Standards of English Language Ability*
CW/W	Proportion of content words to the total number of words
DBF	Differential bundle functioning
DC/C	Dependent clauses per clause
DC/T	Dependent clauses per T-unit
DET	Duolingo English Test
DIF	Differential item functioning
DTF	Differential test functioning
EFA	Exploratory factor analysis

EFL	English as a Foreign Language
EPT	English Proficiency Test
ESLPE	English as a Second Language Placement Exam
ETS	Educational Testing Service
FET	Fudan English Test
GPA	Grade Point Average
GSEEE	Graduate School Entrance English Examination
GSL	General Service List
IELTS	International English Language Testing System
IRT-LRT	Item-response-theory-likelihood-ratio test
JCTP	Joint Committee on Testing Practices
JLPT	Japanese-Language Proficiency Test
KMO	Kaiser-Meyer-Olkin
KSAs	Knowledge, skills, and abilities
L1	First language
L2	Second language
L2SCA	L2 Syntactical Complexity Analyzer
LC	Long conversation
LP	Listening passage
LR	Logistic regression
LSA	Latent Semantic Analysis
M/NP	Modifiers per noun phrase
MANOVA	Multivariate analysis of variance
MATTR	Moving average type-token ratio
MG-CFA	Multiple-group confirmatory factor analysis
MH	Mantel-Haenszel
MIMIC	Multiple indicators multiple causes
ML-AS	Mean length of AS-unit
MLC	Mean length of clause

MLR	Maximum Likelihood estimation with Robust standard errors
MLS	Mean length of sentence
MLT	Mean length of T-unit
MSRT	Ministry of Science, Research, and Technology
NMET	National Matriculation English Test
PCA	Principal Component Analysis
PRETCO	Practical English Test for Colleges
RMSEA	Root mean square error of approximation
RP	Reading passage
RUC-TOPE	Renmin University of China-Test of Oral Proficiency in English
SC	Short conversation
SIBTEST	Simultaneous Item Bias Test
SRMR	Standardized root mean square residual
TA	Test administrator
TD	Test developer
TestDaF	Test Deutsch als Fremdsprache
TFF	Test Fairness Framework
TT	Test taker
TU	Test user
TOEFL® iBT	Test of English as a Foreign Language™ Internet-based test
VETS	Vocational English Test System
WLSMV	Weighted Least Squares Mean and Variance adjusted
ZJU-EPT	Zhejiang University English Proficiency Test

Chapter 1 Introduction

1.1 Background of the study

This study focuses on the fairness of a high-stakes in-house English proficiency test (hereafter referred to as "EPT") administered at a comprehensive university located in eastern China. The EPT is locally developed to evaluate the undergraduates' ability to use English for general purposes across four language skills: listening, reading, writing, and speaking. It functions as an exit requirement in accordance with the university's regulations for a majority of undergraduate degree programs. In other words, the EPT is a high-stakes test used to make decisions which has life-changing impact on its intended test takers. Given the seriousness of the consequences associated with the EPT, the fairness of the EPT becomes a critical issue.

Fairness has recently grown to be an important dimension of test evaluation along with validity and reliability (American Educational Research Association [AERA] et al., 2014; Geisinger, 2015; Jonson & Geisinger, 2022). It has also been one of the core values pursued in the field of language testing and assessment (Nisbet & Shaw, 2020; Worrell, 2016). The inquiry of test fairness, however, has proven to be a challenging endeavor. This can be attributed in part to a lack of consensus on its definition (Cole & Zieky, 2001; Fischer et al., 2013; Gipps & Stobart, 2009; Kunnan, 2000; Zieky, 2006, 2015).

Test fairness has been defined either in a narrow or a broad sense. In a narrow sense, test fairness is defined as comparable validity for identifiable and relevant test-taker groups across all assessment stages (McArthur, 2018; Willingham, 1999; Xi, 2010). While in a broad sense, test fairness is perceived as an inherent part of social, cultural, historical, political, and philosophical fabric (Kunnan, 2013, 2018; Shaw & Imam, 2013; Stobart, 2005). Despite these efforts, universally agreed-upon definition of test fairness has yet to be established, in part owing to its multifaceted (Camilli & Newton, 2022; Opesemowo et al., 2023), socially-constructed (Camilli, 2006; Huggins-Manley et al., 2022; Jonson & Geisinger, 2022; Moss et al., 2005; Stobart,

2005; Tierney, 2014), and contextually-situated (Jang, 2002; Nisbet & Shaw, 2020) nature. Test fairness seems to emerge as a dual construct which is characterized by both objective and subjective aspects. The former suggests that test fairness is measurable and can be evaluated by collecting quantitative evidence, whereas the latter suggests that the evaluation of test fairness goes beyond a technical or psychometric issue as test fairness itself is subject to individual interpretations.

In response to the duality of test fairness, research efforts in language testing and assessment have been dedicated to evaluating the objective and subjective aspects of fairness. One stream of research has evaluated the objective aspects of test fairness from a psychometric perspective, operationalizing test fairness as absence of construct-irrelevant bias. At the very heart of this perspective lies the crucial aim of ensuring measurement comparability among test takers with diverse demographic characteristics (Ercikan, 2006; Nisbet & Shaw, 2019; Song, 2018) and across different cohorts of test takers who take different test forms [1] across test administrations. A prominent area of research focuses on the comparability of test scores at item, testlet, or test levels across gender (e.g., Ahmadi & Bazvand, 2016; Amirian et al., 2020; Aryadoust, 2018; Min & He, 2020; Pae, 2012; Takala & Kaftandjieva, 2000; Zhu & Aryadoust, 2019), age (Geranpayeh & Kunnan, 2007), grade (Liao & Yao, 2021), academic background (Chen & Zeng, 2021; Liu, 2011; Pae, 2004), first language (L1) background (Abbott, 2007; Kang et al., 2024), ethnical background (Uiterwijk & Vallen, 2005), and so forth. Previous empirical investigations have also examined the comparability of task difficulty across multiple test forms administered during a single test session (Liu & Zheng, 2022) or across test sessions (Shi, 2019). It should be noted that many high-stakes language tests employ multiple test forms in a single administration or across administrations to prevent cheating and ensure test security, making the comparability of task difficulty a major fairness concern (Jin & Wu, 2017). The evaluation of the objective aspects

[1] A "test form" is defined as "[a] specific version of a test" (Davies et al., 1999, p. 200). In the context of high-stakes language tests, it is common practice to administer a new test form for each administration for test security purposes.

of test fairness primarily employs quantitative methods and relies heavily on statistical analyses. Common analytical approaches include the detection of differential item, bundle, and test functioning (DIF, DBF, and DTF) and the identification of incomparable characteristics of the input materials used across different test forms (e.g., Liao, 2020). However, quantitative approaches tend to limit the scope of fairness research to test quality, neglecting the aspects preceding and following test administration (Klenowski, 2014). It is worth noting that test-takers' opportunities to access construct-relevant knowledge and skills, score interpretation, and score use consequences are also important aspects of test fairness that warrant research attention (Gipps & Stobart, 2009).

The other stream of research has evaluated the subjective aspects of test fairness by incorporating stakeholder perspectives (e.g., Butler et al., 2021; Deygers, 2017; Fox & Cheng, 2007; Heeren et al., 2021; Song, 2014, 2018; Wallace & Ng, 2023; Yao, 2023). "Stakeholders" refer to a variety of participants involved in assessment practices who assume different roles (Rea-Dickins, 1997). Typical stakeholders include but are not limited to test takers, teachers, test administrators, and test users. Research conducted so far has touched upon "felt fairness"—the "perceptions of fairness or unfairness by the people concerned" (Nisbet & Shaw, 2020, p. 7)—from the perspectives of test takers (Amirian et al., 2020; Fan, 2018; Heeren et al., 2021; Song, 2018; Sonnleitner & Kovacs, 2020; Tsai & Tsou, 2009; Wallace & Qin, 2021), teachers (Fan, 2018; Song, 2014, 2018), and test administrators (Malone & Montee, 2014; Song, 2014, 2018) as well. Much of the scrutiny along this stream of research has been placed upon fairness issues such as test information transparency (Bazvand & Rasooli, 2022), test-takers' opportunity to learn (Saito et al., 2022) and to take a test (Papageorgiou & Manna, 2021), test administration (Jang, 2002; Song, 2018), and test accommodations (Guzman-Orth et al., 2023; Motteram et al., 2023; Randez & Cornell, 2023). This stream of research acknowledges that test fairness, as a social construct, should be evaluated through a social process. As demonstrated by existing research, mainstream methods for evaluating the subjective aspects of test fairness include questionnaire survey (e.g., Fan, 2018; Hamid et al., 2019; Malone & Montee,

2014; Rasooli, 2021; Sonnleitner & Kovacs, 2020; Tsai & Tsou, 2009; Wallace & Qin, 2021) and interviews (e.g., Amirian et al., 2020; Butler et al., 2021; Fox & Cheng, 2007; Rasooli, 2021; Song, 2014, 2018). However, up until now, there has been a paucity of research adopting a multi-stakeholder approach to address the inherent subjectivity that arises from focusing solely on the perspective of a single stakeholder group. Furthermore, there is limited understanding of the factors that influence stakeholders' perceptions of test fairness.

Despite the growing body of research on test fairness, there has been remarkably little research that systematically evaluates its objective and subjective aspects using a mixed-methods approach. Such an approach is essential to avoid the predicament of "the blind men and the elephant" (Cheng & Sultana, 2022). By integrating quantitative and qualitative evidence, a relatively comprehensive and objective evaluation of test fairness is expected to be achieved.

To date, a plethora of studies have focused on the fairness of international high-stakes language tests. Cases in point are the Test of English as a Foreign Language™ Internet-based test (TOEFL® iBT), the International English Language Testing System (IELTS), and the Duolingo English Test (DET). Specifically, an argument-based framework has been proposed for the TOEFL® iBT to systematically guide fairness investigations (Xi, 2010). Empirical research articles and technical reports have been published to demonstrate the fairness of the above tests (Cardwell et al., 2023; Duolingo, 2021; Jin & Yan, 2017; Kang et al., 2024; Liu, 2011; Noori, 2022). Additionally, testing institutions or agencies have established and released fairness standards, guidelines, or principles for language tests administered across the world, exhibiting their dedication to advancing test fairness (Burstein, 2023; Educational Testing Service [ETS], 2013, 2014, 2016, 2022). However, there has been little research into the fairness of high-stakes local language tests (Fan et al., 2022). Local tests "represent the values and priorities within a local instructional program" and "address problems that emerge out of a need within the local context in which the test will be used" (Dimova et al., 2020, p. 1). Enriching the conceptual and empirical understanding of test fairness within local contexts is essential. This is because

connotations of and value judgments on test fairness vary from one context to another (Brown, 2008; Jang, 2002; Nisbet & Shaw, 2020; Song, 2018). Fairness evaluation of local language tests can provide insights into what fairness means in a local testing context, what stakeholders' perceptions of test fairness are, and what contextual factors shape their beliefs about test fairness.

Over the past two decades, a number of top-ranking institutions of higher learning in China have developed their in-house English proficiency tests. Each of these institutions adopts unique strategies for student development, with a key focus on students' English language ability. Tailor-made English language tests have been developed and administered in those institutions to address local instructional and assessment needs. For example, Zhejiang University strives to nurture "future leaders and useful citizens with a global vision and social responsibility" (Zhejiang University, n.d.). In pursuit of this mission, the university developed Zhejiang University English Proficiency Test (ZJU-EPT), aiming to bring about a positive washback on English teaching and learning. Similarly, Fudan English Test (FET) was developed in accordance with the student development objectives of Fudan University. Renmin University of China developed Renmin University of China—Test of Oral Proficiency in English (RUC-TOPE) to ensure that its undergraduates possess a certain level of oral communication ability as specified by internal standards for oral English language ability upon graduation. It should be noted that several local language tests in China (e.g., the ZJU-EPT, the FET, and the RUC-TOPE) are used as exit requirement for undergraduate students. Students who do not pass these tests will not obtain their Bachelor's degrees. Despite their high-stakes nature, the fairness of these local English tests has not yet been systematically evaluated. Overlooking the fairness of high-stakes local language tests can lead to serious consequences, such as jeopardizing the test-takers' future educational and occupational opportunities and compromising the tests' credibility and acceptability.

The EPT examined in this study is a representative of the locally-developed high-stakes language tests in China. The objective and subjective aspects of the fairness of the EPT have not been systematically evaluated from a combination of

psychometric and stakeholders' perspectives. In terms of the objective aspects, previous research has touched upon the comparability of writing and speaking prompts used across different test forms (Lv, 2018; Shi, 2019). However, two major fairness concerns still loom over the listening and reading subtests of the EPT. The first concern is related to the comparability of test scores among test takers with different academic backgrounds. The intended test takers of the EPT are undergraduates from different academic disciplines in the university. Given that test scores from the EPT are used to inform graduation decisions that carry substantial weight for the test takers, empirical research is in pressing need to identify flawed items, testlets, or even subtests that might disadvantage a test-taker group with a certain academic background. The other fairness concern lies in the comparability of listening and reading input materials used in different test forms. To ensure test security, the EPT employs multiple test forms for different test sessions in a single administration. Nevertheless, it remains uncertain whether the input materials are of comparable difficulty levels across test forms. The lack of comparability in input materials could result in differing levels of task difficulty and (dis)advantages for a cohort of test takers in a particular test session. Meanwhile, the subjective aspects of the fairness of the EPT have been overlooked since its launch in 2013. Over the years, the stakeholders' expectations and concerns regarding the EPT's fairness have remained unexplored. Understanding the perceived fairness of the EPT and the factors influencing their perceptions can contribute significantly to shaping the EPT into a culturally-responsive and ethically-accountable local language test.

1.2 Aims and research questions of the study

1.2.1 Research aims

To address the research gaps and needs identified in Section 1.1, this study aims to evaluate the fairness of the EPT from the perspectives of both psychometrics and stakeholders and to propose a model of test fairness evaluation.

1.2.2 Research questions

Two overarching research questions and four sub-questions this study attempts to address are listed below:

- RQ1: To what extent is the EPT fair from a psychometric perspective?
    - RQ1.1: To what extent are the test scores comparable across test-taker groups with different academic backgrounds, as indicated by differential item functioning, differential testlet functioning, and differential test functioning?
    - RQ1.2: To what extent are the input materials comparable in terms of difficulty across different test forms, as indicated by input characteristics?
- RQ2: To what extent is the EPT fair from a multi-stakeholder perspective?
    - RQ2.1: How do stakeholders perceive the fairness of the EPT?
    - RQ2.2: What are the underlying factors, if any, that might influence the stakeholders' perceptions of test fairness?

1.3 Significance of the study

The study is expected to offer theoretical, practical, and methodological insights into the field of language testing.

In terms of theoretical significance, the researcher first proposed a tentative conceptual model of test fairness evaluation based on relevant theoretical frameworks and empirical findings from previous studies. Drawing on the tentative conceptual model and the empirical findings of this study, a model of test fairness evaluation is proposed toward the end of this dissertation (see Section 5.2). The final model is hoped to illustrate: (1) the dimensions or research scopes of test fairness and (2) the underlying factors influencing the stakeholders' perceptions of test fairness. The model is expected to guide a comprehensive and systematic evaluation of the fairness of language tests.

In terms of practical aspects, findings from this study would help shed light on fair practices at various assessment stages, ranging from test design and development,

test administration, scoring and score reporting, to score interpretation and use. The results from score and input material comparability analyses could help identify flawed input materials, items, testlets, and even subtests of the EPT. These results are expected to foster fairer practices in test development, including the training of item writers. Item writers will become more aware of and sensitive to potentially biased items and inappropriate input materials used to measure test-takers' receptive language skills. Furthermore, accounts from the stakeholders can be used to inform fair practices across all assessment stages. The interviews conducted in this study seek to capture stakeholders' perceptions of test fairness prior to, during, and subsequent to test administration. Throughout the interviews, the stakeholders evaluated the fairness of the EPT and its relevant practices, offer detailed reasoning, while also expressing their expectations and concerns about test fairness. It is hoped that their perspectives will facilitate: (1) beneficial modifications to current assessment practices at the local institution, (2) the development of context-specific fairness standards or guidelines, and (3) fair development and use of local, national, and international English proficiency tests.

Methodologically, this mixed-methods study approached test fairness by collecting, analyzing, and integrating both quantitative and qualitative evidence. It demonstrated how to evaluate test fairness in a systematic and comprehensive manner. Quantitatively, the fairness of the EPT was evaluated through comparability analyses of test scores and input materials. Score comparability across test takers with different academic backgrounds was subjected to DIF, DBF, and DTF analyses. In addition, characteristics of the input materials from different test forms were compared following a tailored analysis scheme. Meanwhile, a questionnaire, developed based on the tentative conceptual model of test fairness evaluation, was administered to explore test-takers' perceptions of the fairness of the EPT. The analysis scheme and the questionnaire helped contribute to the evaluation of test fairness. Qualitatively, stakeholders' perceptions on the fairness of the EPT were collected through individual interviews with representatives from each stakeholder group. The involvement of different stakeholders facilitated effective communication among the

test takers, teachers (also test developers in this study), test administrators, and test users in the institution. Their accounts contributed greatly to a comprehensive evaluation of the fairness of the EPT.

1.4 Organization of the dissertation

This dissertation consists of six chapters. Chapter 1, an introductory chapter, gives an overview of the background, aims, research questions, and significance of this study. Following a delineation of basic senses of fairness and the dimensions of test fairness, Chapter 2 reviews the theoretical think-pieces and empirical research on test fairness. The chapter ends by proposing a tentative conceptual model of test fairness evaluation. Chapter 3 lays out the research design, presenting detailed information on the participants, instruments, data collection procedures, and data analysis approaches. In Chapter 4, the findings related to the research questions are presented. Chapter 5 begins with a discussion of the objective and subjective aspects of the fairness of the EPT. Based on a discussion of empirical evidence collected in this study, a model of test fairness evaluation is presented in Section 5.2. Chapter 6 concludes the dissertation by summarizing key findings, discussing implications, pointing out the limitations of the study, and suggesting avenues for further research.

Chapter 2 Literature Review

This chapter reviews existing think pieces on test fairness and relevant empirical studies regarding the fairness of language tests. Section 2.1 reviews the basic senses of fairness and the dimensions of test fairness. Section 2.2 presents frameworks and empirical efforts devoted to evaluating test fairness. This section includes a scrutiny of theoretical frameworks of test fairness evaluation, followed by an overview of empirical investigations into both the measurement and value judgement attributes of test fairness. From the perspective of psychometrics, empirical studies on the comparability of test scores across test-taker groups and the comparability of input materials across test forms will be reviewed. From the stakeholders' perspective, empirical efforts devoted to their perceptions of test fairness and the factors influencing their perceptions will be reviewed. Building on the reviewed literature, Section 2.3 proposes a tentative conceptual model of test fairness evaluation. Lastly, Section 2.4 summarizes the key points covered in this chapter.

2.1 Introducing test fairness

2.1.1 Basic senses of fairness

Examining the etymology of "fair" is essential before defining "test fairness". The meanings of "fair" in Old English had something to do with appearance, weather, and morality (Harper, n.d.). In terms of appearance, "fair" meant "pleasing to the sight (of persons and body features, also of objects, places, *etc.*)", "beautiful", "handsome", and "attractive". In reference to weather, "fair" was defined as "bright, clear, and pleasant", signifying the absence of rain. In the moral sense, "fair" in late Old English conveyed the idea of being "morally good". By the early 13th century, "fair" had evolved to mean "according with propriety" and "according with justice". By the mid-14th century, the definition of "fair" had further expanded to include meanings such as "equitable", "impartial", "just", and "free from bias". In the 19th century, "fair" became associated with moral values in sporting contexts (Harper, n.d.) and

behavioral standards in competitions (Tierney, 2017). As an illustration, a “fair play” implies that all participants should adhere to the rules of a competition.

Shifting away from the earlier emphasis on appearance and weather in Old English, the definitions of “fair” provided by modern dictionaries foreground its connotations associated with moral values. The first definition of “fair”, according to the online Merriam-Webster Dictionary (Merriam-Webster, n.d.), is “marked by impartiality and honesty: free from self-interest, prejudice, or favoritism”. Its second definition is “conforming with the established rules” and “consonant with merit or importance”. “Fair” in the Cambridge English Dictionary (Cambridge University Press, n.d.) is defined as “treating someone in a way that is right or reasonable, or treating a group of people equally and not allowing personal opinions to influence your judgment”. In addition, something can be described as fair if “it is reasonable and is what you expect or deserve”. Also, a fair game or competition “is done according to the rules”. Collins English Dictionary (n.d.) defines “fair” as being “reasonable, right, and just”.

The definitions of “fair” in the aforementioned dictionaries provide a basis for understanding its connotations in the field of educational measurement. Seven senses of fairness are closely connected to assessment. First, “fair” is associated with “formal”, which means the correct and proper application of a rule. Rule-breaking behaviors will result in formal warnings or disciplinary actions. The second sense of fairness can be referred to as an “implied contractual sense”. Should something go against stakeholders’ expectations on a test, they might claim that the test is unfair. Another sense can be labeled as a “relational sense”, denoting “treating (relevantly) like cases alike” (Nisbet & Shaw, 2020, p. 4). In this sense, it is fair to discriminate test takers based on the intended construct and nothing else. The fourth sense of “fair” can be described as “retributive”. This sense is grounded in the concept of “desert” (i.e., the state of deserving reward or punishment) which suggests that assessment outcomes should match an individual’s merits, abilities, or efforts. The fifth sense of fairness is related to the consequential effect of an assessment. Unethical use of assessment outcomes is unfair in this regard. The sixth sense of fairness, known as

the "retrospective" sense, suggests that an assessment can be labeled as unfair if its outcome is influenced by past social actions or policies (e.g., racial segregation). The last sense denotes "respect", referring to the "equality of esteem" (Nisbet & Shaw, 2020, p. 5). Disrespectful content or language embedded within an assessment could be perceived as unfair by an individual test taker or test-taker subgroups. The basic senses of "fair" offer valuable insights into the definition of test fairness.

To clarify, presenting a clear-cut definition of "test fairness"—an elastic concept—is beyond the capability of the author. This is because "test fairness" can accommodate different conceptualizations and interpretations across disciplines and across sociocultural contexts. In the next section, the author introduces four key dimensions of test fairness in the field of language testing.

2.1.2 Dimensions of test fairness

2.1.2.1 Comparability

Test fairness can be understood in terms of "comparability". Nisbet and Shaw (2020) highlighted the close link between fairness and comparability, noting that they are "intertwined and interconnected" (p. 28) in most cases. The relational sense of "fair" (i.e., "treating like cases alike"), as mentioned at the beginning of Section 2.1.1, is typically discussed in scenarios where two similar cases are compared. "Comparability" is one of the representative fairness dimensions that captures the relational sense of test fairness. "Comparability" can be further operationalized as the "absence of measurement bias" and an aspect of validity—"comparable validity". The former represents a negative approach to defining test fairness, acknowledging the difficulty of achieving a perfect and comprehensive definition for "test fairness", whereas the latter adopts a proactive approach by providing a clear definition of what constitutes fairness by resorting to established concept in the field of language testing.

Central to the "absence of measurement bias" is the comparability in the construct being measured across test-taker groups. "Construct" is a psychometric term used to describe the knowledge, skills, abilities or other attributes (KSAs) that a test purports to measure (Walters, 2022; Zieky, 2015). "Bias", a technical and

psychometric term in educational measurement, refers to the advantage or disadvantage an item, a task, or a test offers to a certain group of test takers due to deficiencies in a measurement instrument (AERA et al., 1999, 2014; Deygers, 2019; Jang, 2002; McNamara & Ryan, 2011; Stobart, 2005). Overall, "absence of measurement bias" denotes a sense of impartiality—one of the essential properties of being "fair" according to the dictionaries mentioned above (e.g., the online Merriam-Webster Dictionary). To put it simply, test takers with the same level of the KSAs intended to be measured by an unbiased test should receive the same score (Opesemowo et al., 2023). A fair language test is thus the one that only assesses the test-takers' language competence as specified by the test specifications without being influenced by the construct-irrelevant attributes of the test takers associated with their gender, academic background, cultural background, and so forth (Deygers, 2019; Sabbaghan & Fazel, 2023). Measurement bias undermines test fairness by leading to different interpretations of scores for test takers from different subgroups (AERA et al., 2014; Shepard, 1987). That is why bias is also known as "invalidity" (Shepard, 1987, p. 179) or "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (Cole & Moss, 1989, p. 205). The next paragraph will explain how comparability is related to validity issues.

Comparability is closely linked to validity, especially "comparable validity" (Huggins-Manley et al., 2022; Willingham, 1999; Willingham & Cole, 1997; Xi, 2010). According to Messick's unitary view, validity is "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1989, p. 20). Under this view, construct validity, relevance and utility of test use, value implications, and the social consequences of testing practices are integral to the considerations of test validity. Based on the concept of validity, test fairness is defined as "the extent to which inferences and actions based on test scores are equally valid for a diverse population of test takers" (ETS, 2022, p. 12). Willingham and Cole (1997) further conceived test fairness as ensuring "comparability" or "comparable validity" (p. 6) for individuals or groups of test

takers. To ensure test fairness, they advocated for comparable validity across all assessment stages (i.e., test design, test development, test administration, and test use) and outlined three criteria in deciding what is a fair test. A test is considered fair if it ensures: (1) comparable opportunity for test takers to show their knowledge and skills, (2) comparable test tasks and scores, and (3) comparable treatment of test takers in test interpretation and use (Willingham & Cole, 1997, p. 11). They further reiterated that "[f]airness at one stage of assessment does not guarantee fairness at another" (Willingham & Cole, 1997, p. 8). Drawing on Willingham and Cole's (1997) conceptual framework, Xi (2010) defined test fairness as "comparable validity for identifiable and relevant groups across all stages of assessment" (p. 154). According to her definition, it is crucial that "construct-irrelevant factors, construct under-representation, inconsistent test administration practices, inappropriate decision-making procedures or use of test results have no systematic and appreciable effects on test scores, test score interpretations, score-based decisions and consequences for relevant groups of examinees (p.154)". The idea of framing test fairness as comparable validity is embraced by technical professionals in language testing community. Ensuring comparability in measurement is thus essential for both validity and fairness evaluation (Ercikan, 2006).

2.1.2.2 Accessibility

"Test fairness" has been viewed as "accessibility"—an issue typically considered preceding test administration. "Accessibility" covers a wide range of considerations. First, it pertains to test-takers' learning opportunities, denoting that test takers have equal rights to access quality instructional resources and opportunities to learn the content, knowledge, and skills assessed by a test (AERA et al., 2014; Camilli, 2006; Christensen et al., 2023; Gipps & Stobart, 2009; Kunnan, 2000, 2018; McNamara & Roever, 2006; Rasooli et al., 2018; Solano-Flores, 2019). Ensuring test-takers' equal opportunity to learn is deemed a fair practice in that differential educational access, to some extent, has a role to play in test performance (AERA et., 2014; Kunnan, 2000; Rasooli et al., 2018). Those with limited access to learning resources would be greatly

disadvantaged in acquiring the requisite knowledge and skills that a test is intended to measure (Gipps & Stobart, 2009). Test takers with different cultural backgrounds might have differential exposure to the content or knowledge targeted by a test (Berman et al., 2020; Solano-Flores, 2019). As a result, fair and valid score interpretation would be compromised.

Test-taking opportunity is also one of the important considerations of "accessibility". In a fair testing scenario, test takers should be provided with "multiple, varied, equitable, and meaningful opportunities to demonstrate their learning" (Tierney, 2017, p. 798). "Multiple" test-taking opportunities ensure that adequate information can be gathered to inform decision-making. Using "varied" task types can effectively avoid test method effect while eliciting information about the test-takers' abilities. "Equitable" language testing endeavors to provide opportunities for test takers to exhibit their proficiency in the construct(s) being measured. According to the *Standards for Educational and Psychological Testing* (hereafter referred to as the 2014 *Standards*; AERA et al., 2014), "accessible testing" is defined as enabling "all test takers in the intended population, to the extent feasible, to show their status on the target construct(s) without being unduly advantaged or disadvantaged by individual characteristics (e.g., characteristics related to age, disability, race/ethnicity, gender, or language) that are irrelevant to the construct(s) the test is intended to measure" (p. 52). "Accessibility" also stands as a legal requirement. For instance, to enhance access to the target construct, accommodations are provided for test takers with disabilities. This could include providing individuals with visual impairments the braille or magnified version of the test paper (China Disabled Persons' Federation, Ministry of Education of the People's Republic of China, 2017). Finally, "meaningful" opportunities to demonstrate learning denotes a sense of respect for test-takers' varying learning styles and the knowledge they have acquired. Ensuring equal opportunities for all test takers to demonstrate their best performance is at the very heart of developing "high-quality and fair tests" (Carlsen & Rocca, 2022, p. 606).

The opportunity to access test-related information is also a key aspect of accessibility. Test takers should be provided with transparent information about

assessment content and procedures prior to test-taking (Bachman & Palmer, 2010/2016). Additionally, evaluation criteria and the scoring process should be made as accessible as possible to test takers (Camilli, 2006; Tierney, 2017).

The dimension of "accessibility" also includes considerations such as the affordability of the test, the selection of the test locations, and the opportunities for test takers to familiarize themselves with test environment and equipment, particularly in the case of computer-based tests (Bachman & Palmer, 2010/2016; Berman et al., 2020; Kunnan, 2000).

2.1.2.3 Consistency

"Test fairness" is closely associated with "consistency". "Consistency" is recognized as one of the six prerequisites of procedural fairness in non-legal settings, alongside bias suppression, accuracy of information, correctability, representation, and ethicality (see Leventhal, 1980). Procedural fairness, a necessary but not sufficient requirement for test fairness (Nisbet & Shaw, 2020), concerns the fairness of the procedures used to reach a decision or outcome. In the context of language testing, procedural fairness requires that "all test takers be treated in essentially the same way, that they take the same test or equivalent tests, under the same conditions or equivalent conditions, and that their performances be evaluated using the same (or essentially the same) rules and procedures" (Kane 2010, p. 178). Procedural fairness could be achieved if test administration is perceived as unbiased toward any test-taker group (Wallace & Qin, 2021). Inconsistencies in administration can lead to bias in assessment (Sabbaghan & Fazel, 2023). Therefore, "consistency" represents a "level playing field" for test takers (Kane, 2013, p. 39), which emphasizes the uniformity and impartiality of assessment procedures before, during, and after test administration (Kunnan, 2004; Wollack & Case, 2016).

Prior to administering a test, test developers should ensure its psychometric quality—linchpin of maintaining "procedural equality for individual and subgroups of test takers" (McNamara & Ryan, 2011, p. 163). According to Kunnan (2020), an assessment should maintain consistency across multiple tasks and forms. This

consistency can be achieved through practices such as training item writers, conducting multiple rounds of item review, pilot/field testing, and so forth. These practices in test design and development ensure that assessment tasks provide an accurate and adequate representation of the construct being measured, thereby avoiding favoritism toward any individual or group of test takers.

"Consistency" during test administration is equally important to advance test fairness. Providing uniform test environment for test takers is a key aspect of maintaining consistency, as this practice enables test takers to fully demonstrate the measured abilities during test-taking (AERA et al., 2014; Cohen & Wollack, 2006; Dorans et al., 2022). For example, factors such as unfamiliar or uncomfortable test environment, hardware and software malfunctioning, and security breaches can inadvertently threaten the "equivalence of testing conditions" (Ercikan & Lyons-Thomas, 2013, p. 211) or "administrative reliability" (Petour, 2015, p. 38). Therefore, inconsistent testing conditions or administrations would adversely influence test-taking experience and performance of test takers, jeopardizing the validity of score interpretation and use (Stobart, 2009). In contrast, uniform test environment and consistent administration procedures can nurture greater trust in test results by test users (Wollack & Case, 2016).

Post-administration consistency includes considerations of the uniformity of scoring, score reporting, and score interpretation practices. First, consistent scoring rests with clearly articulated rating criteria (Rasooli et al., 2018) and acceptable procedures for generating test results for all test takers (Bachman & Palmer, 2010/2016; Dorans, 2012). Second, rater reliability (including inter-rater and intra-rater reliability) is pivotal for ensuring fair treatment of test takers throughout the scoring process (AERA et al., 2014; Deygers, 2019; Sabbaghan & Fazel, 2023). Intra-rater reliability is used to describe a rater's consistency in assigning the same score to the same script across multiple occasions, and inter-rater reliability refers to the agreement among different raters in assigning scores to the same script (Weigle, 2002). It is worth mentioning that rater reliability can be influenced by various factors, including but not limited to rater training, experience, rater severity, and raters'

interpretation of the rating scale (e.g., Barkaoui, 2010; Davies, 2016; Lumley, 2002; McNamara & Roever, 2006; Neittaanmäki & Lamprianou, 2024). Failing to consider these factors could potentially result in rating bias. Third, consistency in score interpretation is contingent upon sound standard-setting procedures (Kunnan, 2020). Standard-setting refers to systematic procedures used to set benchmarks for test performance, assisting in making defensible pass/fail decisions and judgments on test-takers' language proficiency levels.

2.1.2.4 Accountability

"Test fairness" is also linked to accountability. Being accountable means taking responsibility for one's actions or decision-making and providing explanations or justifications when necessary. With the professionalization of the field of language testing, the demand for accountable assessments has become more pronounced. This raises two fundamental questions: who should be accountable for test fairness, and what mechanisms should be in place to ensure accountability?

Regarding the first question, existing studies have discussed the role of different stakeholder groups in achieving test fairness. These stakeholders include academic institutions, testing institutions or agencies, testing professionals (e.g., test designers, test developers, and examiners), test users, and test takers. The main responsibility of academic institutions and universities lies in formulating and, if necessary, revising decision-making policies to make sure that they are "ethically and socially just" (Sabbaghan & Fazel, 2023, p. 176).

Testing institutions or agencies roll multiple roles into one. They are expected to be held accountable across all assessment stages, including test design, item writing, test assembly, administration, rating, post-administration performance data analysis, and the evaluation of test consequences and impact (Fan, 2014). At the stage of test design, they should ensure that the test aligns with the latest theories of language competence, that the knowledge and skills assessed by the test align with those in the target language use domain, and that all test takers are provided with equal opportunities to demonstrate their language proficiency levels. In terms of item

writing or test assembly, testing institutions or agencies are expected to ensure the impartiality of both test content and format. Regarding test administration, they should ensure test-takers' access to relevant information about the test prior to test-taking (e.g., test preparation materials, test structure, and time allotment; see Karami, 2013), safeguard test security during administration, and treat test takers with disabilities fairly throughout test administration. It is also the responsibility of testing institutions or agencies to ensure scoring consistency by developing rating criteria that are theoretically and empirically tested, by training raters, by introducing multiple ratings, and by monitoring the quality of ratings (see Sabbaghan & Fazel, 2023). Additionally, after collecting test performance data, testing institutions or agencies should conduct DIF analysis and examine whether the test structure is as expected and comparable across different groups of test takers. Lastly, they are also expected to conduct research on differential washback, consequences, and impact across test-taker groups.

Testing professionals, often affiliated with testing institutions or agencies, have responsibilities that closely align with those of the institutions or agencies. Test designers should strive to incorporate ethical principles into their design practices. According to Weideman (2017), irresponsible designers tend to make design choices that are "illogical, invalid, not transparent, uninterpretable, not useful, unaccounted for, uncaring and uncompassionate" (p. 10). Test developers, a key group of testing professionals, should be held accountable for the psychometric quality of an assessment by minimizing item bias and any flaws embedded in test design (e.g., Gujord, 2023; Sabbaghan & Fazel, 2023; Spaan, 2000; Weideman, 2017). Moreover, they should ensure transparency in test-related information, including but not limited to the nature of the test, test content, test method, test difficulty level, test-takers' rights and responsibilities, the appropriate use of scores, procedures for resolving challenges to scores, and detailed information about test administration and scoring (Joint Committee on Testing Practices [JCTP], 2004; Spaan, 2000). Other responsibilities of test developers include maintaining standardized test delivery, accurately reporting and interpreting test scores, and collecting feedback from both

test users and test takers (JCTP, 2004; Spaan, 2000). Nevertheless, opinions are divided on whether test developers should be held accountable for test use and the subsequent consequences. Some argue that test developers should take a more proactive stance in preventing test misuse. They believe that test developers should assist in interpreting and using test scores by providing easy-to-understand score interpretations and test use guidelines for test users (Carlsen & Rocca, 2022; JCTP, 2004; Spaan, 2000). Opponents argue for limited responsibility of language developers in test use. Karami (2013), for example, suggested that holding test developers accountable for all consequences of test misuse or abuse is impossible, particularly in cases where test developers and users are independent groups. Test developers aside, raters should assign scores strictly following rating criteria instead of their intuition (Sabbaghan & Fazel, 2023).

The responsibilities of test users lie in test selection, score interpretation, and test use. First of all, test users play a key role in selecting tests that are appropriate for the test takers and that can satisfy the intended test purpose (Fan, 2014). In addition, test users should accurately interpret the meaning of test scores by referring to relevant empirical evidence or interpretation guidelines provided by test developers (JCTP, 2004). Test users are also responsible for ensuring that tests are not misused or abused (Spaan, 2000). They are expected to follow the test use guidelines provided by the developers, make informed decisions based on accurate test interpretation, and ensure consistency in test use (JCTP, 2004). Test users should also pay attention to the consequences and impact of test use on different test-taker groups (Fan, 2014; Nisbet & Shaw, 2020).

Test takers, in Spaan's (2000) tripartite "social contract" scheme (the other two parties being test developers and test users), should proactively familiarize themselves with the test content and methods before taking a test. On top of that, they should engage actively in test preparation and equip themselves with the knowledge and skills assessed by the test (Fan, 2014). During test administration, test takers should follow the regulations set by testing institutions or agencies and must not

engage in any form of cheating. They should also report any observed misconduct that could jeopardize test fairness.

Having discussed who should be held accountable for test fairness, the following paragraphs will focus on the mechanisms contributing to an accountable assessment. These mechanisms include sensitivity reviews, test evaluation, stakeholder engagement, the development of professional standards, and legal protections, to name a few.

Sensitivity reviews are usually conducted by testing institutions or test developers to remove any construct-irrelevant bias embedded in test content (e.g., text and visual representations) that may offend certain test-taker groups or provoke strong negative emotions among them (AERA et al., 2014). For example, taboo topics for language tests listed by *ETS Guidelines for Developing Fair Tests and Communications* (ETS, 2022) include slavery, religion, stereotypes, violence, and so forth. Ideally, the members in the review board should remain internationally and culturally diverse (Deygers, 2019). Though test items are typically reviewed after they are written (Sireci & Gándara, 2016), maintaining sensitivity to fairness concerns during item-writing process is also essential (Zieky, 2015).

Fairness has been an important dimension of test evaluation. Test evaluation can be conducted on a variety of aspects by collecting both quantitative and qualitative data (Dimova et al., 2020). For instance, test performance data in either pilot or operational stages could be used to examine the quality of a test (Geisinger, 2015; Sabbaghan & Fazel, 2023). The appropriateness of test purpose and test use consequences could be evaluated by interviewing stakeholders. However, there is no consensus on who should lead or conduct these evaluations. One solution for testing institutions or agencies is to invite external experts to evaluate the tests they develop. This move has been seen as “an important step towards accountability” (Karami, 2013, p. 165), especially as the stakeholders and public increasingly advocate for transparency and openness in testing practices (Boyd & Davies, 2002; Davies, 1997; Weideman, 2017).

Additionally, the stakeholder consultation has gained growing attention. It can take various forms, such as allowing test takers to request a score review after taking a test and collecting feedback from test takers about their experience through questionnaire surveys or interviews, *etc*. There is evidence that when relevant stakeholders are given an opportunity to express their opinions, they are more likely to perceive the test and decision-making procedures as fair (Bøggild, 2016; De Cremer et al., 2008). Stakeholders' feedback can be valuable in improving testing practices (Sonnleitner & Kovacs, 2020).

There are alternative mechanisms that can be employed to ensure accountability in assessment. One possibility is to develop professional standards or guidelines for testing practices (Fan, 2014; Yang & Gui, 2007). Once these standards are in place to guide the professionalization of the language testing field, test takers who have been unfairly treated in an assessment could seek legal recourse to hold testing agencies accountable (Karami, 2013).

2.2 Evaluating test fairness

2.2.1 Frameworks of fairness evaluation

A number of studies have proposed theoretical frameworks to guide the fairness evaluation of language tests. Three types of frameworks emerge in the existing literature, differing in the conceptualization of test fairness and approaches to evaluating it.

The first type of framework focuses on the psychometric qualities of language tests. Technically speaking, test fairness is not given primacy in its own dedicated evaluation framework; rather, it is considered an important aspect of validity. Evaluation of test fairness could help justify the real-world use of language tests. For example, test fairness evaluation could be guided by the general claims in the Assessment Use Argument (AUA; Bachman & Palmer, 2010/2016). The AUA, as essentially a justification process for the use of language tests, is intended to: (1) "[guide] the development and use of a given language assessment and provides the basis for quality control throughout the entire process of assessment development"

and (2) "[provide] the basis for test developers and decision makers to be held accountable to those who will be affected by the use of the assessment and the decisions that are made" (Bachman & Palmer, 2010/2016, p. 97). There are four types of *a priori* claims in the AUA. "Claim", a term in Toulmin's (1958/2003) model of argumentation, refers to a statement that asserts a specific position or viewpoint. Evidence needs to be collected to substantiate, compromise, or refute a claim. Specifically, Claim 1, Claim 2, and Claim 3 of the AUA touch upon test fairness in an implicit way. Claim 1 states that: "[the] *consequences* of using an assessment and of the decisions that are made are **beneficial** to stakeholders" (Bachman & Palmer, 2010/2016, p. 119; italics and bold in original). This claim underscores the importance of responsibly using assessment results to yield beneficence for test takers, education systems, and society as a whole. Claim 2 requires that: (1) score interpretation and decision making should be value sensitive and respectful of the basic rights of test takers and (2) decision making should not favor any test-taker group based on their demographic backgrounds or their opportunities to learn the assessed materials. Impartiality in Claim 3 refers to "the degree to which the format and content of the assessment tasks and all aspects of the administration of the assessment are free from bias that may favor or disfavor some test takers" (Bachman & Palmer, 2010/2016, p. 119). The AUA addresses key considerations within the scope of test fairness, including test-takers' opportunity to learn, test bias, and the consequences of test use, while emphasizing the accountability of both test developers and users. Despite the effectiveness of the AUA in guiding test development and validation, it is not a fairness-centered evaluation framework. Given that fairness is a foundational concern in language testing, the AUA may not serve as an ideal framework for evaluating test fairness.

The other case in point is Xi's (2010) proposal to establish fairness arguments in validity arguments based on Chapelle et al.'s (2008) validation framework. A validity argument, consisting of a chain of inferences, can guide test users toward making appropriate interpretations of test scores and decisions. Xi endorsed Willingham and Cole's (1997) view of conceptualizing test fairness as "comparable

validity for *identifiable* and *relevant* groups across all stages of assessment, from assessment conceptualization to use of assessment results" (Xi, 2020, p. 154; italics in original). Further, she put forward the criteria for test fairness which states that "construct-irrelevant factors, construct under-representation, inconsistent test administration practices, inappropriate decision-making procedures or use of test results have no *systematic* and *appreciable* effects on test scores, test score interpretations, score-based decisions and consequences for *relevant* groups of examinees" (Xi, 2020, p. 154; italics in original). To operationalize fairness evaluation, Xi made an attempt to embed fairness arguments within the established validity arguments. Specifically, rebuttals (i.e., counterclaims) to test fairness can be seen as rebuttals to comparable validity. Therefore, failures to repudiate fairness-related rebuttals might further weaken a validity argument. Following Xi's approach, fairness evaluation could be charted by collecting evidence that may support or refute fairness rebuttals. Weaknesses in test fairness also put at risk the validity for test score interpretation and use. It should be noted that probing into all aspects of fairness would never bring an end to fairness evaluation. By building fairness arguments in validity arguments, Xi's framework enables priority setting in fairness evaluation. It is worth mentioning that research priority of test fairness should be placed on identifying fairness threats with the most serious consequences (Willingham, 1999) and on collecting evidence to examine fairness-related rebuttals that may jeopardize the subsequent inferential links and finally compromise validity. Despite the strong practicality of Xi's approach, which can be used by testing professionals or researchers to systematically address issues of fairness and validity simultaneously, the approach has several limitations. First, Xi's framework does not do justice to fairness. By conceptualizing fairness as comparable validity (i.e., a facet of validity), the research scope of fairness is limited to construct comparability across test-taker subgroups in every inference chain of validity argument. In particular, considerations of test fairness begin with "domain description" in the inference chain, ignoring the fact that fairness issues can emerge prior to test development (Stobart, 2005). This is understandable because the inference chain for validity arguments starts with

"domain definition", and fairness issues can only be addressed within a validity argument in the form of fairness-related rebuttals. Second, relegating fairness as a facet of validity runs the risk of making it inaccessible to a wider group of stakeholders (Huggins-Manley et al., 2022). With imaginably limited level of language assessment literacy, stakeholders would find it difficult to participate in evaluating test fairness which is in the disguise of "comparable validity". Third, Xi's framework does not clearly specify how to prioritize fairness evaluation and by whom. It seems that testing professionals, the framework's intended audience, would be the most likely to set evaluation priorities. However, given the socially-constructed nature of test fairness, the perspectives of other stakeholders (e.g., test takers, test administrators, and test users) should also be considered in determining evaluation priorities.

As can be seen from the review above, this type of framework conceptualizes fairness more or less as a "derivative" of validity. In other words, fairness has been addressed within the discussions of validity—a somewhat technical issue that is within the expertise of language testers but beyond the reach of other stakeholder groups (e.g., test takers and test users).

The second type of framework, acknowledging the multifaceted nature of test fairness, adopts a relatively broad approach to conceptualizing test fairness. In this type of framework, research focuses go beyond the psychometric qualities of language tests. Since the official debut of test fairness in the field of language testing in the 1990s, Kunnan (2000) made one of the first attempts to outline the scope of test fairness research, placing validity, access, and justice on the evaluation agenda. The consideration of validity centers on the evaluation of construct validity, potential bias in test content or format, item- and test-level psychometric qualities, and insensitive language or content embedded in test materials that might stereotype any societal groups. The consideration of access revolves around the affordability of test fees, the distance of test locations, test accommodations, and test-takers' opportunity to learn the target construct and to familiarize themselves with test environment. The two focuses of justice are: (1) the role of test in promoting societal equity and (2)

legal challenges proposed by test takers who feel discriminated against due to their construct-irrelevant backgrounds.

In order to incorporate the underpinnings of moral philosophy into test evaluation, Kunnan (2004) introduced two general principles (i.e., the Principle of Justice & the Principle of Beneficence) and four sub-principles as the foundation of what would become the Test Fairness Framework (TFF). Following these principles, the TFF brings to the forefront five major qualities of a fair language test: validity, absence of bias, access, administration, and social consequences. The five components included in the TFF are identified from the *Code of Fair Testing Practices in Education* (JCTP, 1988) and the 1999 *Standards* (AERA et al., 1999). Specifically, supportive of the validity of score interpretation is evidence surrounding content representativeness or coverage, construct or theory-based validity, criterion-related validity, and reliability. The second quality, absence of bias, pertains to the review of bias in test content, items, and decision-making processes. Compared to Kunnan's (2000) earlier work, the TFF approaches the quality of "access" very similarly. However, the TFF introduces a new component—"administration"—which focuses on ensuring the comfort of test environment and the consistency of administration practices across test occasions. The TFF also advocates evaluating the social consequences of a test, emphasizing the potential of a language test to generate positive washback and advance social equity.

Another theoretical contribution for evaluating test fairness is an ethics-based assessment evaluation framework. Drawing insights from philosophical theories, Kunnan (2018) put forward the Principle of Fairness and the Principle of Justice to guide test evaluation. The Principle of Fairness revolves around the rights of individual test takers. It asserts that "[an] assessment *ought* to be fair to all test-takers; that is, there is a presumption of treating every test-taker with equal respect" (Kunnan, 2018, p. 80; italics in original). The Principle of Justice emphasizes the role of testing institutions or agencies in fostering positive values and advancing social justice. It suggests that "[an] assessment institution *ought* to be just, bring about benefits in society, promote positive values, and advance justice through public reasoning" (ibid).

Each principle is accompanied by sub-principles that outline specific fairness and justice concerns to be addressed throughout test evaluation. Judging from the sub-principles of Principle 1, key components of test fairness include test-takers' opportunities to learn (sub-principle 1), consistency and meaningfulness in score interpretation (sub-principle 2), absence of bias (sub-principle 3), and accessibility, uniformity in administration, along with defensible grounds for score interpretation and decision-making (sub-principle 4). The sub-principles for Principle 2 suggest that testing institutions or agencies should uphold justice in distributing benefits to test takers and should shoulder the responsibility of promoting beneficence either within testing communities or within society as a whole. To operationalize test evaluation, Toulmin's (1958/2003) argumentation model is applied to build fairness and justice arguments. Relevant standards about fairness and justice in language assessment could be used to link principles and claims in an argument. Kunnan further noted that these standards could be developed by "experts in the field of assessment" or derived from "local standards based on best practices in the field" (Kunnan, 2018, p. 89). However, fairness-related standards from authoritative documents (e.g., the 2014 *Standards*) might be inapplicable in a local language assessment context. Furthermore, there might be no existing local standards available for reference. As a result, applying Kunnan's framework to guide fairness evaluation in a local context could be challenging. In addition, claims and sub-claims of test fairness are produced in a top-down manner, as it is the "test[ing] agencies or researchers" (Kunnan, 2018, p. 90) who articulate them according to established professional standards. Other stakeholders such as test takers and test users are silenced in the articulation of fairness claims or sub-claims. Despite elevating test fairness to a key focus of research, this ethics-based framework overlooks the socially-constructed nature of fairness and lacks a mechanism for gathering stakeholders' fairness concerns.

In a nutshell, the second type of the framework approaches test fairness broadly from a moral philosophy perspective, conceptualizing it as fundamentally an ethical issue that deserves its own research agenda. Nevertheless, testing professionals are

still placed in a central position for articulating fairness claims or sub-claims and navigating fairness evaluation.

The third type of framework acknowledges the socially-constructed nature of test fairness and encourages stakeholder participation in fairness evaluation. Huggins-Manley et al. (2022) introduced an argument-based framework to guide fairness evaluation. Similar to the aforementioned argument-based validation and fairness evaluation frameworks, Huggins-Manley et al. applied Toulmin's (1958/2003) model of argumentation in their framework. In contrast to the conceptualization of fairness as an all-encompassing concept that subsumes validity (Kunnan, 2000, 2004) or vice versa (Xi, 2010), Huggins-Manley et al. (2022) assigned equal importance to test fairness and validity. They argued that fairness arguments can exist as independent of and complementary to validity arguments. This framework is therefore flexible in addressing the evolving nature of fairness and validity, and the intricate relationship between them as well. Further, in response to the varying views on test fairness among different stakeholders, this framework enables proactive stakeholder engagement in collaboratively constructing the claims in a fairness argument with testing professionals. Despite a promising framework, there has been a paucity of empirical efforts devoted to the co-construction of fairness arguments with stakeholders. This lack of research may result from insufficient understanding of stakeholders' perceptions and expectations of test fairness.

Based on the review above, the third type of framework recognizes the socially-constructed nature of test fairness and emphasizes the importance of stakeholder involvement in fairness evaluation.

A review of the theoretical frameworks offers multiple insights into fairness evaluation. First, test fairness has become an essential dimension of test evaluation. While there are some overlaps between the research scopes of test fairness and validity, fairness evaluation can serve as both an independent process from validation and a complementary process to validation. Secondly, a broad spectrum of fairness concerns has been unveiled for evaluation. These concerns can be divided into three categories: psychometric concerns, administrative concerns, and ethical concerns.

Psychometric concerns, which are primarily the research focus of test developers, include ensuring comparable validity across test-taker subgroups. Administrative concerns are about the consistency of test administrations. Ethical considerations encompass test-takers' access to tests and the accountability of testing institutions to the language testing community and society at large. Thirdly, involving stakeholders in test fairness evaluation could help reveal their expectations of test fairness within a specific test context.

2.2.2 Evaluating test fairness from psychometric perspective

This section aims to review two streams of studies that have adopted quantitative approaches to evaluating test fairness. The first stream of studies has examined score comparability across test-taker groups, and the second has focused on the comparability of characteristics of the input materials from different test forms.

2.2.2.1 Studies on the comparability of test scores across test-taker groups

From a psychometric perspective, empirical studies on test fairness can be viewed as attempts to examine the comparability of measurement or test scores across different test-taker groups. Three psychometric techniques—differential item functioning (DIF), differential bundle functioning (DBF), and differential test functioning (DTF) analyses—are frequently employed by both researchers and practitioners in language testing and assessment to evaluate the comparability of test scores obtained by different groups of test takers. DIF, DBF, and DTF analyses can provide valuable evidence regarding score comparability at item, bundle (cluster of items), and test levels across test-taker groups defined by either manifest variables such as gender (e.g., Min & He, 2020), language background (e.g., Drackert & Timukova, 2020), academic background (e.g., Pae, 2004), or latent variables (e.g., Aryadoust, 2015).

An item exhibits DIF when test takers with comparable levels of ability but from different groups have different probabilities of responding to the item correctly (Dorans & Holland, 1993; Shepard et al., 1981). There are two types of DIF: uniform DIF and nonuniform DIF. Uniform DIF is present when an item consistently favors

one group of test takers across all ability levels, whereas nonuniform DIF occurs when the direction of this favoritism varies along the ability continuum (Li & Stout, 1996). Two approaches to DIF investigation have been documented in the literature: exploratory and confirmatory approaches. In the exploratory approach, substantive explanations for the presence of DIF are provided following the statistical identification of DIF items. The confirmatory approach involves proposing *a priori* hypotheses about potential causes or sources of DIF. The hypotheses are typically generated based on existing theories, previous findings, or item content analysis. Subsequently, the hypotheses are tested using statistical procedures. DIF indicates multidimensionality in assessment (Pae, 2004). A large DIF value suggests that the item measures additional dimensions that can lead to performance differences across test-taker groups (Abbott, 2006; Angoff, 1993; Camilli & Shepard, 1994; Roussos & Stout, 1996a, 2004; Shealy & Stout, 1993). These additional dimensions can be further categorized as either construct-relevant or construct-irrelevant. Construct-relevant dimensions, being part of the intended construct, are considered auxiliary. DIF items measuring auxiliary dimensions are benign and can be used to reflect true group differences in the ability targeted by a test. On the contrary, construct-irrelevant dimensions are regarded as nuisances which introduce item bias to a test (Douglas et al., 1996; Roussos & Stout, 1996a). It should be pointed out that the presence of DIF does not necessarily indicate item bias (Li et al., 2022; Shimizu & Zumbo, 2005). Further investigation is needed to determine whether DIF items are truly biased or not (Liu, 2011). In other words, DIF is a necessary but not sufficient condition for item bias (Zumbo, 1999). DIF analysis is only one component of an entire fairness and bias review process (Bowles, 2022).

DBF occurs when test takers of comparable ability levels but from different groups exhibit differences in performance on item bundles (Douglas et al., 1996). Item bundles refer to “any set of items chosen according to some organizing principles” (Douglas et al., 1996, p. 466). For example, a set of items related to the same input material can be grouped into a bundle for DBF analysis (see Min & He, 2020). In this case, these items constitute a testlet (Wainer & Kiely, 1987). DBF can

be considered as a variant of DIF in which the unit of analysis evolves from a single item to a set of items (Liu, 2011). It should be noted that the presence of DBF does not necessarily imply DIF in individual items within that bundle (Lakin et al., 2012). Item-level DIF, even if statistically nonsignificant or negligible, may accumulate at the bundle level and lead to DBF (Douglas et al., 1996; Kim & Jang, 2009; Min & He, 2020). In other cases, DIF items within an item bundle that favor different test-taker groups may cancel each other out at an aggregated level (Liu, 2011).

A test exhibits DTF when it fails to maintain measurement invariance across different groups of test takers (e.g., Zumbo, 2003). A test is considered biased when test takers of comparable ability but from different groups obtain different test scores. The investigation into DTF is valuable and necessary. DTF detection helps determine whether DIF and DBF offer a sizable (dis)advantage to a particular group of test takers at the test level (Elosua, 2024). Previous studies have shown that item-level DIF and bundle-level DBF do not necessarily result in test-level DTF (e.g., Ercikan & Por, 2020; Min & He, 2020).

Previous studies have examined DIF, DBF, and DTF in listening assessment (see Appendix 1 for a full list of the reviewed studies). Among the nine studies reviewed, eight investigated item-level score comparability using DIF techniques and reported the presence of DIF, with DIF ratios ranging from 3.33% (Henning, 1990) to 66.67% (Yang et al., 2022). Uniform DIF was identified in all the eight studies, with half of these studies additionally reporting the presence of nonuniform DIF (Chen, 2013; Pae, 2004; Semiyari & Ahangari, 2022; Xiao, 2013). In some cases, items flagged for uniform DIF did not show nonuniform DIF (e.g., Chen, 2013; Xiao, 2013). However, several items in the English subtest of the Korean National Entrance Exam for Colleges and Universities (Pae, 2004) and the Ministry of Science, Research, and Technology (MSRT) proficiency test (Semiyari & Ahangari, 2022) exhibited both types of DIF. Moreover, the direction and number of DIF items varied across studies. Some studies found that uniform DIF items consistently favored science students (Chen, 2013; Henning, 1990), while other studies, such as He (2022), reported an equal distribution of DIF items favoring either humanities or science

students. Pae (2004) identified more DIF items favoring science students, whereas Xiao (2013) and Zhang and Jin (2012) reported more items favoring humanities students. Furthermore, only three of the eight studies examined the magnitude of DIF and reported the presence of negligible, moderate, and large DIF in test items (Aryadoust et al., 2024; Chen, 2013; Pae, 2004). The presence of large DIF in items would raise practical concerns. Among the nine studies on listening comprehension, only one study (i.e., Yang et al., 2022) examined testlet- and test-level score comparability using DBF and DTF techniques. In the study, two out of three testlets in the listening section of the Practical English Test for Colleges (PRETCO) exhibited negligible DBF, both favoring science students. In addition, the listening section of the PRETCO showed negligible DTF, suggesting that testlet-level DBF may manifest at the test level.

In most of the studies on listening comprehension, attempts were made to explain the occurrence of DIF, DBF, or DTF (see Appendix 1). These explanations revolved around test and test-taker characteristics. Regarding test characteristics, item content and discrimination emerged as potential sources of DIF. Pae (2004) found that science students performed better on items dealing with number counting and job interviews, while humanities students excelled on items concerning human relationships. Additionally, item discrimination appeared to play a role in DIF occurrence. He (2022) suggested that poor discrimination might contribute to DIF. Pae (2004) found that items exhibiting nonuniform DIF demonstrated consistently higher discrimination among humanities students, regardless of their content domains. Beyond test characteristics, test-takers' background knowledge was frequently reported as a likely contributor to DIF, DBF, or DTF in listening comprehension tests. Several researchers proposed that science students' greater exposure to and interest in science-related texts might provide them with an advantage when encountering familiar test content (Chen, 2013; He, 2022; Xiao, 2013; Zhang & Jin, 2012). Strategy use also emerged as a potential source of DIF, DBF, or DTF. Zhang and Jin (2012) found that humanities students were favored in the listening section of the College English Test Band 4 (CET-4). The test-takers' self-reports indicate that humanities

students tended to make inferences based on the texts and were more likely to utilize discourse markers when responding to test items, whereas science students often responded to test items without fully attending to the detailed information in the listening texts. In another study, Yang et al. (2022) attributed the advantage gained by science students to their greater use of metacognitive strategies.

A number of studies have examined DIF, DBF, and DTF in reading assessment (see Appendix 2 for a full list of the reviewed studies). Of the 12 studies investigating item-level score comparability between humanities and science groups, 11 identified DIF, with DIF occurrence ratios varying substantially from 2.5% (Song et al., 2015) to 96.43% (Liu, 2011). This wide range can be attributed to the variations in research contexts and DIF detection methods across studies. With regard to DIF types, 11 studies identified uniform DIF, with five studies also detecting nonuniform DIF (Alavi et al., 2011; He, 2018; Min, 2011; Pae, 2004; Xiao, 2013). Two studies reported items exhibiting both uniform and nonuniform DIF (He, 2018; Pae, 2004). Moreover, research findings regarding the direction and number of uniform DIF items were mixed. Three studies identified uniform DIF items that exclusively favored science students (Jafaripour et al., 2024; Min, 2011; Song et al., 2015). Another three studies reported a greater number of DIF items favoring science students (Brati et al., 2006; Ghaemi & Khorami, 2024; Pae, 2004). In contrast, one study found more DIF items that favored humanities students (Xiao, 2013). Furthermore, 10 studies examined the magnitude of DIF. Four studies reported the presence of small DIF in test items (Alavi et al., 2011; Ghaemi & Khorami, 2024; He, 2018; Liu, 2011), six identified moderate DIF in test items (Ghaemi & Khorami, 2024; He, 2018; Liu, 2011; Pae, 2004; Song et al., 2015; Xiao, 2013), and only two found large DIF in test items (Liu, 2011; Xiao, 2013). Of the 14 studies on reading assessment, four studies extended their analyses beyond the item level to examine testlet-level score comparability (Chen & Zeng, 2021; Liu, 2011; Song et al., 2015; Yang et al., 2022). Liu (2011) found no testlets exhibiting DBF despite the presence of item-level DIF. This finding suggests that DIF items favoring the focal or reference group may lead to DIF cancellation at the testlet level. Chen and Zeng (2021) and

Yang et al. (2022) identified testlets exhibiting DBF that favored humanities students, while Song et al. (2015) identified one testlet that favored science students. At the test level, Yang et al. (2022) observed only negligible DTF in the PRETCO reading section, with a slight advantage for humanities students.

The studies on reading comprehension provided various explanations for the observed differential item, testlet, and test functioning (see Appendix 2). These explanations can be classified into two categories: those related to test characteristics and those associated with test-taker characteristics. Regarding test characteristics, the presence of DIF or DBF can be attributed to the topics of the reading input materials (Ghaemi & Khorami, 2024; Pae, 2004; Song et al., 2015; Xiao, 2013), the subskills assessed by certain items (Brati et al., 2006; Yang et al., 2022), item discrimination (He, 2018; Pae, 2004), and response formats (Xiao, 2013). DIF was observed in the reading passages featuring science-related topics such as physics and geology (Ghaemi & Khorami, 2024), the effects of snow on animals (Pae, 2004), DNA (Song et al., 2015), and energy (Xiao, 2013). In these content domains, the DIF items consistently favored science students. In contrast, DIF items present in reading passages about human relationships (Pae, 2004), language, and literature (Ghaemi & Khorami, 2024) favored humanities students. Regarding subskills, DIF items assessing the subskills of understanding main ideas and making inferences favored students with mathematics and science backgrounds (Brati et al., 2006), while items targeting the retrieval of detailed information advantaged humanities students (Yang et al., 2022). Additionally, item discrimination patterns in nonuniform DIF items showed variations across studies. He (2018) found that a nonuniform DIF item was more discriminating for the group of science students, whereas Pae (2004) reported greater DIF item discrimination among humanities students. Response format emerged as another potential source of DIF. Xiao (2013) identified a uniform DIF item favoring science students in a banked cloze task which required test takers to select words from a bank of possible options. This response format was hypothesized to contribute to DIF occurrence.

In addition to test characteristics, test-taker characteristics also contributed to the presence of DIF, DBF, or DTF in reading comprehension tests. One such contributor is the test-takers' background knowledge of the subject matter presented in reading texts. Students in science disciplines were reported to gain an advantage on items associated with science-related passages (Liu, 2011; Min, 2011), likely due to their greater exposure to reading texts on science and technology throughout their academic studies. This disciplinary immersion helped them foster greater familiarity with domain-specific vocabulary and terminology (Song et al., 2015). In a similar vein, humanities students demonstrated advantages on items associated with reading texts that addressed social issues (Chen & Zeng, 2021). According to Liu (2011), students were more likely to achieve a higher level of understanding of a reading text when they were familiar with its topic. Moreover, students' cognitive abilities also played a role in DIF occurrence (Jafaripour et al., 2024). Students in mathematics were reported to possess stronger cognitive abilities and outperform their humanities counterparts on items assessing gist understanding (Brati et al., 2006). Additionally, gender distribution among students in different disciplines was identified as a contributor to DTF. For example, the advantage held by the humanities group in the reading section of the PRETCO was attributed to the group's gender distribution, with females comprising 87.6% of the cohort (Yang et al., 2022). Furthermore, the observed DIF might reflect genuine differences in English proficiency between humanities and science groups. This is evident when DIF analyses reveal unexpected directions of DIF. For instance, items initially hypothesized to favor science students were found to favor humanities students (Ghaemi & Khorami, 2024). Content analyses of these identified DIF items indicate that they assessed construct-relevant language abilities rather than introducing bias into tests (Ghaemi & Khorami, 2024; Liu, 2011; Min, 2011).

A review of the studies examining the measurement attribute of test fairness has identified three research gaps. First, while previous studies have employed various detection methods to examine test-, testlet-, and test-level score comparability across test takers with different academic background, few studies have examined multiple

levels of measurement invariance using DIF, DBF, and DTF techniques simultaneously. Conflicting results have been documented in existing literature regarding the relationship between DIF and DBF (Liu, 2011; Song et al., 2015), between DIF and DTF (Elosua, 2024; Pae, 2004; Raquel, 2019; Zumbo, 2003), and between DBF and DTF (Min & He, 2020; Yang et al., 2022). Given the intricate relationships among different levels of measurement invariance, further research is needed to examine DIF, DBF, and DTF simultaneously. Second, while statistical techniques (e.g., DIF, DBF, and DTF) are instrumental in identifying potential bias in tests across test-taker groups (Camilli, 2016), they are unable to address fairness concerns at the level of individual test takers (Nisbet & Shaw, 2020). Third, quantitative approaches tend to narrow the research scope of test fairness to test quality, overlooking aspects that precede and follow test administration (Klenowski, 2014). A comprehensive evaluation of test fairness should address aspects such as test-takers' opportunities to learn construct-relevant knowledge and skills, the validity of test score interpretation, and the consequences of score-based decisions (Gipps & Stobart, 2009).

2.2.2.2 Studies on the comparability of input materials across test forms

The use of multiple test forms in a single administration or across different administrations is a common practice for many large-scale language tests (Jin & Wu, 2007; Langenfeld, 2020; Weir & Wu, 2006). While this practice enhances test security, it has raised concerns regarding the comparability of different test forms and presented a challenge for test developers (Deygers, 2019). To ensure test fairness, test tasks in different test forms must be made as comparable as possible (Willingham, 1997). An important consideration in this regard is the comparability of task difficulty, as task difficulty can influence test-takers' performance (Weir & Wu, 2006). If the difficulty levels of tasks in different test forms are not comparable, it would result in unfairness for test takers completing different test forms at different time and locations.

The complexity of input materials is a key factor contributing to task difficulty (Skehan, 1998). For listening tasks, the characteristics of the input materials can influence the processing difficulty of spoken texts. These characteristics include lexical complexity (e.g., Brunfaut & Révész, 2015; Pan, 2021; Révész & Brunfaut 2013), syntactic complexity (e.g., He et al., 2018; Kostin, 2004; Pan, 2021), discourse complexity (e.g., Brunfaut & Révész, 2015), and speed of delivery (e.g., Brindley & Slatyer, 2002; Green, 2017; Griffiths, 1992; Yoon et al., 2016; Zhao, 1997). For reading tasks, the difficulty level of the input materials is influenced by a few textual characteristics such as lexical complexity, syntactic complexity, discourse complexity, and readability (Liao, 2020; Zeng, 2022). These input characteristics play an important role in determining the cognitive demands placed on test takers, thereby shaping the overall difficulty of the listening and reading tasks.

There has been remarkably little research on the comparability of the difficulty levels of input materials used across different test forms. To the best of the author's knowledge, Liao (2020) is the only study to examine text difficulty across parallel academic IELTS reading test forms. By analyzing the characteristics of the reading passages, the study found that text difficulty was partially comparable in terms of word concreteness, syntactic complexity, readability, and cohesion. As Liao (2020) pointed out, the comparability of test forms should be evaluated on a case-by-case basis, as results may vary depending on the specific test being studied. As a high-stakes test, the EPT aims to administer test forms that are comparable in difficulty levels across test sessions. Nevertheless, the extent to which the input materials used across different test forms are comparable in terms of complexity and difficulty remains an area that requires further research attention.

2.2.3 Evaluating test fairness from stakeholders' perspective

A number of studies have examined stakeholders' perceived fairness of language tests and testing practices, employing methods such as questionnaire surveys (e.g., Heeren et al., 2021; Sonnleitner & Kovacs, 2020; Tofighi & Safa, 2023), interviews (e.g., Fox & Cheng, 2007; Huang & Garner, 2009; Song, 2018), or a combination of both

(e.g., Fan & Ji, 2014; Jang, 2002; Lu et al., 2023). This stream of study is rooted in the premise that "[a] test is fair if the stakeholders...perceive it to be fair" (Wallace & Ng, 2023, p. 530). Studies in this area have drawn valuable insights from a variety of stakeholders, including test takers (e.g., Fan & Ji, 2014; Jang, 2002; Song, 2018), teachers (e.g., Fan, 2018; Safari & Rashidi, 2018; Wallace & Ng, 2023), test developers (e.g., Moghadam & Nasirzadeh, 2020), test administrators (e.g., Hawkey, 2008; Huang & Garner, 2009; Song, 2018), and university officials (e.g., Moghadam & Nasirzadeh, 2020).

Previous studies have examined stakeholders' overall perceptions of the fairness of language tests (e.g., Hawkey, 2018; Song, 2018; Yao, 2023). For instance, Hawkey (2008) found that the majority of test takers considered the IELTS to be a fair measure of their English proficiency. Similarly, Song (2018) reported that stakeholders, including test takers, teachers, and program administrators, endorsed the use of the Graduate School Entrance English Examination (GSEEE), as it provided fair opportunities for all test takers to gain admission to graduate programs in China. However, stakeholders in her study raised concerns about item quality, test administration, and scoring practices. These findings suggest that negative perceptions of specific aspects of a test or testing practices do not necessarily affect stakeholders' judgments of the test's overall fairness.

Relevant findings from the empirical studies fall into four categories which align with the four dimensions of test fairness outlined in Section 2.1.2—comparability, accessibility, consistency, and accountability.

Several studies have highlighted various comparability issues in language tests and testing practices. A few studies have addressed the comparability of opportunities for different test-taker subgroups to demonstrate relevant language proficiency. For instance, in Jang's (2002) study, some test takers expressed concerns that the TOEFL might favor individuals with a natural science background, as a number of reading passages featured topics related to this field. Similarly, in Yao's (2023) study, Chinese university students questioned the appropriateness of test content in the DET, arguing that the test materials did not adequately align with university contexts. As a

result, they felt disadvantaged when taking the DET. A key implication drawn from these studies is that test content can be a source of potential bias in language tests. Construct underrepresentation has also emerged as a fairness concern. Some test takers in Song's (2018) study perceived the GSEEE as unfair because the test did not assess their listening and speaking skills which are essential components of English proficiency. Consequently, test takers with jagged English profiles, particularly those who excel in reading and writing but are weak in listening and speaking skills, could be unfairly advantaged. In contrast, test takers of the FET, a full-skill English proficiency test, held a positive attitude toward the overall design of the test, as it assesses test-takers' English proficiency in a comprehensive manner (Fan & Ji, 2014). Additionally, a handful of studies have explored stakeholders' perceptions on the comparability of parallel test forms. For example, a test taker in Fan's (2018) study expressed concerns about the use of parallel test forms in a single administration, noting that the test forms often varied in terms of difficulty. Hamid et al. (2019) reported that a repeat test taker observed significant differences in IELTS scores across two test sittings within a two-week period. These findings underscore the importance of ensuring comparability in difficulty levels and test scores across parallel test forms.

Empirical studies have also shed light on the stakeholders' views of accessibility issues. Studies regarding the accessibility of test-related information have yielded mixed findings. Questionnaire data (Fan & Ji, 2014; Lu et al., 2023; Moghadam & Nasirzadeh, 2020; Yao, 2023) and interview data (Choi, 2016; Lu et al., 2023) indicate that test takers had access to test-related information (e.g., test construct, test administration procedures, rating criteria, and past test papers) prior to test-taking. However, Tofighi and Safa (2023) found that teachers administering a classroom-based assessment did not disclose the rubrics to their students in advance. Similarly, for the FET, test-related information was not made transparent to test takers before the test day (Fan & Ji, 2014). With regard to test-taking opportunities, test takers reported feeling fairly treated when they were allowed to choose their preferred test dates (Yao, 2023) and were provided with multiple opportunities to take the test

(Tofighi & Safa, 2023). In terms of learning opportunities, it was reported that test takers had access to sufficient resources and opportunities to learn the knowledge and skills required for the construct being measured (Lu et al., 2023; Moghadam & Nasirzadeh, 2020). As for financial accessibility, locally-developed language tests were reported to be more affordable for test takers compared to large-scale international ones (Moghadam & Nasirzadeh, 2020; Weir, 2019). With respect to geographical accessibility, test takers had equal access to the location of a local reading comprehension test (Moghadam & Nasirzadeh, 2020). Lastly, concerning the accessibility of test equipment and delivery system, test takers of the Vocational English Test System (VETS) and the National Matriculation English Test (NMET) (administered in Shanghai) were given opportunities to take a mock test and familiarize themselves with the test equipment and delivery system before the test day (Lu et al., 2023; Xu, 2019). In contrast, test takers of the computer-delivered speaking component of the FET reported feeling nervous during the test due to their lack of familiarity with the test delivery mode (Fan & Ji, 2014).

Consistency issues in test administration and scoring have also been a focus of research. Regarding test administration, evidence from two studies suggests that test takers expressed a positive attitude toward administration procedures (Butler et al., 2021; Fan & Ji, 2014), which is possibly due to the consistent and equal treatment they received during test-taking. However, Fox and Cheng's (2007) study revealed inconsistencies in the test environment across secondary schools in Ontario, Canada. Jang (2002) found that low-quality test equipment and distracting noises made it difficult for several TOEFL test takers to perform well on the test day. Furthermore, research has documented a variety of measures implemented to ensure test security and standardization in test administration (e.g., Moghadam & Nasirzadeh, 2020; Song, 2018). Despite these efforts, instances of cheating among test takers remain a persistent concern (He, 2015; Huang & Garner, 2009). He (2015) attributed cases of cheating to the lack of attentiveness on the part of proctors. With respect to scoring, test takers highlighted the significance of clear rating criteria in ensuring consistency when scoring compositions and oral responses (He, 2015; Sonnleitner & Kovacs,

2020). However, it should be noted that the descriptors used in rating criteria may lack clarity and are often open to various interpretations. As a result, test takers in the study by Sonnleitner and Kovacs (2020) expressed a preference for teachers to clarify the rating criteria prior to the speaking test. Similarly, in Choi's (2016) study, test takers wished to have access to model essays along with the corresponding ratings and raters' comments. In addition to concerns about rating criteria, He (2015) found that issues with the scoring procedures for college achievement tests at her university undermined the consistency and quality of the ratings.

Accountability issues have also received research attention. One practice that fosters accountability is involving stakeholders in the evaluation of test quality and testing practices. Their engagement enables test designers and developers to better understand various stakeholders' concerns and expectations about test fairness (Caines et al., 2014). However, stakeholders, particularly test takers, are often excluded from participating in test policy formation, test development, and decision-making (e.g., Barrance & Elwood, 2018; Deygers, 2019). In response, Deygers (2019) advocated for greater stakeholder engagement in testing practices. Furthermore, periodic fairness evaluation constitutes an essential component of accountable assessment. For example, Moghadam and Nasirzadeh (2020) evaluated the fairness of a locally-developed reading comprehension test and identified several validity issues that remained to be addressed.

Although limited in number, existing studies have identified several factors that influence stakeholders' perceptions of test fairness. For instance, disparities in pre-university English teaching have been identified to influence test-takers' perceived fairness of the use of an in-house high-stakes English proficiency test (Fan et al., 2022). Additionally, test-takers' perceptions of test fairness have been shown to be influenced by their language proficiency levels (Heeren et al., 2021; Moghadam & Nasirzadeh, 2020; Tsai & Tsou, 2009). Tsai and Tsou (2009), for example, found that test takers with self-reported satisfactory English proficiency viewed the use of a high-stakes English test as a graduation benchmark more positively than those with lower self-assessed proficiency. Moreover, stakeholders' perceptions of test fairness

can be influenced by whether test takers have opportunities to learn the knowledge and skills being assessed (Butler et al., 2021; Tierney, 2014; Tofighi & Safa, 2023).

While previous studies have offered valuable insights into stakeholders' perceptions of test fairness, three research gaps remain. First, the existing literature has predominantly focused on the perceptions of test takers, whereas those of other key stakeholders—such as test developers, test administrators, and test users—have been largely overlooked. Second, there has been a lack of studies adopting a multi-stakeholder approach, with only a limited number of exceptions (e.g., Hawkey, 2008; Huang & Garner, 2009; Song, 2018). Worthy of note is that the perceptions of a single stakeholder group can be inherently subjective. For instance, test-takers' perceptions of test fairness are often influenced by their demographic background and language proficiency levels (Heeren et al., 2021; Iwashita & Elder, 1997; Moghadam & Nasirzadeh, 2020). More importantly, different stakeholder groups have been reported to possess varying levels of knowledge about a test (Song, 2018). A multi-stakeholder approach is therefore recommended for a relatively objective and comprehensive evaluation of test fairness (Huggins-Manley et al., 2022). Third, there is a paucity of studies exploring the factors that influence stakeholders' perceptions of test fairness, except for a few that have focused on factors related to test-takers' educational opportunities and characteristics (e.g., Fan et al., 2022; Tsai & Tsou, 2009). Bridging these gaps is crucial for enhancing test quality, improving testing practices, and fostering greater public trust in the use of high-stakes language tests.

2.3 Proposing a tentative conceptual model of test fairness evaluation

Test fairness possesses a dual nature, encompassing both measurement and value judgment attributes. From one perspective, test fairness can be measured through the collection of quantitative evidence using psychometric methods. From another perspective, test fairness, as a value-laden concept, is subject to stakeholders' perceptions and interpretations across different contexts. In recognition of the dual nature of test fairness, this study proposes a tentative conceptual model of test fairness evaluation.

The tentative conceptual model seeks to capture the duality of test fairness by adopting a two-layered concentric-circle structure (see Figure 2.1). Drawing on insights from the dimensions of test fairness (Section 2.1.2), existing fairness evaluation frameworks (Section 2.2.1), and empirical investigations (Section 2.2.2 & Section 2.2.3), four dimensions of test fairness are identified to form the inner circle of the tentative conceptual model.

The inner circle, consisting of a set of key focuses and essential criteria for test fairness, reflects the multifaceted nature of test fairness:

- **Comparability** embodies the "relational" sense of fairness. For a test to be considered fair, it must adhere to the criterion of comparability throughout all assessment stages, from test design to score interpretation. Key considerations underpinning this criterion include, but are not limited to, the comparability of: (1) test results across test-taker subgroups and across different test forms, (2) difficulty levels across different test forms, and (3) opportunities for different test-taker subgroups to demonstrate their language proficiency. This criterion helps ensure that test takers with similar levels of ability receive comparable scores regardless of their demographic characteristics and across different test administrations.
- **Accessibility** addresses the extent to which test takers have equal access to test-related information, learning and preparation resources, test-taking opportunities, test locations, and, when applicable, test delivery equipment (in the case of computer-delivered tests). Ensuring accessibility aims to remove any barriers that may hinder test takers from taking a test or performing to the best of their abilities during the test.
- **Consistency** refers to the uniformity of testing practices throughout the testing cycle, from test development, test administration, scoring, to score interpretation. This criterion ensures standardization in testing practices.
- **Accountability** refers to the mechanisms, procedures, and practices that uphold the fairness of a test. For instance, systematic fairness evaluation conducted by testing professionals helps identify potential concerns

threatening test fairness. Additionally, score review procedures allow test takers to challenge or seek reconsideration of test results or score-based decisions. Furthermore, seeking input from diverse stakeholder groups offers opportunities for them to provide feedback and suggestions on the test and the associated testing practices. The accountability dimension also emphasizes the collective responsibility of all stakeholders to ensure test fairness.

Surrounding the four dimensions of test fairness is an outer circle that encompasses four key stakeholder groups: test takers, test developers, test administrators, and test users. This outer circle acknowledges that test fairness is a social concern rather than merely a technical issue. Stakeholders play a role in making value judgements about the fairness of a test and its associated testing practices. The outer circle highlights the socially-constructed nature of test fairness.

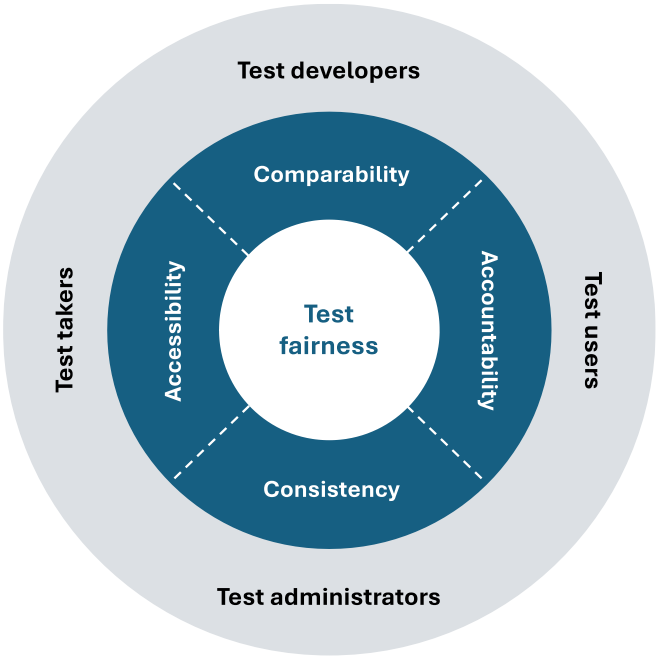

Figure 2.1 A tentative conceptual model of test fairness evaluation.

2.4 Chapter summary

Chapter 2 presents a review of existing literature on test fairness. Section 2.1 reviews the etymology of "fair", the basic senses of fairness, and the dimensions of test fairness (i.e., comparability, accessibility, consistency, and accountability).

Section 2.2 provides an overview of theoretical frameworks and empirical studies devoted to evaluating test fairness. Section 2.2.1 reviews three types of theoretical frameworks and discusses their insights and limitations. Section 2.2.2 reviews empirical studies addressing both the measurement and value judgement attributes of test fairness. Section 2.2.2.1 presents a brief introduction to three psychometric techniques (i.e., DIF, DBF, and DTF) frequently used to assess the comparability of test scores across different groups of test takers. This section then reviews studies that use DIF, DBF, and DTF techniques to detect bias in listening and reading comprehension tests administered to test-taker groups with humanities and science backgrounds. Section 2.2.2.2 reviews studies on the comparability of input materials across parallel test forms, with a focus on the characteristics of these materials. Section 2.2.3 reviews empirical studies that evaluate the value judgement attribute of test fairness.

Drawing on insights from the dimensions of test fairness, theoretical frameworks, and empirical investigations, Section 2.3 proposes a tentative conceptual model of test fairness evaluation. This conceptual model consists of four dimensions of test fairness (i.e., comparability, accessibility, consistency, and accountability) along with four key stakeholder groups (i.e., test takers, test developers, test administrators, and test users). The tentative conceptual model not only recognizes the measurement and value judgment attributes of test fairness but also acknowledges its multifaceted and socially-constructed nature.

Chapter 3 Methodology

This chapter provides an overview and details of the research design employed in this doctoral study. A convergent mixed-methods design (Creswell & Creswell, 2023) was adopted, incorporating a quantitative and a qualitative phase. In response to RQ1.1, RQ1.2, and RQ2.1, the quantitative phase consists of three studies: a test score comparability study, an input material comparability study, and a questionnaire survey. To complement quantitative inquiry, one-on-one semi-structured interviews were conducted with the representatives of different stakeholder groups in the qualitative phase to address RQ2.1 and RQ2.2. The datasets from both phases were analyzed separately before being combined for corroboration. Convergence and contradictions in the datasets were identified, interpreted, and explained in Section 5.1.3. This chapter concludes with a brief recap of the research design for this study.

3.1 Overview of research design

This study employed a convergent mixed-methods design (Creswell & Creswell, 2023) to evaluate the fairness of a high-stakes in-house English proficiency test (EPT) from the perspectives of psychometrics and stakeholders (see Figure 3.1 for an overview of the research design).

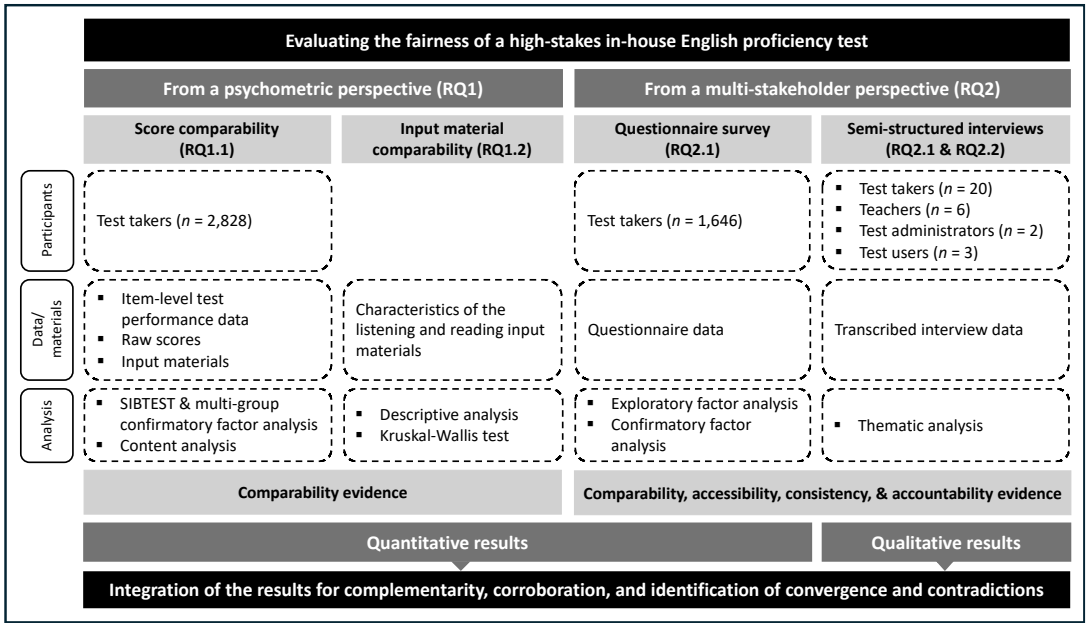

Figure 3.1 A schematic overview of research design.

From a psychometric perspective (RQ1), RQ1.1 focused on item-, testlet-, and test-level score comparability across test-taker groups with different academic backgrounds. Drawing from previous research findings (see Section 2.2.2.1), this study hypothesized that the humanities test-taker group would perform better on items and testlets related to language learning, education, history, and social issues, whereas the science group would excel on items or testlets pertaining to technology, medicine, and other science topics. To test this hypothesis, DIF and DBF analyses were conducted on test-takers' performance in the listening and reading subtests of the EPT. Meanwhile, a DTF analysis was conducted to determine whether DIF and DBF manifested themselves at the test level and compromise the validity of score interpretation. To interpret potential occurrences of DIF, DBF, and DTF, input materials for the two subtests were subjected to content analysis. RQ1.2 aimed to examine input materials in terms of the comparability of their characteristics across four test forms used in a single test administration. To achieve this, characteristics of the listening and reading input materials were extracted following tailor-made schemes (see Section 3.6.1.2 for details) and analyzed using both descriptive and inferential statistical methods.

From a multi-stakeholder perspective (RQ2), RQ2.1 investigated stakeholders'

perceptions of the fairness of the EPT. RQ2.2 sought to identify the factors influencing their perceptions. A questionnaire survey was conducted among test takers to address RQ2.1. Meanwhile, one-on-one semi-structured interviews were conducted with a small sample of test takers, teachers (also test developers), test administrators, and test users. A thematic analysis was conducted for the transcribed interviews to examine stakeholders' perceptions of test fairness (RQ2.1) and to uncover potential factors influencing their perceptions (RQ2.2).

The research design was informed by the tentative conceptual model of test fairness evaluation (see Section 2.3). Among the four dimensions of test fairness, comparability was evaluated by analyzing test scores, characteristics of the input materials, test-takers' questionnaire responses and stakeholder interview data. The accessibility, consistency, and accountability dimensions were evaluated through a questionnaire survey and semi-structured interviews conducted with a sample of stakeholder representatives. The findings from both the quantitative and qualitative phases were integrated for complementarity, corroboration, and identification of convergence and divergence (see Section 5.1.3).

The decision on a convergent mixed-methods approach is based on four considerations. First, employing both quantitative and qualitative approaches can provide a comprehensive understanding of the research questions than using either approach in isolation (Creswell & Clark, 2017). Quantitative and qualitative data are equally important in this study to evaluate the EPT's fairness from psychometric and stakeholders' perspectives. Second, the quantitative and qualitative phases are intended to complement each other. The results derived from one approach can be triangulated and corroborated by using the other. Third, by combining quantitative and qualitative approaches, researchers can better interpret and account for potential unexpected findings obtained from either approach. The last lies in the practical value of the findings (Bryman, 2006). Insights from both quantitative and qualitative strands are expected to enhance the usefulness and applicability of the findings for language testing practitioners and other stakeholders (e.g., test developers, test administrators, and test users) as well.

3.2 Introduction of the EPT

3.2.1 General description of the EPT

The EPT is a high-stakes in-house test developed and administered at a comprehensive university in eastern China. It serves as exit requirement for more than 20,000 non-English-major undergraduate students, ensuring they meet the university's English language proficiency requirements upon graduation. According to the *Handbook for Undergraduate Students* (hereafter referred to as the *Handbook*), undergraduate students are required to earn seven credits in foreign languages prior to graduation. Six credits can be earned by completing English as a Foreign Language (EFL) courses (e.g., College English courses), while the remaining one credit is obtained by passing the EPT. In other words, students who fail to pass the EPT cannot obtain their bachelor's degrees. This EPT-as-exit-test policy therefore makes the test high-stakes in nature.

The EPT, as a full-skill test, aims to assess undergraduate students' general English proficiency across four skills: listening, reading, writing, and speaking. It consists of four subtests, among which the listening, reading, and writing subtests are delivered on computers, while the speaking subtest is conducted in a face-to-face format. The EPT has a total score of 100, with 30% for the listening and reading subtests respectively and 20% for the writing and speaking subtests respectively. To pass the EPT, test takers must achieve minimum scores of 36 on the listening and reading subtests, 12 on the writing subtest, and 12 on the speaking subtest. Students who meet the passing criteria for all subtests are deemed qualified to pass the EPT and earn the required credit. The test content of the EPT features a balanced representation of diverse topics that are relevant to university contexts and accessible to all test takers.

The listening subtest consists of 30 four-option multiple-choice items. It includes three sections. Section A includes 10 items, each based on a short conversation. Section B features long conversation with five items. Section C comprises 15 items, with five items for each of the three passages. The recording is

delivered at a speed of 130 words per minute, with each dialogue, passage, and question played only once. The listening subtest takes about 30 minutes to complete.

The reading subtest consists of 20 items. It includes two sections. Section A comprises two passages (approximately 350 words each), followed by five four-option multiple-choice items. Section B features a banked cloze task with a passage of about 300 words and ten blanks. Test takers must choose 10 appropriate words from a pool of 15 options to complete the passage. The reading subtest, which is administered immediately after the listening subtest, has a time allotment of 25 minutes.

The writing subtest requires test takers to write a composition of at least 160 words within 30 minutes based on a given prompt. According to the test specifications, the compositions are scored by both human raters and an automated writing evaluation (AWE) system. The scores obtained from these two sources are integrated to yield the final score.

In the speaking subtest, four test-takers form a group, with two oral examiners present in the test room to evaluate their performance. The subtest consists of three parts, with the prompts in all three parts revolving around the same topic in each test session. The subtest lasts for approximately 20 minutes. In Part 1, each test taker is randomly assigned a prompt and given one minute to prepare, followed by 1.5 minutes to respond to the prompt. Part 2 features a six-minute group discussion task. In Part 3, the oral examiners ask each test taker one last question.

The listening, reading, and writing subtests have been aligned to the *China's Standards of English Language Ability* (hereafter referred to as the *CSE*; Ministry of Education of the People's Republic of China & National Language Commission of the People's Republic of China, 2018). The alignment results indicate that the students need to achieve Level 5, as specified by the CSE, to pass the three computer-based subtests.

The EPT is administered twice a year (in April and October), with about 7,000 test takers each year. According to the test specifications, students are eligible to register for the test beginning in their second year of studies and may retake any

subtests they do not pass. Although there is no limit on the number of test attempts, test takers are not allowed to re-register for any subtests they have successfully passed. Additionally, students can only register for the speaking subtest after passing the listening, reading, and writing subtests. Students receive their pass or fail status for each subtest one week following the test administration. Test-related information and policies are documented in the test syllabus and the *Handbook*, which are available on the university's official websites and can be accessed to all students.

The university has seven campuses. The EPT is administered on the main campus. All the first-year and second-year undergraduate students, regardless of their major, attend classes and live on the main campus. Starting from the third year, the undergraduate students in certain disciplines are relocated to other campuses for study and living. Among the seven campuses, five are in the same city, while two are in different cities. The university provides shuttle bus services to ensure that students from other campuses arrive at the test location punctually.

Due to the high-stakes nature of the EPT, multiple test forms are administered across different test sessions to maintain test security. To ensure the comparability of the results across test forms, an anchor-item design is employed in the assembly of the test forms for the listening and reading subtests. In this design, a certain number of items, known as "anchor items", are included in all test forms administered during different sessions on the same day. The use of anchor-item design ensures that item parameters estimated from different test forms are calibrated on the same scale. The anchor-item design facilitates post-test equating procedures which compensate for variations in test form difficulty and maintain score comparability across test forms.

3.2.2 Listening and reading subtests of the EPT used in this study

Data of one test administration of the EPT were used in this study, including four forms. Table 3.1 summarizes the tasks in the four forms of the listening subtest. The listening tasks cover a wide range of general topics. The five items associated with LP2 and those associated with LP3 constitute the anchor items used to link Form 1 and Form 2. In other words, these ten anchor items appear both in Form 1 and Form

2. Similarly, the items associated with LC2 and LP4 serve as anchor items to link Form 2 and Form 3. The items associated with LP5 and LP6 are the anchor items linking Form 3 and Form 4.

Table 3.1 Summary of the tasks in the four test forms of the listening subtest.

Test form	Task ID	Item ID	Score range	Topics
Form 1				
Section A	SC1	1–10	0–10	Daily activities; weather; transportation; jobs; health; education
Section B	LC1	11–15	0–5	Education
Section C	LP1	16–20	0–5	Language learning
	LP2	21–25	0–5	Health
	LP3	26–30	0–5	Unemployment
Form 2				
Section A	SC2	1–10	0–10	Education; entertainments; jobs; weather; livings
Section B	LC2	11–15	0–5	Ordering food
Section C	LP4	16–20	0–5	Politics
	LP2	21–25	0–5	Health
	LP3	26–30	0–5	Unemployment
Form 3				
Section A	SC3	1–10	0–10	Education; health; daily activities; pressures
Section B	LC2	11–15	0–5	Ordering food
Section C	LP4	16–20	0–5	Politics
	LP5	21–25	0–5	Marketing a music training program
	LP6	26–30	0–5	Medical sciences
Form 4				
Section A	SC4	1–10	0–10	Business; health; complaints; appearance; clothing; employment
Section B	LC3	11–15	0–5	Retirement age
Section C	LP7	16–20	0–5	Suggestions for sledding
	LP5	21–25	0–5	Marketing a music training program
	LP6	26–30	0–5	Medical sciences

Notes. SC = Short conversation. LC = Long conversation. LP = Listening passage.

Table 3.2 summarizes the tasks in the four test forms of the reading subtest. The reading comprehension tasks include various topics. Specifically, the five items associated with RP2 are used as anchor items to link Form 1 and Form 2. Similarly,

the five items associated with RP3 serve as the anchor items between Form 2 and Form 3. The five items associated with RP4 are used to link Form 3 and Form 4.

Table 3.2 Summary of the tasks in the four test forms of the reading subtest.

Test forms	Task ID	Item ID	Score range	Topics
Form 1				
Section A	RP1	1–5	0–10	Multitasking
	RP2	6–10	0–10	Driverless cars
Section B	BC1	11–20	0–10	Traveling
Form 2				
Section A	RP3	1–5	0–10	Animation
	RP2	6–10	0–10	Driverless cars
Section B	BC2	11–20	0–10	Boarding schools
Form 3				
Section A	RP3	1–5	0–10	Animation
	RP4	6–10	0–10	Gun control
Section B	BC3	11–20	0–10	Enterprises
Form 4				
Section A	RP5	1–5	0–10	Smart devices
	RP4	6–10	0–10	Gun control
Section B	BC4	11–20	0–10	Extinction of Neanderthals

Notes. RP = Reading Passage. BC = Banked Cloze. The item types for each reading comprehension task are multiple-choice questions.

3.3 Participants

Participants in this study include test takers, teachers (also test developers), test administrators, and test users, all of whom are key stakeholders in the context of the EPT. Test takers are key stakeholders due to the high-stakes nature of the test. Teachers in this study, who are also test developers, have multiple roles. They deliver credit-bearing College English courses designed to enhance students' proficiency in listening, reading, writing, and speaking—the key English abilities and skills assessed by the EPT. They are also involved in the development of the EPT, leveraging their expertise in language testing and assessment. Additionally, some of them are responsible for rating the test-takers' writing scripts. Test administrators play a critical role in maintaining consistent test administration and managing score reviews. Test users rely on the test results to make graduation decisions for the target

test takers. The following sections will outline the sampling strategies, sample sizes, and demographic information for each stakeholder group.

3.3.1 Test takers

A total of 2,828 test takers took the written test of the EPT, which included listening, reading, and writing subtests. As shown in Table 3.3, the test was administered in four sessions on a single day, with slightly different number of test takers per session (663 for Session 1, 719 for Session 2, 716 for Session 3, and 730 for Session 4). The test takers comprised 63.47% males (n = 1,795) and 36.53% females (n = 1,033). In terms of grade, 33.80% of the test takers were sophomores (n = 956), 41.09% were juniors (n = 1,162), and 21.71% were seniors (n = 614). A small percentage of the test takers, 3.39% in total, were either in their fifth year of study (n = 75) or beyond (n = 21). The test takers were from seven faculties of the university[1]: 203 (7.18%) from the Faculty of Arts and Humanities, 311 (11%) from the Faculty of Social Sciences, 262 (9.26%) from the Faculty of Science, 955 (33.77%) from the Faculty of Engineering, 477 (16.87%) from the Faculty of Information Technology, 335 (11.85%) from the Faculty of Agriculture, Life and Environment Sciences, and 285 (10.08%) from the Faculty of Medicine and Pharmaceutical Sciences.

[1] Although a comprehensive university, the number of programs and undergraduate students in science, engineering, information technology, agriculture, and medicine far exceed those in humanities and social sciences. The gender and discipline distribution of the sample of test takers was representative of the target test-taker population at the university.

Table 3.3 Demographic profiles of the test takers across test sessions.

	Session 1	Session 2	Session 3	Session 4	Total
Gender					
Male	413 (14.60%)	429 (15.17%)	490 (17.33%)	463 (16.37%)	1,795 (63.47%)
Female	250 (8.84%)	290 (10.25%)	226 (7.99%)	267 (9.44%)	1,033 (36.53%)
Program year					
2nd	325 (11.49%)	288 (10.18%)	195 (6.90%)	148 (5.23%)	956 (33.80%)
3rd	222 (7.85%)	307 (10.86%)	308 (10.89%)	325 (11.49%)	1,162 (41.09%)
4th	92 (3.25%)	112 (3.96%)	180 (6.36%)	230 (8.13%)	614 (21.71%)
5th	18 (0.64%)	9 (0.32%)	26 (0.92%)	22 (0.78%)	75 (2.65%)
Other	6 (0.21%)	3 (0.11%)	7 (0.25%)	5 (0.18%)	21 (0.74%)
Field of study					
Arts and humanities	48 (1.70%)	61 (2.16%)	39 (1.38%)	55 (1.94%)	203 (7.18%)
Social sciences	74 (2.62%)	100 (3.54%)	65 (2.30%)	72 (2.55%)	311 (11.00%)
Science	69 (2.44%)	62 (2.19%)	62 (2.19%)	69 (2.44%)	262 (9.26%)
Engineering	189 (6.68%)	222 (7.85%)	293 (10.36%)	251 (8.88%)	955 (33.77%)
Information technology	105 (3.71%)	101 (3.57%)	108 (3.82%)	163 (5.76%)	477 (16.87%)
Agriculture, life and environment sciences	96 (3.39%)	89 (3.15%)	89 (3.15%)	61 (2.16%)	335 (11.85%)
Medicine and pharmaceutical sciences	82 (2.90%)	84 (2.97%)	60 (2.12%)	59 (2.09%)	285 (10.08%)
Total	663 (23.44%)	719 (25.42%)	716 (25.32%)	730 (25.81%)	2,828

Upon completing the EPT, 1,646 test takers, more than 50% of the whole test-taking population in this administration, participated in a paper-based questionnaire survey on the fairness of the test on a voluntary basis. Following data entry, verification, and cleaning, the responses from 1,134 test takers were used for subsequent factor analysis. The demographic profiles of these test takers are

presented in Table 3.4. The questionnaire survey included 723 males (63.76%) and 411 females (36.24%), aged between 18 and 25. Most participants (82.80%) were in their second or third year of study. 17.02% of the participants were from arts and social sciences disciplines, whereas the others majoring in sciences, engineering, or information technology, *etc*. As shown in Table 3.4, the majority of test takers (83.69%) took the test for the first time, some (11.64%) took it twice. Less than 5% of the test takers took the test three or more times.

A purposive sampling strategy (Dörnyei, 2007) was employed to recruit test takers for the semi-structured interview. Considering the potential impact of demographic characteristics on the participants' test-taking experiences and their perceptions of test fairness, inclusion criteria were carefully designed to account for the test-takers' demographic information. To ensure a balanced representation, 20 test takers among those who volunteered for the interview were included in the study. The inclusion criteria included factors such as gender, grade, academic background, and the number of test attempts. Overall, the test takers who participated in the interviews reflect a well-balanced demographic profile of the test-taking population of the EPT (see Table 3.5).

Table 3.4 Demographic profiles of the test takers in the questionnaire survey.

	Frequency	Percent (%)	Cumulative percent (%)
Gender			
Male	723	63.76	63.76
Female	411	36.24	100
Age			
18	12	1.06	1.06
19	177	15.61	16.67
20	418	36.86	53.53
21	289	25.49	79.01
22	191	16.84	95.86
23	40	3.53	99.38
24	4	0.35	99.74
25	3	0.27	100
Program year			
2nd	519	45.77	45.77
3rd	420	37.04	82.80
4th	168	14.82	97.62
5th	22	1.94	99.56
Other	5	0.44	100
Field of study			
Humanities	78	6.88	6.88
Social sciences	115	10.14	17.02
Sciences	197	17.37	34.39
Engineering	341	30.07	64.46
Information technology	184	16.23	80.69
Agriculture, life and environment sciences	113	9.97	90.65
Medicine and pharmaceutical sciences	106	9.35	100
Number of test attempts			
Once	949	83.69	83.69
Twice	132	11.64	95.33
Three times	34	3.00	98.33
Four times	12	1.06	99.38
Five times	3	0.27	99.65
Other	4	0.35	100
Total	1,134	100	100

Table 3.5 Demographic profiles of the test takers in the semi-structured interview.

Interviewees*	Gender	Age	Program year	Field of study	Test attempt
TT1	Female	20	2nd	Humanities	1
TT2	Female	20	2nd	Humanities	1
TT3	Male	21	4th	Humanities	3
TT4	Male	18	2nd	Humanities	1
TT5	Male	20	3rd	Sciences	1
TT6	Male	22	3rd	Sciences	1
TT7	Male	20	3rd	Sciences	1
TT8	Male	22	4th	Sciences	2
TT9	Male	19	2nd	Sciences	2
TT10	Female	19	2nd	Sciences	1
TT11	Male	20	2nd	Sciences	1
TT12	Male	21	4th	Sciences	1
TT13	Female	20	3rd	Humanities	1
TT14	Male	21	3rd	Sciences	1
TT15	Male	22	3rd	Sciences	1
TT16	Female	23	4th	Sciences	1
TT17	Female	19	2nd	Sciences	1
TT18	Male	21	3rd	Sciences	1
TT19	Female	21	3rd	Humanities	1
TT20	Male	22	4th	Sciences	1

Note. To ensure participant anonymity, test takers were assigned unique alphanumeric identifiers consisting of "TT" followed by a number (e.g., TT1).

3.3.2 Teachers

The teachers (n = 6), aged between 44 and 57 (M = 49, SD = 5.18), were a convenience sample from the Foreign Languages Teaching Center at the university where the EPT was administered (see Table 3.6). They were experienced teachers responsible for the delivery of College English courses to non-English-major students, with an average of 26 years of teaching experience (SD = 7.40). Four of them held doctoral degrees and two had master's degrees; five were in linguistics or applied linguistics and one was in economics. In addition to their teaching responsibilities, they also contributed to the development of the EPT. They were involved in item writing and served as raters or oral examiners.

Table 3.6 Demographic profiles of the teachers in the semi-structured interview.

Interviewees*	Gender	Age	Teaching experience	Educational background	Roles regarding the EPT
TD1	Female	44	24 years	PhD in literature or linguistics	Item writer; oral examiner
TD2	Female	52	26 years	PhD in economics	Item writer; oral examiner
TD3	Female	45	19 years	MA in literature or linguistics	Item writer; oral examiner
TD4	Female	57	40 years	PhD in literature or linguistics	Item writer & reviewer; oral examiner
TD5	Female	45	21 years	PhD in literature or linguistics	Item writer; oral examiner; rater
TD6	Female	51	26 years	MA in literature or linguistics	Test blueprint designer; item writer; oral examiner

Note. To ensure participant anonymity, teachers (also test developers) were assigned unique alphanumeric identifiers consisting of "TD" followed by a number (e.g., TD1).

3.3.3 Test administrators

Two test administrators[1] were recruited for the semi-structured interviews using a convenience sampling approach. TA1, aged 59, was the Vice Dean of Academic Affairs at the university, with over 20 years of working experience. As a university-level test administrator, TA1 was instrumental in early-stage discussions regarding practicality issues related to administering the EPT. Additionally, she ensured the smooth administration of the test by coordinating test locations, rooms, and equipment and addressing complaints of the test takers. The other participant, TA2, aged 39, held a master's degree in education. She has had 16 years of working experience in a school-level management department for undergraduate education programs. TA2 served as one of the coordinators of the EPT. Her primary responsibilities included arranging test schedules, coordinating test locations, checking equipment (hardware and software) in the test rooms, managing registrations, and releasing test scores. She also helped with hiring and training oral

[1] To ensure participant anonymity, test administrators were assigned unique alphanumeric identifiers consisting of "TA" followed by a number (e.g., TA1).

examiners and proctors, resolving score review requests, and helping test takers who may have questions about the test or seek advice on preparing for it.

3.3.4 Test users

In this study, three test users[1] were invited for the interview. All of them are responsible for general English language teaching, with teaching experience ranging from 20 to 42 years. Two of them, TU1 and TU2, held doctoral degrees in linguistics and applied linguistics. They were experienced researchers in language testing and assessment, with substantial professional experience in developing and validating real-world language tests. They also delivered courses on language assessment and statistics for the graduate students in the applied linguistics program. TU3 had a master's degree in linguistics and applied linguistics. In addition to teaching, they all assumed administrative responsibilities as well. TU1 served as Vice President of the university at the time when the data were collected; TU2 was the Vice Dean of the school; and TU3 was the Head of the Foreign Languages Teaching Center. Regarding their roles in relation to the EPT, TU1 identified herself as the "initiator" and a key policymaker. TU2 was responsible for test assembly and post-test score equating. TU3, a senior item writer and reviewer, also managed the training of oral examiners and proctors.

3.4 Instruments

One of the key instruments for this study is the EPT which has been introduced in Section 3.2.2 and will not be elaborated in this section. This section will only introduce test fairness questionnaire for test takers and the semi-structured interview guide for the four groups of stakeholders.

[1] To ensure participant anonymity, test users were assigned unique alphanumeric identifiers consisting of "TU" followed by a number (e.g., TU1).

3.4.1 Test fairness questionnaire for test takers

A questionnaire was developed in this study to elicit test-takers' perceptions of the fairness of the EPT. Existing questionnaires or scales of test fairness reported in the literature (see Table 3.7) have revealed various dimensions of test fairness in the context of classroom-based assessment (Rezai, 2022; Wallace & Qin, 2021) or large-scale language tests (Fan, 2018; Jang, 2002; Li, 2021). Test fairness is also discussed within the domain of critical language assessment literacy (Tajeddin et al., 2022). However, the conceptualization of test fairness in these studies differs in one way or another from the present study. Furthermore, test-taker participants in this study might have difficulty responding to certain items about "test design", "score interpretation", and "validity" covered in existing instruments due to their limited assessment literacy. These concerns highlight the need for a tailored questionnaire that: (1) aligns well with the conceptualization of test fairness in this study and (2) includes items that are clear and accessible for the test-taker participants in this study. The development process and the structure of the questionnaire are detailed below.

Table 3.7 Summary of existing scales and questionnaires relevant to test fairness.

Author(s)	Instrument	Number of items	Scale	Constructs/Dimensions
Fan (2018)	A student perception questionnaire on English language testing practices in China	29	Five-point Likert-type scale	- Test design - Test information - Test impact - Test administration - Test fairness
Jang (2002)	A test fairness perception questionnaire	25	Five-point Likert-type scale	- School policy - Test bias - Test administration - Validity
Li (2021)	A scale for evaluating fairness of language testing	20	Five-point Likert-type scale	- Absence of bias - Score interpretation - Validity - Equal opportunity
Rezai (2022)	A student perception questionnaire on fairness in classroom-based assessment	110	Five-point Likert-type scale	- Learning materials and practices - Test design - Opportunities to demonstrate learning - Test administration - Grading - Offering feedback - Test result interpretation - Decisions based on test results - Test result consequences - Students' fairness-related beliefs and attitudes
Tajeddin et al. (2022)	A critical language assessment literacy scale	46	Five-point Likert-type scale	- Teachers' knowledge of assessment objectives, scopes, and types - Assessment use consequences - Fairness - Assessment policies - National policy and ideology
Wallace & Qin (2021)	A questionnaire on classroom fairness and justice in L2 assessment	18	Five-point Likert-type scale	- Distributive fairness - Procedural fairness - Interactional fairness - Entity justice

The development procedures followed the scale development guidelines proposed by DeVellis and Thorpe (2022). The multi-stage process used to develop the questionnaire is documented in the subsequent paragraphs (see Figure 3.2 for an overview).

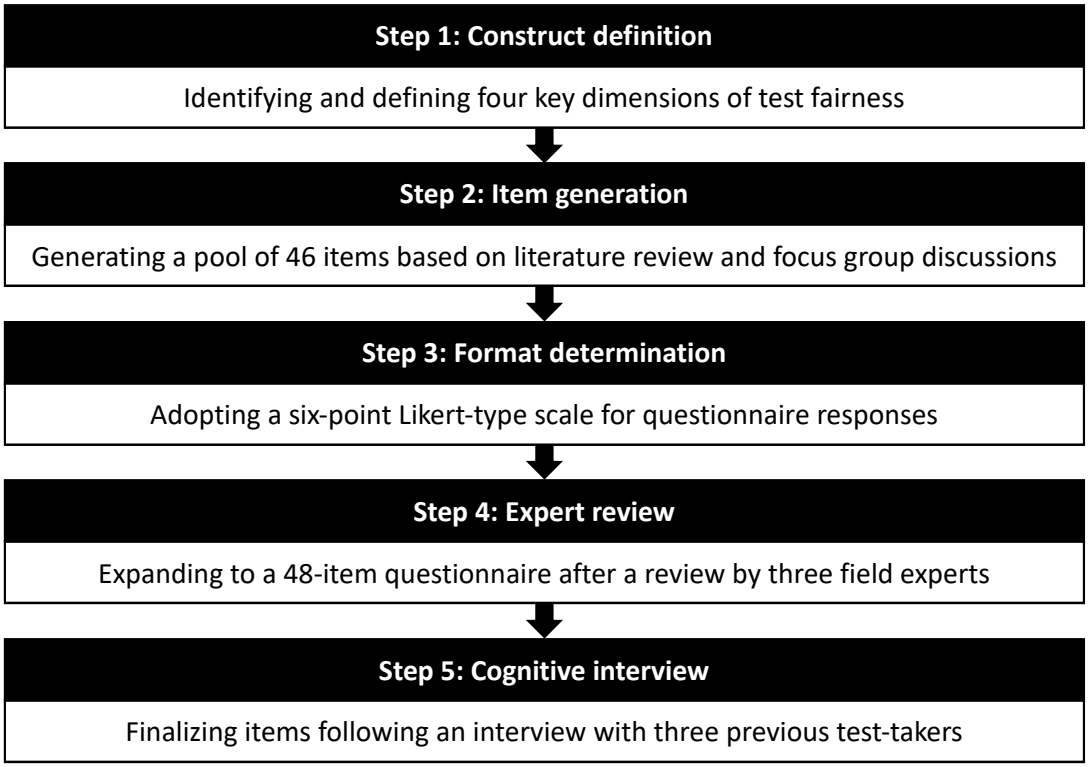

Figure 3.2 Overview of questionnaire development process.

Construct definition. Building on the dimensions of test fairness presented in Section 2.1.2 and the tentative conceptual model presented in Section 2.3, test fairness is characterized as a multifaceted concept encompassing four dimensions: comparability, accessibility, consistency, and accountability. These four dimensions of test fairness are well-supported by existing literature mentioned in Section 2.1.2. In designing the questionnaire, "comparability" refers to the extent to which task types, test content, and administration conditions provide comparable opportunities for all test takers or test-taker subgroups to demonstrate their true language proficiency levels without being unfairly advantaged or disadvantaged by construct-irrelevant factors. "Accessibility" covers aspects such as the availability of test-

related information, opportunities for test takers to prepare for the test, and test-takers' access to test locations as well as the hardware and software for test delivery. "Consistency" in this study is defined as the uniformity of test administration conditions and procedures. Lastly, "accountability" refers to the mechanisms and procedures that allow test takers to raise concerns regarding various aspects of the test or testing practices.

Item generation. In this step, a 46-item pool was developed, grounded in the four dimensions of test fairness outlined above. The items were developed based on published academic sources and insights from focus group discussions. Initially, 35 items were drafted in Chinese, drawing on the literature that supports these dimensions and includes specific items designed to elicit test-takers' perceptions of test fairness. To ensure clarity, testing terminologies in the questionnaire were supplemented with explanations or examples. The item pool was expanded to include 46 items based on individual focus group discussions involving five MA students, four PhD candidates, and four college EFL teachers from an institute of applied linguistics. The graduate students involved have a research interest in language testing and assessment. All the teachers held PhD degrees and had research experience in language testing and assessment. All the participants were familiar with the context of the EPT. The teachers had extensive College English teaching experience, which provided them with a deep understanding of the target test takers. Some of them were also involved in test development. The graduate students assisted with test administration and therefore knew the testing procedures very well. As demonstrated, the focus group participants were highly qualified to develop questionnaire items that were both relevant and accessible for the test takers. Each focus group discussion began with an introduction of the research aim and the purpose. The discussions were guided by the following questions:

- How do you understand test fairness?
- What do you think are the characteristics of a fair language test?
- How do you view the fairness of the EPT?

- In what aspects do you consider the test fair/unfair? Could you elaborate on why you perceive them as fair/unfair?

Notes were taken by the researcher during each focus group discussion. After all discussions were completed, keywords in the notes were carefully examined. This process uncovered key aspects of test fairness, which were further categorized into four dimensions of test fairness specified above (i.e., comparability, accessibility, consistency, and accountability). Based on these key aspects, questionnaire items were then formulated.

Format determination. The questionnaire employed a six-point Likert-type scale, ranging from "strongly disagree" to "strongly agree" (1 = "strongly disagree", 2 = "disagree", 3 = "slightly disagree", 4 = "slightly agree", 5 = "agree", and 6 = "strongly agree"). An even number of responses was chosen to eliminate neutral options like "neither agree nor disagree" or "not sure" (DeVellis & Thorpe, 2022). Test takers were asked to rate their agreement or disagreement with each statement by selecting the appropriate option on the provided scale.

Expert review. The preliminary questionnaire items were reviewed by a panel of three highly experienced professors specializing in language testing and assessment research. Each of them independently evaluated the items, identified any ambiguities or redundancies, provided suggestions for item rewording, and proposed new items where needed. This review process led to the addition, removal, and refinement of items, ultimately enhancing the items' relevance, clarity, and comprehensiveness. Upon completion of the review, the professors' suggestions were incorporated, yielding a 48-item questionnaire (see Appendix 3).

Cognitive interviewing[1]. For this study, two sophomore students and one senior student were recruited to participate in the interview. All participants had taken the EPT within one year prior to the interview. During the interview, participants carefully read each questionnaire item and provided ratings on a scale from 1

[1] Cognitive interviewing is a qualitative method for determining how potential respondents interpret and respond to each item in an instrument (DeVellis & Thorpe, 2022).

("Strongly disagree") to 6 ("Strongly agree"). They were asked to describe their understanding of each item and provide an explanation of their ratings. Participants were encouraged to highlight any uncertainties or confusions they encountered with each item. Additionally, their remarks and suggestions regarding the items were documented. Based on the feedback collected during the interview, minor modifications to the wording of the questionnaire items were made.

The questionnaire for test takers consists of two parts (see Appendix 3). The first part was designed to collect test-takers' demographic information, including gender, age, major, grade, and the number of test attempts. The second part examines test-takers' perceptions of the fairness of the EPT across four dimensions: comparability, accessibility, consistency, and accountability. Table 3.8 details the preliminary dimensions of test fairness alongside the corresponding item numbers in the administered version of the questionnaire. For each of the 48 statements in the second part of the questionnaire, test takers were asked to indicate their level of agreement or disagreement using a six-point Likert scale, ranging from 1 ("strongly disagree") to 6 ("strongly agree"). For the dimension of accessibility, 18 items assessed the transparency of test-related information prior to test-taking (Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12), the availability of learning and test preparation opportunities (Items 13, 14, 15, and 16), and access to the test location (Items 17 and 18). For the dimension of comparability, ten items (Items 28, 29, 30, 31, 32, 33, 34, 35, 36, and 37) were designed to investigate whether test takers thought they are given equal opportunities to demonstrate their English proficiency. The consistency dimension was assessed through 15 items (Items 19, 20, 21, 22, 23, 24, 25, 26, 27, 38, 39, 40, 41, 42, and 43) aimed at capturing test-takers' perceptions of the uniformity in test administration procedures. Lastly, five items (Items 44, 45, 46, 47, and 48) were included in the questionnaire to examine the test-takers' views on the availability of opportunities to express their concerns regarding testing practices.

Table 3.8 Hypothesized questionnaire dimensions and corresponding item numbers.

Dimensions	Item numbers (see Appendix 3)
Comparability	28, 29, 30, 31, 32, 33, 34, 35, 36, 37
Accessibility	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
Consistency	19, 20, 21, 22, 23, 24, 25, 26, 27, 38, 39, 40, 41, 42, 43
Accountability	44, 45, 46, 47, 48

3.4.2 Semi-structured interview guides

To address RQ2 (i.e., "To what extent is the EPT fair from a multi-stakeholder perspective?"), two interview guides, one for test takers and one for the other three stakeholder groups, were developed to gain a thorough understanding of the stakeholders' perceptions of test fairness.

The test-taker interview guide (see Appendix 4 for details) included four parts. Part 1 was designed to build rapport with participants (including clarifying the study's purpose, outlining the interview procedures and expected duration, explaining participants' rights, and obtaining participants' consent). Test takers were invited to introduce themselves by providing details such as their current grade, major, experience in learning English, test-taking history with the EPT (e.g., number of attempts and test results), and the importance of English proficiency in their post-graduation aspirations. Part 2 included three general questions aiming to encourage test takers to share their thoughts and feelings about the fairness of the EPT. These open-ended questions aimed to uncover test-takers' overall perceptions of test fairness and delve into the reasoning behind their viewpoints. Informed by the tentative conceptual model (see Section 2.3), Part 3 organized the interview questions into four dimensions of test fairness: comparability, accessibility, consistency, and accountability. Part 4 included a five-item post-interview questionnaire survey designed to gather additional information about the test-takers' family and educational background. This information was intended to facilitate the analysis and interpretation of the interview data, as the test-takers' family and educational background might influence their views on test fairness. The interview guide

concluded with a closing statement encouraging participants to share any final thoughts, questions, or comments about the interview or the study.

The interview guide for teachers, test administrators, and test users included three parts (see Appendix 5 for details). Part 1 aimed to build rapport by explaining the study's purpose, describing interview procedures and duration, outlining participants' rights, and securing consent. Participants were invited to introduce themselves, sharing information such as their educational background, work experience, and their roles and responsibilities regarding the EPT. Part 2 included three general questions aimed at exploring participants' overall perceptions of the fairness of the EPT. Informed by the four key dimensions of test fairness (see Section 2.3), Part 3 organized questions around comparability, accessibility, consistency, and accountability. A closing statement at the end of the interview guide encouraged participants to provide any questions, comments, or final thoughts they had regarding the interview or the study.

3.5 Data collection

The quantitative and qualitative data were collected concurrently.

3.5.1 Quantitative data collection

3.5.1.1 Administration of the EPT

The written test of the EPT, including the listening, reading, and writing subtests was administered to 2,828 test takers across four sessions. Four test forms, with an anchor-item design, were used in the four sessions held at different times of the same day.

3.5.1.2 Administration of the questionnaire

Toward the end of each test session, a paper-based questionnaire was administered to explore the test-takers' perceptions of the fairness of the EPT. In each test room, the two proctors informed the test takers that participation in the questionnaire survey was voluntary. Among the 2,828 test takers, 1,646 agreed to participate in the survey. Six MA students helped to input the questionnaire responses into a Microsoft Excel

spreadsheet. They cross-checked each other's entries, and the researcher further verified each input to ensure accuracy.

3.5.1.3 Collection of performance data, test materials, and test-takers' demographic information

Test-takers' performance data, their background information, and test materials were accessed with consent after the test administration. Specifically, the following materials were provided by the Research Center for Language Development and Assessment at the university: (1) item-level performance data and raw scores, (2) four test papers (one for each session), and (3) audio recordings for the four test forms of the listening subtest. Additionally, demographic information of the test takers, including their grade, gender, and majors, was provided by the university's Academic Affairs Office. The raw scores and demographic information would be used to identify potential DIF, DBF, and DTF in the listening and reading subtests.

3.5.2 Qualitative data collection

To examine the stakeholders' perceived fairness of the EPT (RQ2), 31 one-on-one semi-structured interviews were conducted with 20 test takers, six teachers (also test developers in this study), two test administrators, and three test users. Participation in the interviews was voluntary. Prior to the interviews, potential participants were provided with an information sheet (Appendix 6) outlining the purpose of this study, researcher details, interview procedures, and so forth. Written consent was then obtained from those willing to participate in the interviews (See Appendix 7 for the consent form).

Due to COVID-19 restrictions, 17 interviews were conducted online via Tencent Meeting, while the remaining interviews were conducted face-to-face. Each interview was structured around an interview guide (see Appendix 4 & 5) and audio-recorded for subsequent thematic analysis. All interviews were conducted in Chinese, with a median duration of 53 minutes.

Following the interviews, the audio recordings were transcribed verbatim by the researcher. The transcripts were verified for accuracy by a PhD student who replayed the recordings and corrected any inaccuracies. Subsequently, all transcripts were sent back to the participants for their review and revision. The interview excerpts presented in Section 4.2.2 were translated from Chinese into English and were edited slightly to enhance readability while keeping the original meaning unchanged.

3.6 Data analysis

3.6.1 Quantitative analysis

3.6.1.1 DIF, DBF, and DTF

In this study, DIF is defined as the difference in item performance between humanities and science test-taker groups after controlling for their true latent English proficiency. DBF refers to the difference in testlet performance between the two groups with the same latent English proficiency. A "testlet" is a cluster of items based on a common input material (Wainer & Kiely, 1987). For example, in the listening subtest of the EPT, the five items based on the same long conversation form a testlet. As detailed in Table 3.1, there are four five-item testlets in the listening subtest. The reading subtest includes two five-item testlets and one ten-item testlet. DTF is the difference in overall test performance between the two groups of test takers with the same latent English proficiency. DTF is present when the listening and reading subtests exhibit a lack of measurement invariance across these two groups.

The dataset used for DIF, DBF, and DTF analyses comprised raw scores from 2,828 test takers on the listening and reading subtests of the EPT. The test takers were divided into two groups based on their fields of study. The humanities group included 505 test takers (17.86%), while the science group consisted of 2,323 (82.14%) test takers. Throughout the investigation of DIF, DBF, and DTF, the humanities group was designated as the focal group (i.e., a potentially disadvantaged group in DIF investigation), and the science group as the reference group. The distribution of test takers across the four test forms is detailed in Table 3.9.

Table 3.9 Summary of test-taker distribution by academic discipline and test form.

Group	Form 1	Form 2	Form 3	Form 4	Total
Humanities	119	158	103	125	505
Sciences	544	561	613	605	2,323
Total	663	719	716	730	2,828

Prior to conducting differential functioning analyses, descriptive statistics were calculated for the test performance of the humanities and science groups on the listening and reading subtests via IBM SPSS Statistics for macOS (Version 29.0; IBM Corp., 2024). To determine whether there were significant differences in test performance between the two groups, independent samples *t*-tests were conducted. However, preliminary analyses revealed violations of normality in the performance data. Subsequently, the homogeneity of variances was assessed using the Brown-Forsythe test due to its robustness against non-normal distributions in test performance data. Given the violations of both assumptions, the non-parametric Mann-Whitney *U* test was employed to examine performance differences between the two groups across all test forms of both subtests.

DIF and DBF analyses. To identify item-level DIF and testlet-level DBF in the listening and reading subtests, the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993) was performed using the *mirt* package (Chalmers, 2012) in R (Version 4.3.2; R Core Team, 2023). The SIBTEST was selected as opposed to other DIF and DBF detection methods for several reasons. First, as a non-parametric procedure, the SIBTEST is particularly useful when addressing potential violations of local independence assumption in testlet-based tests (Lee et al., 2009; Min & He, 2020; Song et al., 2015). Second, the SIBTEST can be used to detect both DIF and DBF. Third, unlike observed score methods, the SIBTEST uses a regression estimate of the true score to match test takers on their English proficiency levels (Shealy & Stout, 1993). This allows the SIBTEST to account for the effect of measurement errors on the DIF statistic, thereby reducing statistical bias in DIF detection (Noble et al., 2023). Additionally, the SIBTEST can be used to identify both uniform and nonuniform DIF. Uniform DIF occurs when an item consistently favors one group of test takers over

another across all ability levels, while nonuniform DIF arises when an item alternately favors one test-taker group at different ability levels (Li & Stout, 1996). More importantly, the SIBTEST procedure is robust under the following conditions: (1) the sample size of test takers is small (e.g., $n = 100$); and (2) there are distributional differences in the construct being measured across test-taker groups (Roussos & Stout, 1996b). Lastly, an established effect size guideline exists to assist in determining the magnitude of DIF (Roussos & Stout, 1996a).

A two-step analysis was performed to detect DIF and DBF across each test form of the listening and reading subtests. In the first step, a standard one-item-at-a-time DIF anchor purification procedure was conducted to identify a set of DIF-free anchor items. Specifically, for each form of the listening subtest, each of the 30 items was sequentially treated as a suspect item, while the others served as a temporary matching subtest. Items that did not exhibit DIF were subjected to another round of anchor purification, still, through the one-item-at-a-time DIF analysis. This purification procedure was repeated until a stable set of DIF-free items was identified. The DIF-free anchor items then constituted the final matching subtest used to match the abilities of the test takers from the humanities and science groups. The same purification procedure was conducted on the 20 items in each form of the reading subtest. In the second step, the items and testlets in each form of the listening and reading subtests were examined for DIF and DBF using the final matching subtest identified for each form. In the listening subtest, for each test form, the 10 stand-alone short conversation items were examined for DIF, while the four testlets were examined for DBF. Similarly, in the reading subtest, the three testlets for each test form were scrutinized for DBF.

The presence and effect of DIF and DBF are determined by: (1) the results from a significance test and (2) the direction and magnitude of the $\hat{\beta}_{\text{uni}}$ statistic. According to the guidelines proposed by Roussos and Stout (1996b), DIF/DBF is considered present when the null hypothesis, which posits that no DIF/DBF exists, is rejected. A positive $\hat{\beta}_{\text{uni}}$ value indicates that the reference group (i.e., the science group) is favored over the focal group (i.e., the humanities group) along the ability continuum

of the test takers on the items/testlets flagged for DIF/DBF. Specifically, a $|\hat{\beta}_{\text{uni}}|$ value of less than .059 suggests negligible DIF; a range of $.059 \leq |\hat{\beta}_{\text{uni}}| < .088$ indicates moderate DIF; and a $|\hat{\beta}_{\text{uni}}|$ value of .088 or greater signifies large DIF. Currently, there is no established effect size guidelines for determining the magnitude of DBF.

DTF analysis. To explore whether the cumulative effect of item-level DIF and testlet-level DBF lead to bias at the test level, a multiple-group confirmatory factor analysis (MG-CFA) was performed using the *lavaan* package (Rosseel, 2012) in R (Version 4.3.2; R Core Team, 2023). MG-CFA is a commonly used technique for examining measurement invariance, which can determine whether the same construct is consistently measured across test-taker groups with different fields of study. Measurement non-invariance indicates the presence of DTF. A two-step procedure was followed to detect DTF: (1) establishing and estimating the baseline CFA model for the listening and reading subtests and (2) examining measurement invariance between the humanities and science groups.

The baseline CFA model specifies two distinct yet correlated latent factors—listening comprehension and reading comprehension—each measured by multiple testlet-level observed variables (see Figure 3.3). Each variable represents an item bundle or a testlet. Specifically, the ten short conversation (SC) items constitute an item bundle, whereas the items associated with the same listening or reading input form testlets (e.g., LC, LP1, RP1, BC). These variables were treated as polytomous items, with scores reflecting the sum of the individual items associated with each item bundle or testlet. The CFA model was tested for the entire sample, as well as for the humanities and science groups. The Maximum Likelihood estimation with Robust standard errors (MLR) was used to estimate model parameters. The Satorra-Bentler scaled Chi-square test was used to adjust for non-normality in the data. The model fit was evaluated using the following criteria: (1) a non-significant Satorra-Bentler scaled Chi-square ($SB\chi^2$) goodness-of-fit statistic, (2) the comparative fit index (CFI) $\geqslant$ .90 (Hu & Bentler, 1995), (3) the root mean square error of approximation (RMSEA) $\leqslant$ .08 (Hu & Bentler, 1999), and (4) the standardized root mean square residual (SRMR) $\leqslant$ .08 (Hu & Bentler, 1999). The chi-square test is easily affected

by sample size. In cases of large sample sizes, the test may indicate significance even when the CFA model fits the data reasonably well (Babyak & Green, 2010; Bollen, 1989). Therefore, the model fit was evaluated using this statistic alongside other fit indices.

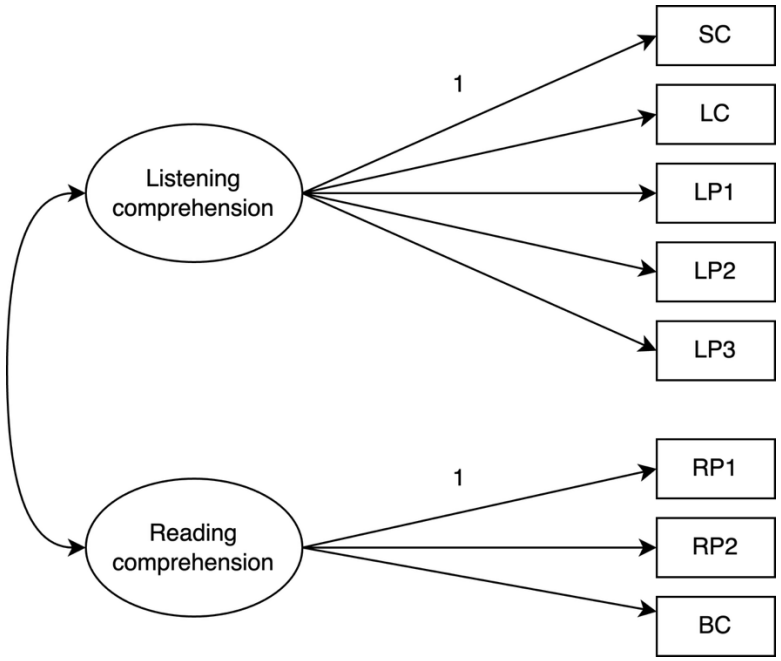

Figure 3.3 CFA model for the listening and reading subtests.

Measurement invariance testing was conducted for the listening and reading subtests of the EPT to evaluate the equivalence of the test structure across the humanities and science groups. The analyses began with testing for full measurement invariance based on the baseline CFA model. Four increasingly restrictive measurement invariance models were examined sequentially: the configural invariance model, the metric invariance model, the scalar invariance model, and the strict invariance model. Configural invariance examines whether the same general factor structure could be identified across test-taker groups, without imposing equality constraints on model parameters. Building on configural invariance, metric invariance was tested to determine whether the constructs measured by the subtests could be interpreted in the same way across the two groups. The metric invariance model constrained the factor loadings to be equal across the two test-taker groups.

Upon confirming metric invariance, the scalar invariance model was tested by additionally adding the equality constraint of item intercepts. Establishing scalar invariance enabled meaningful comparisons of latent means between the humanities and science groups. Lastly, the strict invariance model was examined by constraining the item residual variances to be equal across groups, in addition to the constraints imposed in the scalar invariance model. Strict invariance implies that the amount of item variance not accounted for by the latent factor is equivalent across groups. This level of invariance suggests that the scale measures the construct with the same degree of precision in both groups and is often considered a prerequisite for comparing observed scores across groups. It should be noted that, to ensure score comparability across test-taker groups, establishing scalar invariance is a minimum requirement (Ercikan & Por, 2020). The determination of invariance at each level was based on the model fit indices (i.e., CFI, RMSEA, and SRMR) and the change in the CFI value between the nested models. A change in the CFI (ΔCFI) of less than .01 between the less restricted and more restricted models was considered evidence of invariance (Cheung & Rensvold, 2002).

3.6.1.2 Characteristics of the input materials

To address RQ2.2 (i.e., "To what extent are the input materials comparable in terms of difficulty across different test forms, as indicated by their input characteristics?"), a series of statistical analyses were performed on the characteristics of the listening and reading input materials used across the four test forms.

I. Characteristics of the listening input materials

A review of the literature indicates that both textual and acoustic characteristics of listening input materials are related to the difficulty of listening tasks (e.g., Brunfaut & Révész, 2015). In this study, the characteristics of the input materials were operationalized as their lexical complexity, syntactic complexity, discourse complexity, and speed of delivery.

Table 3.10 provides an overview of all the selected measures and the analysis tools or procedures. The 27 measures were derived from the input materials used across all the four test forms. Each test form consisted of 10 short conversations, one long conversation, and three passages. As a result, each test form yielded 14 observations for each of the 27 measures.

Lexical characteristics. In terms of lexical complexity, this study adopted Brunfaut and Révész's (2015) analytical framework, focusing on measures of lexical frequency, diversity, density, and the concreteness of content words. Lexical frequency measures were obtained from VocabProfiler (Web VP Classic v.4 version; Cobb, n.d.), including the proportion of K1 words, K1 function words, K1 content words, K2 words, K1 + K2 words[1], academic words[2], and off-list words[3]. The Moving Average Type-Token Ratio (MATTR), a measure of lexical diversity, was extracted using TAALED 1.4.1 (Kyle et al., 2021). MATTR was chosen among various lexical diversity measures because it is less sensitive to or independent from text-length (Bestgen, 2024; Fergadiotis et al., 2015; Sung et al., 2024; Zenker & Kyle, 2021). Lexical density, an indicator of information processing difficulty (Bloomfield et al., 2011; Révész & Brunfaut, 2013), was measured as the proportion of content words to the total number of words. This measure was also obtained using VocabProfiler (Web VP Classic v.4 version; Cobb, n.d.). Concreteness of content words was computed using Coh-Metrix 3.0 (McNamara et al., 2014) to indicate the cognitive demands placed on test takers during the listening subtest.

[1] K1 and K2 words refer to the first and second 1,000 most frequent word families in English based on the General Service List (GSL; West, 1953).

[2] Academic words are identified based on the Academic Word List (AWL; Coxhead, 2000).

[3] Off-list words are those not included in either the GSL or AWL.

Table 3.10 Measures used to operationalize the characteristics of the listening input.

Dimension	Measure	Analysis tool/ procedure
Lexical characteristics		
Lexical frequency	Proportion of:	VocabProfiler (Web VP Classic v.4 version)
	• K1 words	
	• K1 function words	
	• K1 content words	
	• K2 words	
	• K1 + K2 words	
	• academic words	
	• off-list words	
Lexical diversity	Moving average type-token ratio (MATTR)	TAALED 1.4.1
Lexical density	Proportion of content words to the total number of words (CW/W)	VocabProfiler (Web VP Classic v.4 version)
Concreteness	Mean concreteness value for all content words	Coh-Metrix 3.0
Syntactic characteristics		
Length-based	Mean length of sentence (MLS)	Web-based L2SCA
	Mean length of clause (MLC)	
	Mean length of AS-unit (ML-AS)	Manually calculated
Phrase-level	Left embeddedness (Mean number of words before main verb)	Coh-Metrix 3.0
	Modifiers per noun phrase (M/NP)	
	Complex nominals per clause (CN/C)	Web-based L2SCA
	Complex nominals per AS-unit (CN/AS)	Manually calculated
Clause-level	Clauses per AS-unit (C/AS)	Manually calculated
	Clauses per sentence (C/S)	Web-based L2SCA
Discourse characteristics		
Local-level	Incidence score of all connectives	Coh-Metrix 3.0
	Semantic similarity (between sentences)	
Global-level	Semantic similarity (between paragraphs)	
Text-level	Incidence score of causal content	
	Score of temporal cohesion	
	Semantic similarity (across text)	
Speed of delivery	Words per minute	Manual calculation
	Speech rate (syllables per second)	Text Inspector & manual calculation

Syntactic characteristics. The researcher employed length-based, phrase-level, and clause-level measures to capture the syntactic complexity of the listening input materials. Length-based measures included the mean length of sentences, clauses, and the Analysis of Speech Units (AS-units). The first two measures were calculated using the web-based L2 Syntactical Complexity Analyzer (L2SCA; Lu & Ai, 2015), while the mean length of AS-units was manually coded based on the definition of AS-unit (see Foster et al., 2000) and calculated by the researcher. Phrase-level complexity was evaluated by measuring left embeddedness, modifiers per noun phrase (M/NP) and complex nominals per clause (CN/C) with Coh-Metrix 3.0 (McNamara et al., 2014). Left embeddedness is defined as the mean number of words before the main verb. This measure was selected because it is related to the cognitive load imposed by a sentence. Specifically, a higher number of words preceding the verb increases information density and sentence ambiguity (Graesser et al., 2004). Additionally, complex nominals per AS-unit (CN/AS) were calculated by the researcher. Clause-level complexity was assessed using two measures: clauses per AS-unit (C/AS) and clauses per sentence (C/S).

Discourse characteristics. The discourse complexity of the listening input materials was analyzed using local-, global-, and text-level measures obtained through Coh-Metrix 3.0 (McNamara et al., 2014). At the local level, the incidence score of all connectives (including causal, logical, adversative, contrastive, temporal, additive, positive, and negative connectives) and the semantic similarity between adjacent sentences were computed. Global-level cohesion was assessed by calculating the semantic similarity between paragraphs. Text-level cohesion was measured by the incidence score of causal verbs and particles (i.e., causal content), temporal cohesion, as well as semantic similarity across the text.

Speed of delivery. The speed of delivery were measured by words per minute and speech rate. Words per minute were calculated manually to provide a coarse measure of delivery speed. Speech rate, defined as the number of syllables per second, was obtained using a combination of Text Inspector and manual calculations. Text Inspector was used to count the number of syllables in the listening input.

As an initial step, descriptive statistics were computed for the input characteristics, operationalized through 27 measures across four dimensions: lexical characteristics, syntactic characteristics, discourse characteristics, and speed of delivery. The descriptive analyses were conducted using IBM SPSS Statistics for macOS (Version 29.0; IBM Corp., 2024).

Subsequently, to examine whether significant differences existed among the input characteristics across test forms, a multivariate Kruskal-Wallis test and a permutation test with 10,000 iterations were performed using the *npmv* package (Version 2.4.0; Burchett et al., 2017) in R (Version 4.3.2; R Core Team, 2023). The choice of these nonparametric tests was mainly informed by the relatively small sample size in this study (n = 56, with 14 observations per test form). The sample size in this study did not meet several recommended thresholds for conducting parametric multivariate analyses such as Multivariate Analysis of Variance (MANOVA). To ensure robustness, it is advisable to have a sample size of around 20 observations per group (Mardia, 1971). However, the current study included only 14 observations per test form. Additionally, an *a priori* power analysis for a one-way MANOVA, assuming a medium effect size (partial $\eta^2 = 0.06$), an alpha level of 0.05, and a desired statistical power of 80%, suggests a minimum sample size of 212 to reliably detect significant group differences. Considering these factors, the multivariate Kruskal-Wallis test, a nonparametric alternative to one-way MANOVA, was deemed appropriate due to its suitability for small sample sizes. To strengthen the robustness of the analysis, especially given the relatively small sample size, the multivariate Kruskal-Wallis test was complemented with a permutation test. The permutation test was effective with small sample sizes and could provide reliable significance estimates without relying on strict distributional assumptions (Burchett et al., 2017; Good, 2013).

Following the multivariate analyses, separate Kruskal-Wallis tests were conducted for each of the 27 measures to identify specific input characteristics that differed significantly across the four test forms. These univariate analyses were deemed necessary for several reasons. First, they allowed for a more nuanced

understanding of how each linguistic feature varies across test forms. While the multivariate Kruskal-Wallis test was used to identify global differences across all dependent variables collectively, it did not indicate which measure differed significantly across groups. Second, the multivariate test lacked the sensitivity to detect variations in specific input characteristics, especially given this study's limited sample size. Conducting separate Kruskal-Wallis tests increased the likelihood of identifying significant differences in individual linguistic measures that might have been obscured in the multivariate analyses. Given the multiple comparisons involved in examining the 27 linguistic measures independently, the risk of Type I errors increased. To control for the false discovery rate in multiple comparisons, *p*-values were adjusted using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). The Kruskal-Wallis tests and subsequent *p*-value adjustments were performed using the *rstatix* package (Kassambara, 2023) in R (Version 4.3.2; R Core Team, 2023).

II. Characteristics of the reading input materials

Previous studies have shed light on how input characteristics influence test-takers' reading performance (e.g., Brunfaut, 2021; Eslami, 2014). Based on the insights drawn from these studies, the present study analyzed the lexical, syntactic, and discourse complexity, as well as the readability of the reading input materials in the four test forms. All selected measures and analysis tools or procedures are displayed in Table 3.11. In total, 42 measures were obtained for each of the four test forms. It should be noted that each test form included three passages. As a result, three observations were obtained from each test form for each of the 42 measures.

Table 3.11 Measures used to operationalize the characteristics of the reading input.

Dimension	Measure	Analysis tool/ procedure
Lexical characteristics		
Lexical frequency	Proportion of:	VocabProfiler (Web VP Classic v.4 version)
	• K1 words	
	• K1 function words	
	• K1 content words	
	• K2 words	
	• K1 + K2 words	
	• Academic words	
	• Off-list words	
Lexical diversity	Moving average type-token ratio (MATTR)	TAALED 1.4.1
Lexical density	Proportion of content words to the total number of words (CW/W)	VocabProfiler
Concreteness	Mean concreteness value for all content words	Coh-Metrix 3.0
Syntactic characteristics		
Length-based	Mean length of sentence (MLS)	Web-based L2SCA
	Mean length of clause (MLC)	
	Mean length of T-unit (MLT)	
Phrase-level	Left embeddedness	Coh-Metrix 3.0
	Modifiers per noun phrase (M/NP)	
	Complex nominals per clause (CN/C)	Web-based L2SCA
	Complex nominals per T-unit (CN/T)	
	Coordinate phrases per clause (CP/C)	
	Coordinate phrases per T-unit (CP/T)	
Clause-level	Clauses per T-unit (C/T)	Web-based L2SCA
	Clauses per sentence (C/S)	
	Dependent clauses per clause (DC/C)	
	Dependent clauses per T-unit (DC/T)	
Discourse characteristics		
Local-level	Incidence score of all connectives	Coh-Metrix 3.0
	Incidence score of causal connectives	
	Incidence score of logical connectives	
	Incidence score of temporal connectives	
	Overlap between nouns (between sentences)	
	Overlap between arguments (between sentences)	

(*Continued*)

Table 3.11 (Continued).

Dimension	Measure	Analysis tool/ procedure
Local-level	Overlap between stems (between sentences)	Coh-Metrix 3.0
	Semantic similarity (between sentences)	
Global-level	Overlap between nouns (between paragraphs)	Coh-Metrix 3.0
	Overlap between arguments (between paragraphs)	
	Overlap between stems (between paragraphs)	
	Semantic similarity (between paragraphs)	
Text-level	Incidence score of causal content	Coh-Metrix 3.0
	Incidence score of intentional content	
	Score of causal cohesion	
	Score of intentional cohesion	
	Score of temporal cohesion	
	Semantic similarity (across text)	
Readability	Coh-Metrix L2 Readability Index	Coh-Metrix 3.0

Lexical characteristics. The measures used to capture the lexical complexity of the reading input materials were identical to those used for examining the lexical complexity of the listening input materials. Therefore, relevant measures and analytical procedures will not be detailed here.

Syntactic characteristics. The syntactic complexity of the reading input materials was analyzed in terms of length-based, phrase-level, and clause-level characteristics. Length- based measures included the mean length of sentences (MLS), clauses (MLC), and T-units (MLT). At the phrase level, measures such as left embeddedness, modifiers per noun phrase (M/NP), complex nominals per clause (CN/C) and per T-unit (CN/T), as well as coordinate phrases per clause (CP/C) and per T-unit (CP/T), were obtained. In terms of clause-level measures, clauses per T-unit (C/T), clauses per sentence (C/S), dependent clauses per clause (DC/C), and dependent clauses per T-unit (DC/T) were used. Two measures, left embeddedness and modifiers per noun phrase (M/NP), were extracted using Coh-Metrix 3.0

(McNamara et al., 2014), while the remaining measures were automatically obtained through the Web-based L2SCA (Lu & Ai, 2015).

Discourse characteristics. Cohesion measures reflecting the discourse complexity of the reading input materials were automatically extracted using Coh-Metrix 3.0 (McNamara et al., 2014). Local-level cohesion was assessed in terms of the incidence scores of all connectives, as well as causal (e.g., *because*), logical (e.g., *or*), and temporal (e.g., *after*) connectives between adjacent sentences. Overlap between nouns, arguments, and stems, as well as semantic similarity between sentences, were also included. Similarly, global-level cohesion measures included the overlap between nouns, arguments, and stems, along with semantic similarity between paragraphs. At the text level, the incidence scores of causal and intentional content, the score of causal, intentional, and temporal cohesion, and semantic similarity across the text were computed.

Readability. The readability of the reading materials was assessed using the Coh-Metrix L2 Readability Index. According to Crossley et al. (2011), this measure was tailored specifically to mirror text readability for second language (L2) learners.

To examine potential differences in the characteristics of the reading input materials across the four test forms, descriptive analyses were conducted on the 42 measures using IBM SPSS Statistics for macOS (Version 29.0; IBM Corp., 2024). Due to the limited number of observations in each test form, inferential statistical analyses were not performed. Specifically, for each linguistic measure, the number of observations per test form was limited to three, corresponding to the three passages in the reading subtest: Reading Passage 1 (RP1), Reading Passage 2 (RP2), and a passage for the banked cloze (BC) task. The small sample size per group (n = 3) precluded the use of most parametric and non-parametric inferential tests, as these tests typically require larger sample sizes to ensure statistical power and reliability. As a result, inferential statistical analyses were deemed inappropriate for this dataset, and only descriptive statistics were reported to summarize the linguistic characteristics of the reading materials across the four test forms.

3.6.1.3 Questionnaire survey

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were conducted using IBM SPSS Statistics for macOS (Version 29.0; IBM Corp., 2024) to identify and validate the underlying structure of the test fairness questionnaire. Prior to the analysis, the dataset was cleaned and organized. Out of the 1,646 voluntary responses, 331 were discarded due to missing values, resulting in a valid sample of 1,315 (79.89% valid response rate). Negatively worded items in the questionnaire were reverse-coded to ensure a consistent response direction across all items. It should be cautioned that outliers in the dataset can potentially distort factor extraction process in EFA (e.g., Liu et al., 2012). Therefore, the dataset was examined for potential outliers. The Mahalanobis distance was calculated for each questionnaire item and compared against a chi-square distribution with degrees of freedom equal to the total number of variables in the dataset (Mahalanobis, 1936). A total of 181 observations, each with a Mahalanobis distance exceeding 84.17 at a significance level of .001, were identified as outliers and subsequently removed from further analyses. After data cleaning, 1,134 valid samples remained.

The cleaned dataset was randomly divided into two samples for EFA (n = 567) and CFA (n = 567). The homogeneity of the two samples was examined using a chi-square test of independence (for academic background) and Mann-Whitney U tests (for responses on the Likert-scale test fairness questionnaire). The chi-square test of independence ($\chi^2(1) = 1.81$, $p = .18$) revealed no significant difference between the EFA ($n_{Humanities} = 88$, $n_{Science} = 479$) and CFA ($n_{Humanities} = 105$, $n_{Science} = 462$) samples. Subsequent Mann-Whitney U tests also showed no significant differences between the two samples' responses on any of the questionnaire items, with p values ranging from .07 to .96.

The EFA sample was scrutinized to ensure that it met the necessary assumptions for conducting an EFA. The sample size (n = 567) was deemed adequate according to prevailing standards for EFA: minimum sample sizes of 100 (Hair et al., 1995), 300 (Tabachnick & Fidell, 2013), or 500 (Comrey & Lee, 1992), as well as the recommended item-to-respondent ratio of 1:10 to 1:15 (Field, 2009). To further

substantiate the adequacy of the sample, the Kaiser-Meyer-Olkin (KMO) measure was calculated, yielding a value of .946, which exceeds the threshold of .90 and indicates excellent sampling adequacy (Field, 2009). Furthermore, Bartlett's test of sphericity produced a significant result (χ^2 = 17,086.98, df = 1,128, p < .001), confirming that the variables were sufficiently intercorrelated for factor analysis (Kaiser & Rice, 1974). Taken together, these preliminary analyses confirm the dataset's suitability for factor analysis.

The EFA sample was then analyzed with two objectives: (1) to explore the factor structure of the test fairness questionnaire and (2) to identify questionnaire items that might distort the factor structure. Principal Component Analysis (PCA) with Direct Oblimin rotation was employed to extract and rotate factors. PCA was selected over principal axis factoring due to its effectiveness in reducing the dimensionality of questionnaire data (Fabrigar et al., 1999). PCA maximizes the total variance explained, thereby enhancing questionnaire data interpretability. Given that the four proposed dimensions of test fairness (i.e., comparability, accessibility, consistency, and accountability) could potentially be intercorrelated, the oblique rotation method, Direct Oblimin, was deemed appropriate as it accommodates potential correlations between the factors. The decision on factor retention was informed by Kaiser's (1960) eigenvalue-greater-than-one criterion, the scree plot (Cattell, 1966), and parallel analysis (Horn, 1965). With the number of factors fixed at four, the EFA yielded four plausible factors, each with an eigenvalue greater than one. Additionally, both the scree plot and parallel analysis indicate a four-factor solution. Subsequent analyses involved removing items with poor loading performance that could potentially distort the intended four-factor structure. Items that did not load onto any factors were discarded from the dataset. A loading cutoff value of .40 was determined. Accordingly, items with loadings below .40 on a single factor or cross-loadings exceeding .40 on more than one factors, were removed. This process resulted in the removal of 15 items from the original 48-item dataset due to issues such as lack of loading (Items 3, 11, 14, 15, 16, 17, 18, and 22), cross loading (Items 2, 19, and 43), and unexpected loading (Items 6, 12, 38, and 39) that did not align with the

hypothesized questionnaire dimensions (see Table 3.8). After item screening, 33 items were retained for subsequent analysis.

CFA was conducted on the other sample (n = 576) to cross-validate the four-factor structure of the test fairness questionnaire extracted by the EFA. Three CFA models were specified, estimated, and compared to select the best model that fit the CFA sample. Model 1 posited a unidimensional structure, hypothesizing that a single latent factor of test fairness accounted for all item covariances. Model 2, a five-factor correlated-trait model, hypothesized that the five latent factors underlying the test-takers' questionnaire responses were correlated with each other. Model 3 is a second-order factor model with a higher-order general factor (i.e., test fairness) and five first-order factors extracted by EFA. WLSMV estimation method was adopted to obtain parameter estimates. It is important to note that the questionnaire data in this study was collected using a six-point Likert-type scale, meaning that the items in the questionnaire were measured on an ordinal scale. An inspection on the distribution of the dataset indicates a violation of both univariate and multivariate normal distribution. For the ordinal and non-normal type of data in this study, WLSMV is an appropriate estimation method because it does not assume normal distribution for the observed variables and is ideal for handling ordinal type of data (Li, 2016). Model fit was evaluated based on the following criteria: (1) the non-significant chi-square (χ^2) goodness-of-fit statistic, (2) the comparative fit index (CFI) of .90 or above (Hu & Bentler, 1995), (3) the root mean square error of approximation (RMSEA) of .08 or below (Hu & Bentler, 1999), and (4) the standardized root mean square residual (SRMR) of .08 or below (Hu & Bentler, 1999). As the chi-square statistic tends to be sensitive to sample size (e.g., Babyak & Green, 2010; Bollen, 1989), the decision on the best-fitting model also took the other three indices into consideration.

Following factor structure validation, the entire valid sample (n = 1,134) was used for subsequent analysis. To summarize test-takers' responses, descriptive statistics, including means, standard deviations, skewness, kurtosis, and reliability coefficients, were computed for each latent factor. These statistics provided an overview of how test takers perceived various aspects of the EPT's fairness.

3.6.2 Qualitative analysis

Thematic analysis (Braun & Clarke, 2006, 2021) was used to examine the stakeholders' perceptions of the EPT's fairness and identify the factors influencing their perceptions. The analysis was performed using MAXQDA Plus 24 for Mac (Release 24.5.0), following Braun and Clarke's (2021) analytical guidelines (see Figure 3.4).

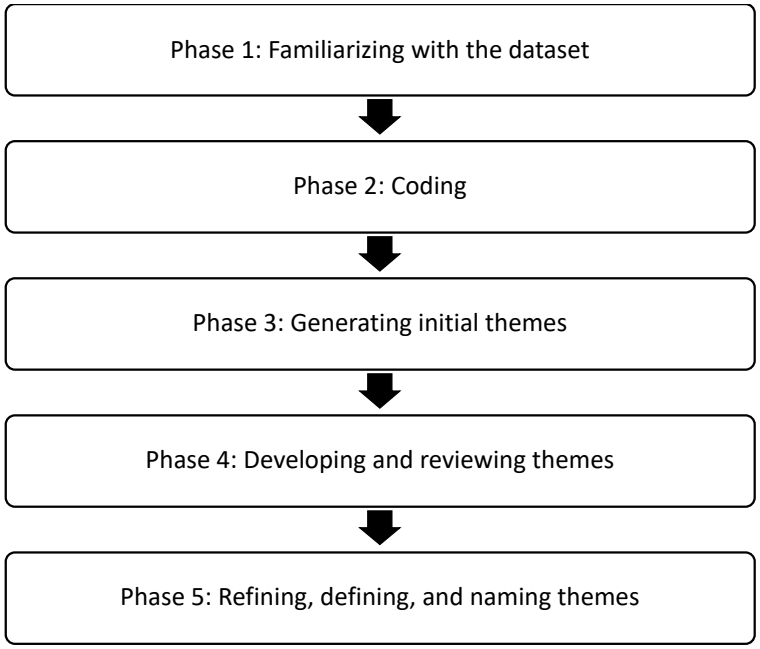

Figure 3.4 Analytical process of thematic analysis.

During the initial phase of analysis, the researcher engaged with the interview dataset by listening to the audio recordings, reading and re-reading the transcripts, paraphrasing content when necessary, and noting any insightful segments.

In the second phase, initial codes were developed and applied to identify data segments relevant to RQ2. For RQ2.1, both deductive and inductive coding strategies were employed to examine the stakeholders' perceptions of the EPT's fairness. Deductive coding was performed based on *a priori* codes derived from the tentative conceptual model integrating the theoretical underpinnings of test fairness and the contextual features of this study. Meanwhile, inductive codes emerged directly from

the interview data were also incorporated throughout initial coding. For RQ2.2, inductive coding was conducted to identify the factors influencing the stakeholders' perceptions of the EPT's fairness. To ensure coding reliability, the researcher and a PhD candidate (with research expertise in language testing and assessment) coded 50% of the interview transcripts independently, using the initial coding scheme. The inter-coder agreement reached 73.11% for the stakeholders' perceptions and 80.28% for the influencing factors, indicating substantial agreement between the two coders (see Hallgren, 2012). Disagreements on the coding categories were discussed and resolved through multiple meetings after the initial coding. After reaching consensus, the researcher coded the remaining 50% of the interview transcripts, using the finalized coding schemes (see Appendix 8 and 9).

In the third and fourth phase, themes were identified and reviewed. Codes with shared meanings were grouped into themes and sub-themes. The initial analysis generated four themes and 15 sub-themes related to the stakeholders' perceptions of the EPT's fairness, along with four themes and nine sub-themes about the factors influencing their perceptions. These candidate themes and sub-themes were then reviewed by the researcher and a professor specialized in language testing and assessment to ensure they: (1) aligned appropriately with the coded segments, (2) accurately represented the entire dataset, and (3) captured key aspects relevant to both RQ2.1 and RQ2.2.

In the fifth phase, the candidate themes and sub-themes were refined to ensure distinctiveness and representativeness, with some being merged, separated, or eliminated. The researcher then assigned names to the finalized themes.

It should be noted that the semi-structured interviews were conducted in Chinese. Thematic analysis, including coding and theme development, was performed using the Chinese transcripts. The results and relevant interview excerpts were translated in English and will be reported in Section 4.2.2.

3.7 Chapter summary

This chapter outlines the methodology used to address the two overarching research questions of this study, covering the research design, participants, instruments, data collection procedures, and data analysis techniques. A convergent mixed-methods design was adopted to assess the fairness of the EPT from the perspectives of psychometrics and stakeholders. This involved the use of both quantitative and qualitative methods. On the quantitative side, an analysis was conducted on test-takers' item-, testlet-, and test-level performance in the listening and reading subtests to identify potential DIF, DBF, and DTF. Additionally, the characteristics of input materials across different test forms were analyzed to assess the comparability of their difficulty levels. A questionnaire survey was administered to test takers to gather their perceptions of the EPT's fairness. On the qualitative side, a multi-stakeholder approach was adopted through semi-structured interviews with representatives of test takers, teachers, test administrators, and test users. The integration of quantitative and qualitative data was expected to provide a comprehensive and objective evaluation of the EPT's fairness. The findings of this study will be presented in the next chapter.

Chapter 4 Results

This chapter presents the findings of this study, examining the fairness of the EPT from both psychometric and stakeholders' perspectives. Section 4.1 reports on the quantitative analyses conducted to assess: (1) the comparability of test scores between humanities and science test-taker groups and (2) the comparability of input materials across test forms of the listening and reading subtests. Section 4.2 summarizes the findings from a test-taker questionnaire survey and the interviews with a sample of stakeholders. The section outlines stakeholders' perceptions of the EPT's fairness and the factors influencing their perceptions.

4.1 Test fairness from the psychometric perspective

Section 4.1 presents the results of psychometric analyses aimed at evaluating the fairness of the EPT. The findings are presented in Section 4.1.1 and Section 4.1.2. First, the comparability of test scores across test-taker groups with different academic backgrounds is reported in Section 4.1.1. The section includes results from DIF, DBF, and DTF analyses on the listening and reading subtests of the EPT. Second, findings on input material comparability across different test forms are presented in Section 4.1.2. The section presents the analysis results of the characteristics of the input materials for listening and reading tasks across different test forms. These findings provide evidence of the EPT's psychometric properties.

4.1.1 Test score comparability across test-taker groups with different academic backgrounds

4.1.1.1 Test score comparability as indicated by differential item/bundle functioning

As shown in Table 4.1, in the listening subtest, the science group outperformed the humanities group in three out of the four test forms. In the reading subtest, the science group outperformed the humanities group in two test forms, while the humanities

group scored higher than the science group in the remaining two test forms. Analysis of the test performance data revealed non-normal distributions across most test forms, as indicated by the Shapiro-Wilk test results. The Brown-Forsythe test results showed that the variances were homogeneous between the humanities and science groups for all test forms except Form 1 of the listening subtest ($F(1, 661) = 4.98$, $p = .03$). Due to these violations of parametric test assumptions, nonparametric Mann-Whitney *U* tests were conducted to examine potential differences in test performance between the humanities and science groups. Mann-Whitney *U* tests revealed comparable performance between the humanities and science groups across all reading test forms and most listening test forms. The only exception was observed in Form 2 of the listening subtest, where the science group performed significantly better than the humanities group ($U = 39{,}580.00$, $p = .04$), with a small effect size ($r = .052$). It is important to note that the observed differences in mean scores do not affect the identification of DIF or DBF. The SIBTEST procedure could mitigate the influence of overall group differences by matching test takers from both groups based on their true latent ability. This matching approach ensures that any detected DIF is due to genuine differences in how items function for different test-taker groups.

Table 4.1 Descriptive statistics and Shapiro-Wilk Test of normality for humanities and science groups' performance on listening and reading subtests.

Subtest	Form	Humanities									Science								
		n	Min	Max	Mean	*SD*	*Skewness*	*Kurtosis*	*W*	*p*	*n*	Min	Max	Mean	*SD*	*Skewness*	*Kurtosis*	*W*	*p*
Listening	1	119	7	25	16.41	4.01	.22	-.52	.97	.023*	544	8	27	16.60	3.43	-.07	-.39	.99	<.001***
	2	158	7	24	16.98	3.30	-.17	-.43	.98	.026*	561	6	30	17.65	3.69	-.10	-.04	.99	.003**
	3	103	9	28	18.32	3.76	-.09	.36	.97	.036*	613	5	28	17.66	3.58	-.25	.14	.99	<.001***
	4	125	7	26	18.37	4.03	-.21	-.28	.98	.07	605	6	30	18.40	4.02	-.34	-.19	.98	<.001***
Reading	1	119	3	19	11.80	3.29	-.19	-.51	.98	.14	544	0	20	11.78	2.98	-.35	.33	.98	<.001***
	2	158	6	20	14.20	2.78	-.40	.12	.98	.007**	561	0	20	14.28	2.98	-.62	.46	.96	<.001***
	3	103	5	20	12.24	3.07	.14	-.33	.98	.13	613	3	20	11.75	3.42	-.11	-.44	.99	<.001***
	4	125	3	17	9.98	3.36	.04	-.81	.97	.008**	605	2	19	10.22	3.09	.03	-.19	.99	<.001***

Notes. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. The university offers a wide range of programs across various fields of study. However, the number of programs and undergraduate students in science far exceed those in humanities. The discipline distribution of the sample of test takers was representative of the target test-taker population at the university.

I. The listening subtest

This section presents the results of the DIF and DBF analyses for the listening subtest. It begins with a detailed account of the SIBTEST analysis for Form 1, which includes a step-by-step explanation of the procedures for anchor purification and DIF/DBF detection. Following this, a concise overview of the DIF/DBF detection results is provided for the remaining three forms of the listening subtest (Forms 2, 3, and 4).

The anchor purification process for Form 1 of the listening subtest involved two rounds of analysis to identify DIF-free items. In the first round, 30 items were examined for both uniform and nonuniform DIF. Only one item (i.e., Item 15) showed a statistically significant DIF ($p < .05$). The remaining 29 items, which showed no significant DIF, were used to form a temporary matching subtest for the subsequent round of anchor purification. The second round of anchor purification was conducted to examine whether the items in the temporary matching subtest exhibited either uniform or nonuniform DIF. The results indicate that none of the items exhibited statistically significant DIF in this round. This two-step purification process resulted in a final matching subtest consisting of 29 DIF-free items, excluding Item 15, for the subsequent DIF and DBF detection.

DIF detection for Form 1 of the listening subtest was intended to focus solely on the 10 stand-alone items in the Short Conversation section. However, since these items were found to be free of significant DIF during the second round of anchor purification, further DIF analyses were not conducted. This finding indicates that the Short Conversation items in Form 1 of the listening subtest function similarly across the focal group (the humanities group) and the reference group (the science group), providing no evidence of item-level bias.

DBF analyses were then conducted on the four testlets in Form 1 of the listening subtest: Long Conversation (LC), Listening Passage 1 (LP1), Listening Passage 2 (LP2), and Listening Passage 3 (LP3), with each testlet containing five items. The matching subtest used for DBF analyses was obtained through the previously described anchor purification procedures. It is important to note that, unlike item-level DIF, there are no established effect size guidelines for determining the

magnitude of DBF. Consequently, in this study, testlets were simply categorized as either exhibiting DBF or being DBF-free.

As shown in Table 4.2, the results indicate that only one testlet, LP2, exhibited statistically significant DBF ($p < .05$), demonstrating both uniform and nonuniform DBF. For uniform DBF, a negative $\hat{\beta}_{\text{uni}}$ value ($\hat{\beta}_{\text{uni}} = -.242$) suggests that LP2 favored the humanities group along the ability continuum. To be specific, this result indicates that the humanities group was expected to score approximately .242 points higher than the science group on LP2, assuming equal ability levels. However, the presence of significant nonuniform DBF suggests that this advantage is not consistent across all levels of underlying abilities. In other words, the performance difference between the humanities and science groups on LP2 varied across the ability spectrum, potentially favoring the humanities group at some ability levels and the science group at others. For instance, LP2 might favor humanities group at lower ability levels and science group at higher ability levels. It is important to note that the SIBTEST procedure does not specify the exact ability levels at which group performance differences occur for nonuniform DBF. As a result, the precise points along the ability continuum where the advantage shifts between the two groups cannot be identified. The existence of nonuniform DBF revealed a nuanced interaction between test-takers' ability levels and their group membership. The detection of nonuniform DBF thus serves as a valuable complement to uniform DBF analysis, uncovering complexities that might otherwise go unnoticed.

Table 4.2 DBF detection in the listening subtest (Form 1).

Suspect bundle	No. of items	Type	$\hat{\beta}_{\text{uni}}$	p	Type	$\hat{\beta}_{\text{uni}}$	p
LC	5	Uniform	-.036	.733	Nonuniform	.039	.934
LP1	5	Uniform	.067	.626	Nonuniform	-.162	.367
LP2	5	Uniform	-.242	.029*	Nonuniform	-.242	.029*
LP3	5	Uniform	.085	.480	Nonuniform	-.085	.480

Notes. $^{*}p < .05$. LC = Long Conversation. LP = Listening Passage.

The same anchor purification and DIF/DBF detection procedures were conducted on the remaining three forms of the listening subtest. Table 4.3 presents DIF detection results on the stand-alone items in the Short Conversation section across test forms. Among the 40 items associated with the Short Conversation tasks in the four forms of the listening subtest, only two items exhibited DIF, accounting for 5% of the total short-conversation items. In Form 2, the sixth item in the Short Conversation section demonstrated a moderate level of uniform DIF ($\hat{\beta}_{\text{uni}}$ = -.103, p < .05), according to the guidelines proposed by Roussos and Stout (1996b). This negative $\hat{\beta}_{\text{uni}}$ value indicates that a test taker from the humanities group is expected to score approximately .103 points higher than a test taker of equal ability from the science group. Additionally, the tenth item in Form 3 demonstrated a large uniform DIF and a large nonuniform DIF. The presence of both uniform and nonuniform DIF in this item highlights the complex interaction between test-takers' item performance and their group membership.

Table 4.3 Overview of items showing DIF in the listening subtest across test forms.

Form	Type	Item	$\hat{\beta}_{\text{uni}}$	p	DIF level
Form 2	Uniform	6	-.103	.026*	Moderate
Form 3	Uniform	10	-.193	<.001*	Large
	Nonuniform	10	.193	<.001*	Large

Note. *p < .05.

The results of DBF detection for all testlets in the listening subtest across test forms are presented in Table 4.4. Out of the ten testlets examined, four (40%) exhibited statistically significant DBF, all of which demonstrated significant uniform DBF. Specifically, the LP2 in Form 1 favored the humanities group, as indicated by a negative $\hat{\beta}_{\text{uni}}$ value of -.242. In contrast, the long conversations in both Form 3 and Form 4, and the third listening passage in Form 4 favored the science group, as denoted by positive $\hat{\beta}_{\text{uni}}$ values. The uniform DBF results indicate that one testlet (10%) favored the humanities group, while three testlets (30%) advantaged the science group. The LC in Form 3 showed the largest amount of uniform DBF ($\hat{\beta}_{\text{uni}}$ = .375), suggesting that a randomly selected test taker with a humanities background

could score .375 points higher than a science student of equal ability on this testlet. Nonuniform DBF was also significant in all the four testlets, indicating that the testlet-level performance differences between the two groups varied across ability levels. Specifically, LP2 in Form 1 ($\hat{\beta}_{uni}$ = -.242, p = .029), LCs in Form 3 ($\hat{\beta}_{uni}$ = -.375, p = .007) and Form 4 ($\hat{\beta}_{uni}$ = -.379, p = .004), and LP3 in Form 4 ($\hat{\beta}_{uni}$ =.234, p = .041) exhibited a significant nonuniform DBF. This implies that the advantage for the humanities or science group could be pronounced at certain ability levels but not the others. Again, the results indicate the existence of an interaction between testlet performance and group membership. The presence of both uniform and nonuniform DBF in these testlets highlights the importance of examining how testlet functioning varies across different ability levels.

Table 4.4 Overview of testlets showing DBF in the listening subtest across test forms.

Form	Type	Bundle	$\hat{\beta}_{uni}$	p
Form 1	Uniform	LP2	-.242	.029*
	Nonuniform	LP2	-.242	.029*
Form 3	Uniform	LC	.375	.007*
	Nonuniform	LC	-.375	.007*
Form 4	Uniform	LC	.351	.002*
	Nonuniform	LC	-.379	.004*
	Uniform	LP3	.234	.041*
	Nonuniform	LP3	.234	.041*

Notes. *p < .05. LC = Long Conversation. LP = Listening Passage.

II. The reading subtest

This section presents the DBF detection results for the reading subtest. The same procedures for anchor purification and DBF detection for the listening subtest were employed.

Table 4.5 displays the results of the DBF detection for the reading subtest across the four test forms. Out of the nine testlets used, two testlets (22.22%) demonstrated statistically significant DBF, both in Form 2. Specifically, the second reading passage (RP2) exhibited a significant uniform DBF ($\hat{\beta}_{uni} > 0$, $p < .05$), favoring the science group. RP2 also showed a significant nonuniform DBF. Similarly, the Banked Cloze

(BC) task displayed a significant uniform DBF ($\hat{\beta}_{\text{uni}} < 0$, $p < .05$), favoring the humanities group, along with a significant nonuniform DBF ($p < .05$). While the existence of uniform DBF suggests an advantage for either the humanities or the science group in these two testlets, the presence of significant nonuniform DBF indicates that this advantage might not be consistent across all ability levels.

Table 4.5 Overview of testlets showing DBF in the reading subtest across test forms.

Form	Type	Bundle	$\hat{\beta}_{\text{uni}}$	p
Form 2	Uniform	RP2	.229	.015*
	Nonuniform	RP2	-.229	.015*
	Uniform	BC	-.384	.047*
	Nonuniform	BC	-.384	.047*

Notes. $^{*}p < .05$. RP = Reading Passage. BC = Banked Cloze.

4.1.1.2 Test score comparability as indicated by differential test functioning

As shown in Table 4.6, the baseline CFA model (see Figure 3.3) demonstrated good fit for the entire sample as well as for the humanities and science samples across the four test forms. One exception was observed for the humanities group in Form 2, where the Comparative Fit Index (CFI = .787) fell below the accepted threshold for good fit. Despite this exception, other fit indices for this group in Form 2 (RMSEA = .072, SRMR = .069) were within acceptable ranges. Therefore, the baseline CFA model was retained for the subsequent measurement invariance analyses.

Table 4.6 Summary of fit statistics for the CFA model across test forms.

Test form	Sample	SBχ^2	*df*	*p*	CFI	RMSEA	SRMR
Form 1	All	24.05	19	.19	.989	.002	.028
	Humanities	19.37	19	.43	.997	.013	.053
	Science	22.03	19	.28	.991	.017	.030
Form 2	All	14.87	19	.73	1.000	.000	.021
	Humanities	34.67	19	.02	.787*	.072	.069
	Science	11.50	19	.91	1.000	.000	.020
Form 3	All	27.33	19	.10	.985	.025	.027
	Humanities	20.47	19	.37	.982	.027	.057
	Science	24.63	19	.17	.989	.022	.028
Form 4	All	19.20	19	.44	1.000	.004	.022
	Humanities	22.79	19	.25	.969	.004	.057
	Science	14.70	19	.74	1.000	.000	.020

Notes. * indicates poor model-data fit. SBχ^2 = Satorra-Bentler scaled Chi-square. CFI = Comparative fit index. RMSEA = Root mean square error of approximation. SRMR = Standardized root mean square residual.

Table 4.7 presents the results of measurement invariance testing for the listening and reading test forms across the humanities and science groups. Scalar invariance was established for Forms 1 and 3, indicating the absence of DTF in these two forms. In other words, the test scores were comparable across the two groups. For Form 1, the configural, metric, and scalar invariance models demonstrated acceptable model-data fit, with ΔCFI values below the recommended cutoff of .01 (Cheung & Rensvold, 2002). These results indicate that both factor loadings and item intercepts were invariant across groups. However, strict invariance was not supported, as the ΔCFI value between strict and scalar invariance models exceeded the threshold for invariance (Cheung & Rensvold, 2002). It should be noted that strict invariance, the most restrictive form of invariance, is not required for meaningful comparisons of test scores between the two groups. Once scalar invariance is established, any observed group differences in the latent variables can be interpreted as true differences rather than measurement artifacts. Therefore, strict measurement invariance was not pursued in this study. Similarly, Form 3 demonstrated measurement invariance across the two groups, with all invariance models showing satisfactory fit indices and ΔCFI values within the recommended threshold.

Table 4.7 Fit statistics of measurement invariance models for the listening and reading test forms across disciplines.

Test form	Invariance model	Equality constraints	SBχ^2	*df*	CFI	RMSEA	SRMR	$\Delta\chi^2$ (Δdf)	ΔCFI
Form 1	Configural invariance model	-	41.41	38	.992	.016	.030	-	-
	Metric invariance model	Factor loadings	47.67	44	.992	.016	.035	6.26(6)	.000
	Scalar invariance model	Factor loadings & item intercepts	55.89	50	.987	.018	.038	8.22(6)	.005
	Strict invariance model	Factor loadings, item intercepts & residual variances	72.12	58	.970	.026	.053	16.23*(8)	.017*
Form 2	Configural invariance model	-	45.00	38	.984	.022	.028	-	-
	Metric invariance model	Factor loadings	58.80	44	.966	.030	.033	13.80*(6)	.018*
Form 3	Configural invariance model	-	45.28	38	.988	.022	.029	-	-
	Metric invariance model	Factor loadings	47.81	44	.993	.015	.031	2.53(6)	.005
	Scalar invariance model	Factor loadings & item intercepts	55.69	50	.990	.018	.033	7.88(6)	.003
	Strict invariance model	Factor loadings, item intercepts & residual variances	62.17	58	.988	.022	.029	6.48(8)	.003
Form 4	Configural invariance model	-	37.48	38	1.000	.000	.024	-	-
	Metric invariance model	Factor loadings	45.49	44	.998	.010	.029	8.01(6)	.002
	Scalar invariance model	Factor loadings & item intercepts	61.34	50	.981	.025	.032	15.85*(6)	.017*

Notes. SC = Short Conversation. LC = Long Conversation. LP = Listening Passage. $\Delta\chi^2$ = difference in Satorra-Bentler scaled Chi-square values between the nested models. Δdf = difference in degrees of freedom between the nested models. $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$. ΔCFI = Difference in CFI values between the nested models. $^{*}|\Delta CFI| > .01$. Invariance supported if $|\Delta CFI| \leq .01$ (Cheung & Rensvold, 2002).

In contrast, Forms 2 and 4 did not establish measurement invariance at the scalar level, which is necessary for meaningfully comparing latent means across the humanities and science groups. For Form 2, the ΔCFI between the metric and configural invariance models exceeded the recommended threshold (ΔCFI = .018), indicating non-invariance of factor loadings across the two groups. Consequently, scalar and strict invariance models were not tested, as successful establishment of metric invariance is necessary before testing more restrictive models. For Form 4, the ΔCFI between the scalar and metric invariance models was also above the threshold (ΔCFI = .017), indicating a lack of scalar invariance. These results suggest that Forms 2 and 4 might not measure the same construct in the same way across the two groups. It can be concluded that DTF was present in Forms 2 and 4. In other words, for these two test forms, the test scores were not comparable across the humanities and science groups. However, whether the identified DTF introduces bias in the listening and reading subtests remains to be determined through content analysis.

4.1.2 Input material comparability across test forms

4.1.2.1 Comparability of listening input material

Overall, the measures showed relatively consistent patterns across all test forms of the listening subtest. In terms of lexical characteristics, the percentage of high-frequency words (most frequent 1,000 and 2,000 words) was consistently high, ranging from 88.80% to 91.10% across test forms. By contrast, academic and off-list words accounted for less than 12% of the input materials, indicating that the listening input materials were largely composed of commonly used vocabulary. This vocabulary profile aligned well with the purpose of the EPT, which is designed as a general-purpose proficiency test. The values for lexical diversity (measured by the MATTR) and lexical density (measured by the proportion of content words to the total number of words) were also comparable across test forms. Similarly, the word concreteness values, ranging from 362 to 379, exhibited a consistent trend across test

forms. Coh-Metrix assigns concreteness values on a scale from 100 to 700, where higher values indicate a higher frequency of concrete words and lower values a high frequency of more abstract words in a text. The observed mid-range values suggest that the listening input materials maintained a balanced mix of concrete and abstract words. Taken together, the above findings indicate that the listening subtest of the EPT maintained a comparable level of lexical complexity across different test forms.

The syntactic characteristics of the listening input materials demonstrated consistency across the four test forms, with only minor variations. Length-based syntactic complexity measures for each test form were relatively stable. For instance, the values for MLS ranged from 10.50 to 12.80, indicating relatively comparable sentence complexity across test forms. The MLC showed even greater levels of consistency, with the values ranging from 7.34 to 8.80. However, the ML-AS exhibited some variations, with the lowest mean in Form 4 at 9.76 (*SD* = 5.35) and the highest in Form 2 at 12.30 (*SD* = 2.92). Left embeddedness, which can mirror the cognitive demand on listeners by measuring the mean number of words before the main verb, also showed consistency across the four test forms. This suggests that the cognitive load imposed by the syntactic complexity of the listening input was similar across all forms. While there were some variations in phrase-level and clause-level complexity measures, they generally remained comparable across the test forms. Overall, these findings indicate that the listening input materials maintained a consistent level of syntactic complexity across different test forms, ensuring comparable levels of cognitive processing demands imposed by sentence structures.

The discourse characteristics of the listening input materials were generally consistent across the four test forms, with exceptions observed in one measure of local-level cohesion and one measure of text-level cohesion. At the local level, the incidence score of all connectives—including causal, logical, and temporal connectives—varied considerably, with Form 4 showing the lowest mean of 56.10 (*SD* = 44.10) and Form 2 the highest mean of 95.30 (*SD* = 45.70). This variation

suggests potential differences in the explicit cohesion of the input materials across test forms. The semantic similarity between sentences, measured by the Latent Semantic Analysis (LSA) cosine values, remained relatively consistent across test forms. At the global level, semantic similarity between paragraphs showed slight variations across test forms, with means ranging from .07 (Form 2) to .14 (Form 1). The text-level characteristics exhibited both variation and consistency. The incidence scores of causal content (including causal verbs and causal particles) showed variation, with means ranging from 44.40 (Form 1) to 58.70 (Form 4), indicating differences in the amount of causal relationships expressed in the listening input materials across the test forms. In contrast, temporal cohesion remained relatively stable across all test forms. Semantic similarity at the text level was notably consistent, with all forms showing a mean of .24, suggesting similar semantic relationships within the input materials across the four test forms.

The speed of delivery of the audio recordings was consistent across the four test forms. The measure of words per minute demonstrated a high level of comparability across the four test forms, with mean values of 158, 159, 160, and 161 words per minute for Forms 1–4, respectively. The measure of speech rate demonstrated negligible variations across test forms, ranging from 3.59 to 3.78 syllables per second. The minor differences in both measures indicate that the speed of delivery was maintained at a comparable level across the four test forms. Such comparability suggests that the test takers were exposed to comparable listening experiences in terms of the speed of delivery, regardless of which test form they completed.

To examine whether there were significant differences in the characteristics of the listening input materials across the four test forms, a multivariate Kruskal-Wallis test was conducted. This analysis included all 27 measures—representing lexical characteristics, syntactic characteristics, discourse characteristics, and speed of delivery—as dependent variables, with the test form serving as the independent variable. The result indicates no statistically significant differences among the test

forms in terms of the characteristics of the listening input (H = .07, *df* = 3, *p* = .98). Given the constraints of the sample size, a permutation test with 10,000 iterations was performed to further validate this finding. The permutation test validated the results of the multivariate Kruskal-Wallis test, confirming that there were no significant differences among the test forms (*p* = .96). The alignment of the results from both the multivariate Kruskal-Wallis test and the permutation test strongly supported the conclusion that there were no significant differences in the characteristics of the listening input materials across the four test forms.

Following the multivariate analyses, separate Kruskal-Wallis tests were conducted for each of the 27 measures to identify specific characteristics that might differ significantly across the four test forms (see Table 4.8). These univariate analyses were performed to provide a more nuanced understanding of how individual characteristics vary across test forms, and to identify potential differences that might have been masked in the multivariate analysis due to the limited sample size. The analyses revealed no significant differences among the test forms for all measures after adjusting the *p* values for multiple comparisons. Initially, the unadjusted *p*-values indicate potential differences in two measures: the ML-AS (H = 11.34, *p* = .01) and the incidence score of all connectives (H = 9.52, *p* = .02). However, after applying the Benjamini-Hochberg correction, these differences were no longer statistically significant, with adjusted *p*-values of .27 and .31, respectively.

In summary, the above findings suggest that the characteristics of the listening input materials were comparable across the four test forms. The absence of significant differences in lexical characteristics, syntactic characteristics, discourse characteristics, and speed of delivery indicate that the listening input materials maintained a comparable level of complexity, which was essential for ensuring the comparability of listening task difficulty (Mohamadi, 2013; Willingham & Cole, 1997).

Table 4.8 Kruskal-Wallis test results for characteristics of the listening input across test forms.

Measure	H statistic	*df*	*p*	Adjusted *p*	η^2
Lexical characteristics					
Lexical frequency (%)					
K1 words	.30	3	.96	.99	.01
K1 function words	2.78	3	.43	.83	.05
K1 content words	5.73	3	.13	.50	.10
K2 words	1.53	3	.68	.91	.03
K1 + K2 words	1.24	3	.74	.92	.02
Academic words	2.57	3	.46	.83	.05
Off-list words	2.60	3	.46	.83	.05
Lexical diversity					
MATTR	1.21	3	.75	.92	.02
Lexical density					
CW/W	2.79	3	.43	.83	.05
Concreteness					
Concreteness value	2.73	3	.43	.83	.05
Syntactic characteristics					
Length-based characteristics					
MLS	7.70	3	.05	.47	.14
MLC	5.67	3	.13	.50	.10
ML-AS	11.34	3	.01*	.27	.21
Phrase-level characteristics					
Left embeddedness	.74	3	.86	.97	.01
M/NP	7.00	3	.07	.48	.13
CN/C	3.77	3	.29	.83	.07
CN/AS	4.42	3	.22	.74	.08
Clause-level characteristics					
C/AS	2.82	3	.42	.83	.05
C/S	2.09	3	.55	.88	.04
Discourse characteristics					
Local-level characteristics					
All connectives	9.52	3	.02*	.31	.17
Semantic similarity	2.42	3	.49	.83	.04
Global-level characteristics					
Semantic similarity	6.53	3	.09	.48	.12
Text-level characteristics					
Causal content	.97	3	.81	.95	.02
Temporal cohesion	1.68	3	.64	.91	.03
Semantic similarity	.08	3	.99	.99	.00

(*Continued*)

Table 4.8 (Continued).

Measure	H statistic	*df*	*p*	Adjusted *p*	η^2
Speed of delivery					
Words per minute	.14	3	.99	.99	.00
Speech rate	1.65	3	.65	.91	.03

Notes. $^{*}p < .05$. MATTR = Moving average type-token ratio. CW/W = Proportion of content words to the total number of words. MLS = Mean length of sentence. MLC = Mean length of clause. ML-AS = Mean length of AS-unit. M/NP = Modifiers per noun phrase. CN/C = Complex nominals per clause. CN/AS = Complex nominals per AS-unit. C/AS = Clauses per AS-unit. C/S = Clauses per sentence.

4.1.2.2 Comparability of reading input material

Overall, input characteristics across the four test forms did not exhibit a consistent level of comparability. While some characteristics were consistent across test forms, others exhibited variation, potentially leading to differences in the complexity of the reading input materials.

Lexical characteristics across the four test forms exhibited a high degree of comparability. In terms of lexical frequency, the percentage of high-frequency words, specifically the most frequent 1,000 and 2,000 words, were consistently high across all test forms, ranging from approximately 79% to 84%. Although the proportions of academic words in the reading input materials showed some variations across test forms, they consistently remained below 10%. This distribution of academic words suggests that the reading input materials were designed to assess general English reading proficiency rather than academic reading skills. Additionally, measures of lexical diversity, lexical density, and word concreteness were comparable across all test forms. The results indicate that the reading subtest of the EPT maintained a comparable level of lexical complexity across the four test forms.

Syntactic characteristics across the four test forms exhibited both variation and comparability. Among the four test forms, Form 2 included reading passages with the lowest level of syntactic complexity, whereas Form 4 featured the most syntactically complex passages. Length-based characteristics showed notable differences, with Form 4 consistently displaying the highest values for MLS, MLC, and MLT. The

results suggest that, out of the four test forms, the reading passages in Form 4 demonstrated the highest level of syntactic complexity. Phrase-level characteristics demonstrated variations across test forms, as indicated by the measure of left embeddedness. The lowest value was found in Form 2 (M = 4.56, SD = .89), and the highest in Form 4 (M = 7.16, SD = .63). The results indicate potential differences in the cognitive load imposed by word-level syntactic complexity of the reading input materials. By contrast, statistics for M/NP, CN/C, CN/T, CP/C, and CP/T remained relatively consistent across test forms, with only minor variations. This indicates similar levels of complexity regarding noun phrase and coordinate phrase structures. Clause-level characteristics demonstrated both uniformity and divergence across the four test forms. Measures such as C/T and C/S were comparable across all test forms. However, measures of dependent clauses, specifically DC/C and DC/T, exhibited a different pattern. Forms 1, 2, and 3 demonstrated comparable values in terms of DC/C and DC/T, while Form 4 showed notably higher values. The results suggest that Form 4, compared with the other test forms, featured an increased level of subordination. In summary, while some syntactic characteristics were comparable across the test forms, others exhibited variations. These variations could potentially affect the test-takers' perceived difficulty of the reading passages in different test forms.

The discourse characteristics across the four test forms exhibited variation at the local, global, and textual levels, with a few exceptions in local-level measures. For example, the incidence scores of causal connectives and semantic similarity (as measured by LSA cosine values) remained stable across the test forms. This stability indicates comparability in causal relations and semantic coherence between sentences in the reading input materials across test forms. However, other local-level measures displayed notable differences. Form 3, for instance, stood out with the highest incidence scores for all connectives (M = 100.00) and logical connectives (M = 40.50), suggesting more explicit local-level textual cohesion compared to that of other test forms. Additionally, local-level noun, argument, and stem overlap measures,

indicative of the repetition of key ideas between adjacent sentences, varied across test forms. Form 3 generally showed higher values in these measures, reflecting greater referential cohesion between neighboring sentences. Global-level cohesion measures also varied across the test forms. Similar to local-level cohesion, Form 3 exhibited the highest means for the overlap among nouns (M = .34, SD = .22), arguments (M = .45, SD = .27), and stems (M = .42, SD = .21), while Form 4 demonstrated the highest mean for semantic similarity (M = .21), indicating a higher degree of global semantic cohesion. Text-level characteristics fluctuated in terms of the incidence scores of causal content and cohesion, intentional content and cohesion, and temporal cohesion as well. The text-level semantic similarity also varied across the test forms, with the LSA cosine values ranging from .23 (Form 2) to .41 (Form 4). The text-level semantic similarity also varied across test forms, with the LSA cosine values ranging from .23 (Form 2) to .41 (Form 4). These variations in discourse characteristics across test forms may affect the readability of the reading passages, potentially leading to varying levels of comprehension difficulty for test takers who completed different test forms.

The readability measure also showed variation across the four test forms. The Coh-Metrix L2 Readability Index showed that the reading input in Form 2 was the most accessible (M = 14.10) and that in Form 4 the most challenging (M = 8.53) for test takers. The variation in the readability measure suggests that test takers might encounter different levels of comprehension difficulty across the four test forms. These variations could potentially influence their reading comprehension and even their overall test performance.

4.2 Test fairness from the stakeholders' perspective

4.2.1 Questionnaire results

4.2.1.1 EFA results

An EFA was conducted on a random valid sample (n = 567) to identify the underlying

factor structure of the questionnaire designed to measure test-takers' perceived fairness of the EPT. Four factors with eigenvalues greater than one were extracted via Principal Component Analysis with Direct Oblimin rotation. This four-factor solution could explain a cumulative variance of 55.50% in test-takers' responses (see Table 4.9).

Table 4.9 Eigenvalues for the four-factor solution of the test fairness questionnaire.

Factor	Eigenvalue	% of Variance	Cumulative %
Factor 1	6.80	20.60	20.60
Factor 2	4.87	14.76	35.40
Factor 3	3.76	11.40	46.80
Factor 4	2.88	8.72	55.50

Note. $n = 567$.

Factor labels were assigned through a detailed content analysis of the items associated with each factor. Factor 1 included 11 items relating to the uniformity of administration practices during test-taking. Thus, Factor 1 was labeled "Consistency in Administration". The 10 items comprising Factor 2 tapped into the comparability of opportunities for test takers to demonstrate their proficiency in the EPT. Accordingly, Factor 2 was labeled "Comparable Opportunity". Factor 3, consisting of eight items about the transparency of test-related information, was labeled "Test Information Accessibility". Lastly, Factor 4, labeled "Accountability Mechanisms", included four items regarding the test-takers' rights to request a score review and voice concerns about testing services after taking the EPT.

The rotated pattern matrix showing factor loadings and internal consistency of each factor is presented in Table 4.10. The Cronbach's alpha values for each factor, ranging from .746 to .927, indicating acceptable to excellent internal consistency within each factor (DeVellis & Thorpe, 2022). In other words, the items associated with each factor can reliably measure the test-takers' perceptions across the four dimensions of test fairness represented by these four factors.

Table 4.10 EFA results and reliabilities for the four factors of the test fairness questionnaire.

	F1	F2	F3	F4	Cronbach's α
Factor 1: Consistency in Administration					.927
Item 27	.904				
Item 42	.897				
Item 40	.840				
Item 21	.826				
Item 24	.749				
Item 26	.725				
Item 23	.674				
Item 20	.671				
Item 25	.632				
Item 46	.520				
Item 41	.468				
Factor 2: Comparable Opportunity					.873
Item 33		.782			
Item 28		.753			
Item 34		.746			
Item 35		.695			
Item 29		.692			
Item 36		.647			
Item 30		.632			
Item 32		.630			
Item 37		.629			
Item 31		.609			
Factor 3: Test Information Accessibility					.832
Item 8			.800		
Item 7			.778		
Item 9			.727		
Item 10			.662		
Item 4			.577		
Item 13			.532		
Item 5			.461		
Item 1			.459		
Factor 4: Accountability Mechanisms					.746
Item 48				.739	
Item 45				.723	
Item 44				.694	
Item 47				.599	

Notes. $n = 567$. Extraction Method: Principal Component Analysis. Rotation Method: Oblimin without Kaiser Normalization (rotation converged in seven iterations). Coefficients below .40 are not presented.

The EFA revealed a substantial congruence between the initially proposed dimensions of test fairness and the empirically derived factors (see Table 4.11). Each of the four tentative dimensions—comparability, accessibility, consistency, and accountability—yielded a standalone factor in the questionnaire (i.e., Factor 2, Factor 3, Factor 1, and Factor 4, respectively). Thus, the factor structure of the questionnaire demonstrated an alignment with the conceptualization of test fairness in this study.

Table 4.11 Comparison of tentative dimensions of test fairness with data-driven factor structure.

Tentative dimension		Extracted factor
Comparability	→	Factor 2: Comparable Opportunity
Accessibility	→	Factor 3: Test Information Accessibility
Consistency	→	Factor 1: Consistency in Administration
Accountability	→	Factor 4: Accountability Mechanisms

The correlation matrix revealed significant positive relationships among all the four factors, with correlation coefficients ranging from .26 to .61 (see Table 4.12). The use of Spearman's rho, a non-parametric measure, was chosen due to the non-normality of the factor score distribution. The strongest correlation was observed between Factor 1 ("Consistency in Administration") and Factor 4 ("Accountability Mechanisms"), with a coefficient of .61. Factor 1 ("Consistency in Administration") showed moderate correlations with all other factors. Factor 2 ("Comparable Opportunity") emerged as the most distinct factor, as evidenced by its weak Spearman's rho correlation coefficients (.26–.39) with the other factors. Overall, the observed relationships among the four factors, though varying in strength, underscore the multidimensional nature of test fairness.

Table 4.12 Correlation matrix for the four factors of the test fairness questionnaire.

Factor	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1: Consistency in Administration	1.00			
Factor 2: Comparable Opportunity	.39***	1.00		
Factor 3: Test Information Accessibility	.57***	.30***	1.00	
Factor 4: Accountability Mechanisms	.61***	.26***	.50***	1.00

Notes. n = 1,134. ***p < .001. Spearman's rho coefficients are reported due to non-normal distributions of the factor scores.

4.2.1.2 CFA results

CFA with WLSMV estimation was conducted on the other random sample (n = 576) to validate the four-factor structure of the test fairness questionnaire extracted by the EFA. To determine the most appropriate factor structure for the questionnaire, three competing CFA models were specified and tested respectively (see Figure 4.1, Figure 4.2, and Figure 4.3).

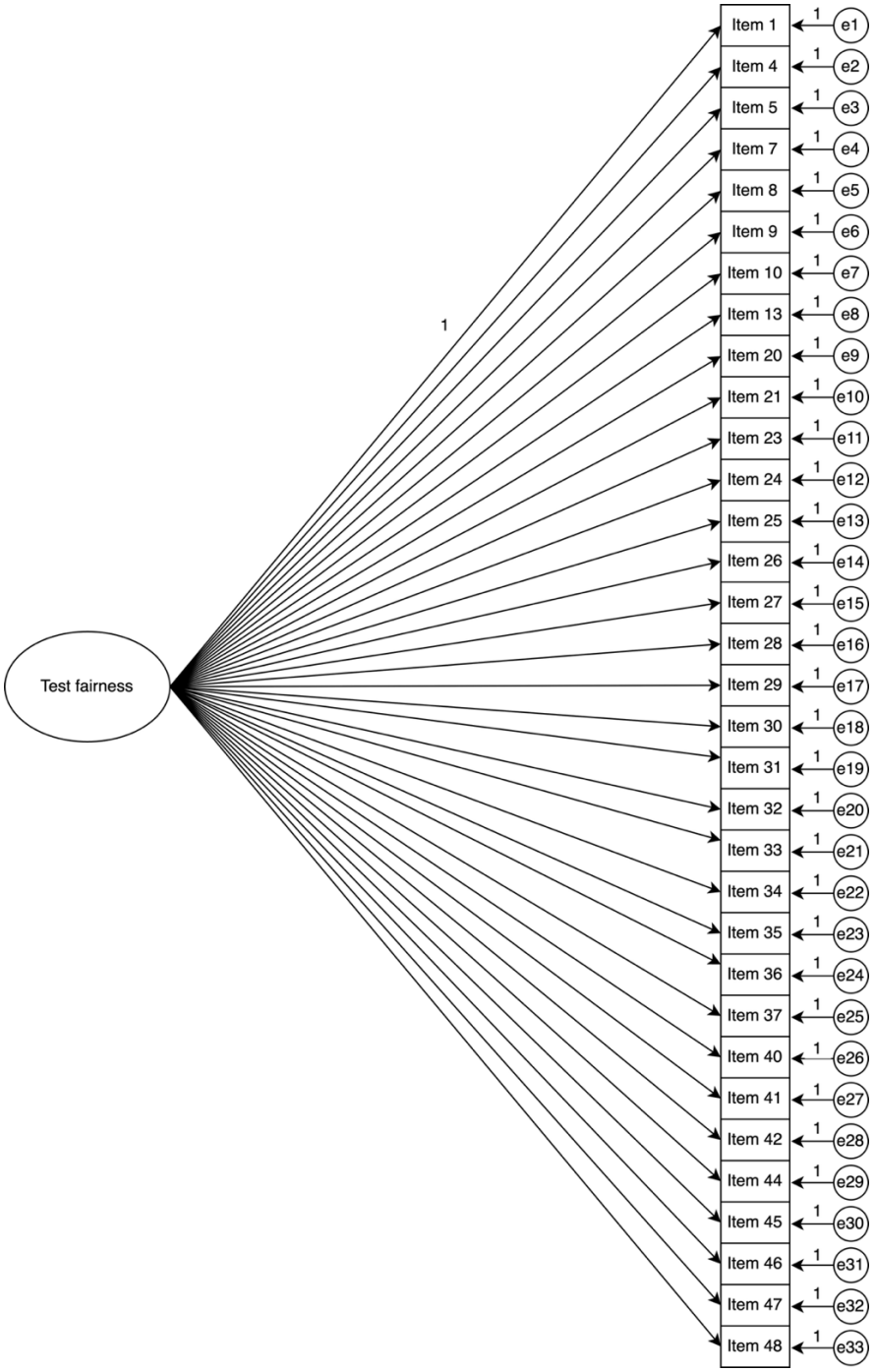

Figure 4.1 Single-factor model for the test fairness questionnaire.

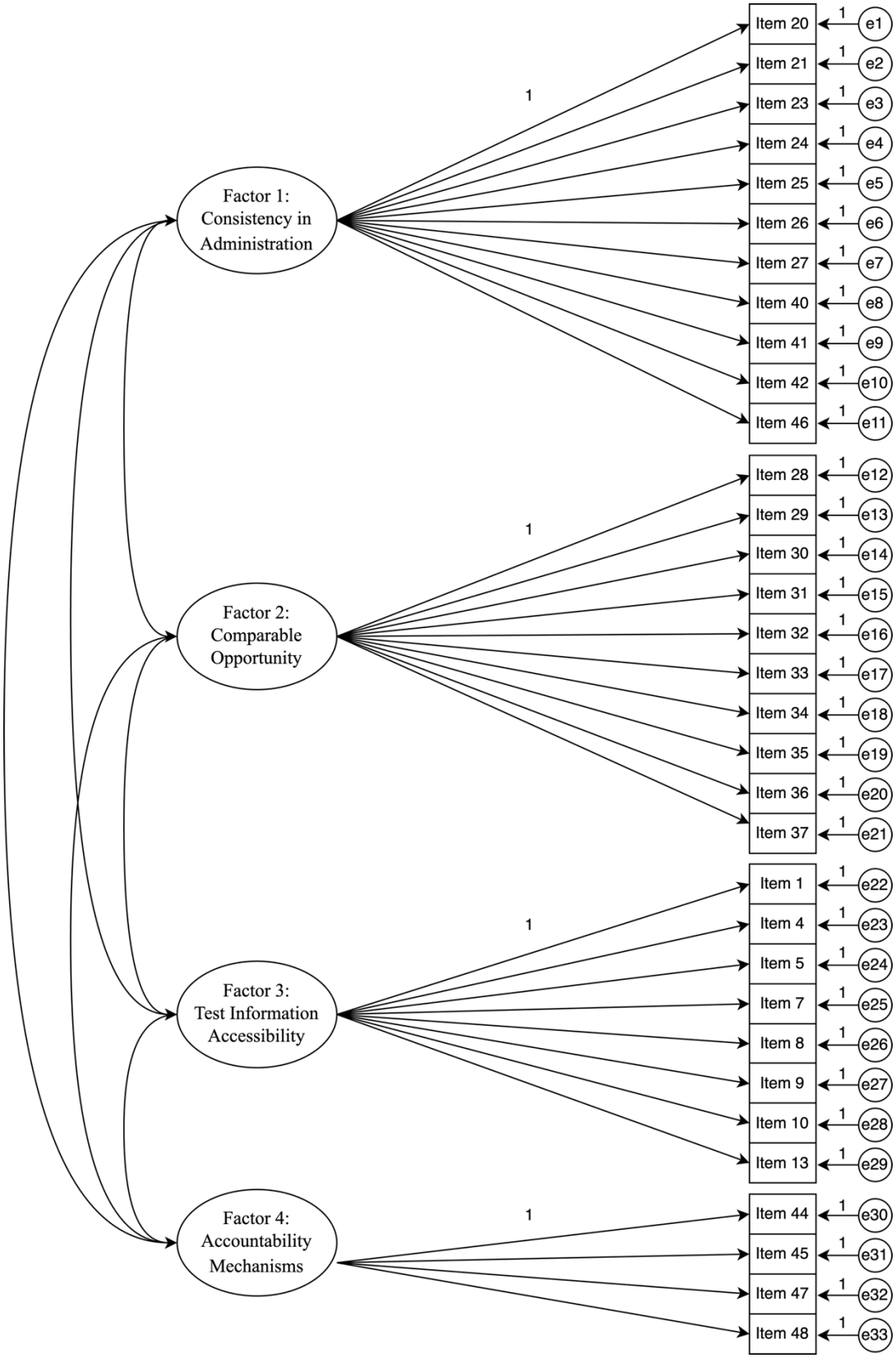

Figure 4.2 Correlated-trait model for the test fairness questionnaire.

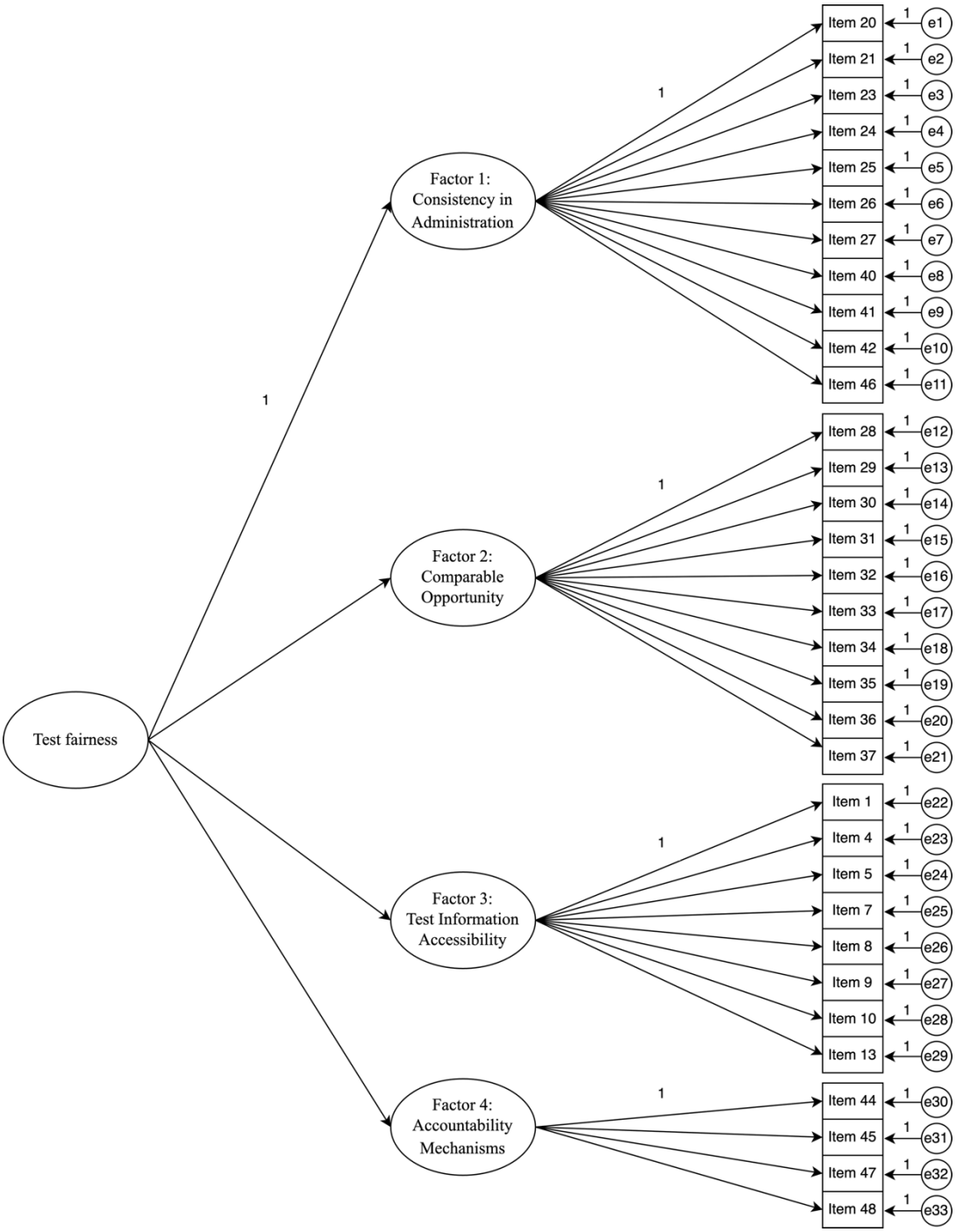

Figure 4.3 Second-order factor model for the test fairness questionnaire.

As shown in Table 4.13, the one-factor CFA model demonstrated poor model-data fit, as indicated by the statistics of CFI, RMSEA, and SRMR. In contrast, the fit indices for the correlated-trait model and the second-order model were satisfactory. The correlated-trait model exhibited slightly better fit indices (χ^2 = 1,448, *df* = 489, CFI = .973, RMSEA = .059, SRMR = .063) compared to the second-order model (χ^2

= 1,451, *df* = 491, CFI = .973, RMSEA = .059, SRMR = .065). Although the difference in fit indices between these two models was minimal, the correlated-trait model was retained as the final and most parsimonious representation of the factor structure underlying the test fairness questionnaire. In summary, the results from the CFA supported the factor structure of the questionnaire identified by the EFA.

Table 4.13 Fit statistics for three competing CFA models of the test fairness questionnaire.

Model	χ^2	*df*	*p*	CFI	RMSEA	SRMR
One-factor model	5,390	495	< .001	.864	.132	.143
Correlated-trait model	1,448	489	< .001	.973	.059	.063
Second-order model	1,451	491	< .001	.973	.059	.065

4.2.1.3 Test-takers' perceptions of test fairness

Overall, the test takers provided moderate to high ratings across the questionnaire items, with item means ranging from 3.61 to 5.90, indicating a generally positive perception of the fairness of the EPT.

Factor-level descriptive statistics can be found in Table 4.14. Factor scores were derived by summing the ratings on the constituent items for each factor and dividing that sum by the total number of items in the respective factor. The means of factor scores, ranging from 4.32 to 5.79, indicate that the test takers held moderately positive to positive perceptions of the fairness of the EPT across the four factors. The standard deviations for the factor scores ranged from .38 to .96, suggesting that the test-takers' responses across the four factors exhibited low variability.

Table 4.14 Descriptive statistics for the four factors of the test fairness questionnaire.

Factor	No. of items	Item No.	*M*	*SD*	*Skew.*	*Kurt.*	Cronbach's α
Factor 1	11	20, 21, 23, 24, 25, 26, 27, 40, 41, 42, 46	5.79	.38	-2.21	4.81	.932
Factor 2	10	28, 29, 30, 31, 32, 33, 34, 35, 36, 37	4.32	.96	-.36	-.15	.875
Factor 3	8	1, 4, 5, 7, 8, 9, 10, 13	5.08	.84	-.65	-.51	.829
Factor 4	4	44, 45, 47, 48	5.27	.72	-.70	-.45	.733

Notes. n = 1,134. Questionnaire items can be found in Appendix 3.

As shown in Table 4.14, the test takers perceived "Consistency in Test Administration" (Factor 1) as the fairest aspect of the EPT. This was followed by "Accountability Mechanisms" (Factor 4), "Test Information Accessibility" (Factor 3), and "Comparable Opportunities" to demonstrate English proficiency during the EPT (Factor 2). The highest mean value for Factor 1—"Consistency in Administration" (M = 5.79, SD = .38)—suggests that the test takers agreed that the EPT was administered consistently. This is understandable due to the standardized administration procedures of the test. For instance, uniform security and ID checks, along with rigorous proctoring measures, all helped ensure similar testing conditions for all test takers. The second-highest mean value (M = 5.27, SD = .72) was observed in Factor 4 ("Accountability Mechanisms"). This suggests that the test takers perceived the EPT as fair, likely because they can request a score review or raise any concerns about unfair testing practices (e.g., test taker or proctor misconduct). The mean value for Factor 3, "Test Information Accessibility" (M = 5.08, SD = .84), demonstrated the test-takers' perceived transparency of test-related information (e.g., test procedures, delivery mode, and rating criteria). The lowest mean value was observed for Factor 2, "Comparable Opportunity" (M = 4.32, SD = .96), suggesting that the test takers only marginally concurred with the statement that they had comparable opportunities to demonstrate their English proficiency during the EPT. A closer look at the items associated with Factor 2 revealed that Item 32 received the lowest ratings (M = 3.61, SD = 1.47) among all items, indicating that the test-takers' performance was likely influenced by their insufficient background knowledge of the topics covered in the test.

4.2.2 Interview results

This section presents the perceptions of various stakeholders regarding the fairness of the EPT. A thematic analysis of interview data revealed four overarching themes which aligned well with the four dimensions of test fairness as identified in the

tentative conceptual model (see Section 2.3). Interviews were conducted with 31 participants, including 20 test takers, six teachers, two test administrators, and three test users[1]. Overall, the participants all perceived the EPT as a fair test. The following sub-sections present the interviewees' perceptions, organized around the four themes: comparability, accessibility, consistency, and accountability.

4.2.2.1 Stakeholders' perceptions of test fairness

I. Theme 1: Comparability

As revealed by the interview data, key considerations underpinning "comparability" include two aspects: (1) comparability of the construct measured by the EPT across test-taker groups and (2) comparability of test results across test forms.

Sub-theme 1: Comparability of the construct measured by the EPT across test-taker groups

This sub-theme refers to the extent to which the construct measured by the EPT is comparable across different groups of test takers. According to the interview data, test takers can be categorized into different groups based on their: (1) academic backgrounds and (2) strengths or weaknesses in certain English skills.

Overall, all the teachers and test users who were involved in the development of the EPT asserted that the EPT was fair for test takers from different academic backgrounds. In designing the EPT, test developers intentionally avoid using input materials that feature "overly specialized, abstruse, or sensitive topics" (TD4; TU2, 3). The topics covered in the test content were reported to be free from bias and did not provide any advantage to test takers from different academic backgrounds.

From the test-takers' point of view, a majority of them, regardless of their majors, reported familiarity with the topics covered in the listening and reading input materials. Drawing from their test-taking experiences, several test takers noted that

[1] Each participant was assigned a unique code for ease of reference in reporting the results. Test takers are coded as TT1–TT20; teachers (also test developers) as TD1–TD6; test administrators as TA1–TA2; and test users as TU1–TU3.

the input materials did not include "terminologies or specialized knowledge" that might be more familiar to test takers from certain disciplines (TT6, 8, 9, 11).

Moreover, two teachers and one test user thought that the EPT did not favor test takers who exhibit weakness in a particular skill. As a "full-skill test" (TD4, 6; TU1), the EPT is designed to evaluate test-taker's general proficiency across four skills: listening, speaking, reading, and writing. Accordingly, the test is divided into four subtests. To pass the EPT, test takers must meet the established standards in each of the subtests. In other words, test takers who were particularly weak in any one skill were likely to fail the EPT.

Sub-theme 2: Comparability of test results across test forms

Four teachers and all three test users maintained that test takers with the same proficiency levels would achieve comparable scores, regardless of which test session they attended. They argued that the test results would not be influenced by "the difficulty levels of different test forms" (TD1; TU2, 3). As TU2 explained, an anchor item design was employed in the development of the listening and reading subtests to link test forms and facilitate post-test equating procedures:

> *"From a technical standpoint, anchor items are embedded in multiple test forms used for each administration. Moreover, since the initial administration of the test, post-test equating method has remained the same. This ensures that test-takers' scores remain comparable across test forms, test sessions, and academic years."* (TU2)

One teacher and all the test users acknowledged the importance of balancing the overall difficulty of different test forms. The assembled test papers were reported to "undergo multiple rounds of review prior to operational administration" (TD6). However, the variations in the overall difficulty levels of different test forms were inevitable. They emphasized that such variations would not compromise the comparability of test scores across different test-taker groups, as an anchor-item design was employed to address this issue. These findings suggest that ensuring test score comparability is a top priority in upholding test fairness.

II. Theme 2: Accessibility

In this study, "accessibility" refers to the extent to which the EPT is designed and administered to ensure that all test takers have:

- access to test-related information (Sub-theme 1);
- opportunities for test taking (Sub-theme 2) and learning (Sub-theme 3);
- access to the test location (Sub-theme 4);
- opportunities to familiarize themselves with the test delivery system (Sub-theme 5).

Sub-theme 1: Test information accessibility

Overall, all the test takers perceived the information related to the EPT as "transparent". They acknowledged the variety of channels for obtaining test-related information.

Test takers reported that they were able to obtain test-related information through official channels. These channels included the university's official websites and the registration notice from the university's Academic Affairs Office. These two channels provided test takers with "authoritative and accurate test information" (TT2, 4, 7, 19) prior to the test day. TT2 further added that relevant official documents could be accessed through the links provided in the test registration notice:

> *"The staff from the university's Academic Affairs Office issues the registration notice for the English Proficiency Test. There are many attachments to the notice, including the test syllabus, administration details, and a sample test paper. By reviewing these documents, I have got a clear understanding of the procedures and the format of the test."* (TT2)

The above-mentioned official documents provided test takers with detailed information about the EPT, including "vocabulary requirement" (TT1, 3, 11), "task types" (TT11, 12), "weighting for each task" (TT4, 9, 11, 20), "delivery mode" (TT19), "test procedures" (TT13), "speech rate of the listening materials" (TT19),

and “scoring criteria” (TT11, 19). In addition, test takers could obtain information about relevant policies by referring to the *Handbook for Undergraduate Students*.

In addition to official documents, test takers relied on other sources to obtain test-related information, including the university’s online forum, social media platforms, and word-of-mouth communication (TT1, 3, 6, 7, 10, 11, 14, 15, 19, 20). For instance, TT20 shared:

> *“I obtained firsthand test-related information from my classmates or roommates who had already taken the test. I also browsed the university’s online forum to find suggestions for test preparation shared by those who had passed the EPT.”* (TT20)

Three test takers reported to have access to test-related information through notices posted on WeChat Official Account and in QQ special interest groups managed by the Department of Student Services, Student Union. Overall, the various information access channels ensured the transparency of test-related information and offered a supportive network to assist prospective test takers in preparing for the test.

Sub-theme 2: Test-taking opportunities

All test takers believed that they were provided with sufficient opportunities to take the EPT. They also reported the sufficiency of test slots. Moreover, they could select test sessions “according to their personal preferences” (TT14, 15, 16, 18).

One test administrator and two teachers believed that the registration policy of the EPT show full respect to the test-takers’ autonomy in deciding when to take the test. Test takers can take the test during any academic year they prefer starting from their sophomore year. This flexibility allows test takers to perform at their best on the test day:

> *“Test takers have the flexibility to register for the test according to their individual preferences. They can choose to take the test when they feel fully prepared and confident in their ability to pass the test.”* (TD3)

The EPT was reported to offer sufficient slots to accommodate the needs of its test takers. According to TA2:

> *“As long as test takers stay tuned with the registration notices for the test, they*

should be able to secure their preferred slot". (TA2)

Nevertheless, due to equipment and personnel constraints, it is not always possible to guarantee every test taker their preferred slots. According to TA2, slots for the less popular sessions were "always available during each administration".

Sub-theme 3: Learning opportunities

The third sub-theme—"Learning opportunities"—focuses on test-takers' access to learning and preparation resources prior to test-taking.

Teachers believed that the learning and preparation resources were adequate for undergraduates. For instance, the university provides a series of College English courses (Levels II-VI) designed to enhance undergraduates' general English proficiency. Although these courses are not intended to prepare them for passing the EPT, the teaching content "aligns well with the knowledge and skills assessed by the test" (TD2, 5, 6).

Consistent with the perceptions of teachers, a majority of test takers also thought that the College English courses helped them enhance their general English proficiency and meet the passing standard for the EPT. In addition to College English courses, test takers could enroll in various English courses like "English Public Speaking", "English Listening and Speaking", and so forth to strengthen any identified weaknesses in their English skills.

Despite the overall positive feedback on learning and preparation resources, one test user voiced a concern with the learning materials available for those who struggled with English learning:

> *"One of my concerns is the lack of support for test takers who find difficulties in English learning and fail the English Proficiency Test. Given that passing this test is a graduation requirement, I think that the learning resources should be tailored to meet the test-takers' individual needs and effectively enhance their English proficiency."* (TU1)

In fact, for undergraduates who struggled with English learning and made multiple unsuccessful attempts at the EPT, the university offered academic support for all the students during their undergraduate years:

> *"Undergraduates are encouraged to attend tutorial sessions, during which teachers of College English courses are available to provide academic support. In addition, the university's English Writing Center provides one-on-one tutorials which can help students improve their writing skills."* (TD2)

TD4 observed that, despite the availability of various learning resources, repeat test takers had not made full use of them. She hoped that the test takers could take full advantage of on-campus learning resources and improve their general English proficiency before retaking the EPT.

Sub-theme 4: Test location accessibility

For a vast majority of test takers who stayed on the main campus where the test was administered, the distance to the test location did not compromise the fairness of the EPT. For those who did not stay on the main campus, the university arranged shuttle buses, so they were able to "arrive at the test location punctually" (TT12, 17, 18).

Sub-theme 5: Delivery system accessibility

"Delivery system accessibility" refers to test-takers' opportunities to familiarize themselves with the software system used to administer the computer-based written test of the EPT. Three test takers reported being unfamiliar with the delivery system prior to taking the test. For example, TT19 voiced her concern as follows:

> *"I don't know how the delivery system works before taking the test. I wish there were a mock system available for me to familiarize myself with it beforehand."* (TT19)

This lack of familiarity arises from the fact that the sample test paper is not made available in a format that mirrors the actual test delivery system. This concern highlights the need for a mock delivery system that enables test takers to gain greater familiarity with the technological aspects of the test environment.

III. Theme 3: Consistency

In this study, "consistency" is conceptualized as the degree of uniformity in testing practices across four stages: test development, test administration, rating, and score interpretation.

Sub-theme 1: Test development

The process of test development for the EPT includes two critical steps: item writing and pilot testing. The interview results will be organized around these two steps.

All teachers believed that item writing followed a consistent and uniform procedure. Initially, all item writers participated in "professional training" (TD2, 4, 6). The professor responsible for the training sessions has expertise in language testing and assessment and has been actively involved in relevant research for more than two decades. She possesses not only a strong theoretical foundation in test design and development but also extensive practical experience. A test developer recalled that the training sessions enhanced her item writing literacy:

> *"After training, I strived to ensure that the items I wrote were neither too difficult nor too easy. I also tried to avoid context where the information in the input materials was insufficient to determine the correct answer."* (TD6)

The training sessions equipped the item writers with necessary knowledge and skills to develop high-quality test items. Following initial item writing, the items underwent multiple rounds of review. The procedures of item writing and review are summarized as follows:

> *"College English teachers work in pairs during the item writing process. First, they write items independently, and then cross-check each other's work. Following the peer review, senior item writers conduct an additional round of review and make modifications to the items if necessary. Finally, all the items undergo a final round of review. I believe that the item writing and review procedures contribute to the high quality of the item bank."* (TD2)

To ensure the quality of the items, all items in the item bank were subjected to pilot testing. Over 5,500 test takers, drawn from the target population, participated in the pilot testing. Following the pilot testing, statistical analyses were performed to

obtain the parameters for each item in the item bank. These statistics were considered valuable for "ensuring the quality and appropriateness of the items before they are finalized for operational testing" (TD3).

Sub-theme 2: Test administration

This section will report the interviewees' perceptions of the administration of the EPT. Their views on equipment functionality, test environment, and test security measures will be presented.

Equipment functionality

Technical personnel were responsible for ensuring the proper functioning of the equipment both before and during the testing process. Prior to the test, technical personnel "thoroughly check all the equipment in each test room" (TA1) to prevent potential technical issues that might disrupt the administration of the test.

A majority of test takers reported that the equipment used for administering the EPT functioned well. However, one test taker noted that the assigned keyboard was not functioning properly before the start of the test. In response, the technical support staff "promptly replaced the keyboard" (TT11) for him.

The quality of test equipment is one of the concerns held by a test user. TU1 recommended that, conditions permitting, the university should regularly update the test equipment:

> *"Devices such as keyboards and monitors require regular updates to ensure optimal performance. If the keyboard is not functioning or not functioning properly, it could adversely influence test-takers' writing performance. Likewise, poor display quality on the monitor will make it difficult for test takers to see instructions, prompts, and input materials clearly."* (TU1)

Test environment

Overall, the test environment was the same for every test taker. Test takers reported that the test rooms were standardized. All the test takers took the test in the test rooms equipped with "the same equipment" (TT3, 10, 16). Four test takers considered the seating arrangement in the test rooms satisfactory. For example, TT1 observed that:

> *"During the test, test takers are seated separately with dividers placed between the seats. This seating arrangement makes me focus on my own computer monitors without interference from my neighboring test takers."* (TT1)

Interview responses revealed mixed perceptions about the noise levels in the test rooms. While TT9 reported that the test room was very quiet, four test takers complained about the noise generated by test equipment. For instance:

> *"The noise generated by keyboard and mouse use can be distracting during the test. I hope each test room could be equipped with silent keyboards and mice."* (TT12)

Additionally, two test takers reported that the dividers between the seats were ineffective in blocking the noise produced by test equipment, making it difficult for them to concentrate on the test content during the test.

Test security

Overall, the stakeholders expressed appreciation for the measures taken to safeguard the security of the EPT. The university-level test administrator highlighted the importance of maintaining confidentiality of the test content. She supported the use of multiple test forms, as this practice "prevents test takers from knowing the exact test content prior to test-taking" (TA1).

Additionally, to prevent item overexposure, items that had been used once were withheld from subsequent administrations "for a period of more than five years" (TU1, 2). In other words, if a test taker takes the EPT six times during his/her university years, he/she would not encounter the same items.

With respect to proctoring, all proctors participated in a training session before the start of the test to become familiar with the test procedures and their responsibilities. Before each session, proctors verified the IDs of the test takers. This procedure was implemented to ensure that "no one is impersonating another to take the test" (TD2). Each test room was staffed with two proctors to ensure discipline, and monitoring cameras were activated to oversee the entire test room. According to a test taker:

"Discipline in the test room is strictly maintained during the test. Test takers are not allowed to whisper to each other." (TT20)

Moreover, the university's *Code of Conduct for Undergraduate Students* served as a deterrent against cheating. Test takers who violate test-taking regulations may "face the risk of dismissal from the university" (TA1). Given that cheating could jeopardize test-takers' chances of obtaining a bachelor's degree, "it is unlikely that they would venture to take such a risk during the test" (TA2). Meanwhile, to decrease the possibility of cheating, the order of options in multiple-choice questions was randomized. This makes it difficult for test takers to "identify the specific answers chosen by others" (TA2; TU2).

Sub-theme 3: Rating

This section will report on the interviewees' perceptions of the rating methods and procedures for the EPT.

Rating methods

Test takers perceived the computer-based rating as objective and fair. The listening and reading subtests consist solely of multiple-choice questions. Test takers believed that using computer software for scoring these questions was "highly efficient" (TT4, 5, 16), ensuring both "impartiality" (TT15, 16, 20) and "accuracy" (TT5, 8, 12) in the rating process.

Teachers endorsed the integration of human and machine rating methods in assigning scores for test-takers' compositions. The use of AWE system was believed to "ensure consistency and reliability in the rating process" (TD5, 6). Human raters could address the limitations inherent in AWE systems to some extent. For instance, the ratings derived from these systems could be sensitive to "the writing format and linguistic errors in compositions" (TD1).

Two test users found the scores generated by the AWE system to be reliable. TU2 examined the rating performance of the system and found that the system did not exhibit central tendency in assigning scores. In other words, the system assigned scores across an entire range of possible scores, rather than over-relying on middle-

range scores. More importantly, the consistency between the scores assigned by human raters and the AWE system was "satisfactory" (TU2). Additionally, a professional team was responsible for the maintenance of the system (TU1).

Rating procedures

Teachers and test users believed that the procedures for rating both written and oral responses were effective in ensuring reliability. Prior to the rating process, all raters received training to "enhance the raters' understanding of the rating scale" (TD1) and to "maintain the same levels of leniency" (TD5) when assigning scores.

To ensure reliability of the rating of compositions, several measures were taken. First, the compositions were scored anonymously to "minimize rater bias" (TD4). Second, an AWE system was introduced to assist human raters. Specifically, following the automated rating process, human raters reviewed the ratings and paid attention to aspects such as "content relevance and the length of the compositions" (TD5; TU3). Human raters would adjust the machine-generated ratings if necessary. TD5 elaborated on the measures taken to ensure the consistency between the ratings assigned by human raters and the AWE system:

> *"If there is a large discrepancy between the two sets of ratings, a second human rater will review or re-score the compositions to ensure the accuracy of the ratings."* (TD5)

To ensure reliability of the rating of oral responses, two measures were implemented. First, each test room was staffed with two oral examiners. This helped reduce the risk of "rater bias" (TD1) in the rating procedures. Second, the speaking subtest was recorded for rating quality control. This measure allowed for "a review of the scores assigned by the examiners" (TD2). Meanwhile, the awareness that the speaking subtest would be recorded made the examiners "exercise greater caution" when assigning scores (TD6).

Sub-theme 4: Score interpretation

The EPT has been aligned with the *CSE*, allowing both test takers and test users to interpret test scores with reference to a national English proficiency standard. TU2, who was involved in EPT-CSE alignment, pointed out that:

> *"Following the release of the CSE, the test has been aligned to it. The alignment results indicate that test takers must achieve Level 5 on the CSE (CSE-5) to pass the test. By referring to the cut scores used for classifying test-takers' performance into corresponding CSE levels, we know whether test takers have attained CSE-5 and determine their pass or fail status on the test."* (TU2)

The above excerpt indicates that the alignment of the EPT with the *CSE* enables a consistent interpretation of the test scores. The EPT-CSE alignment also enhanced the test-takers' understanding of their English proficiency levels. Additionally, the *CSE* could serve as a reference for setting learning goals. For instance, if a test taker did not pass the EPT, the performance descriptors in the CSE-5 could serve as "learning goals" (TU3).

IV. Theme 4: Accountability

In this study, the theme of "accountability" refers to the policies, procedures, or practices that ensure the fairness of the EPT. The findings are presented in terms of four aspects: score review procedures, stakeholder engagement, fairness evaluation, and the stakeholders' responsibility for test fairness.

Sub-theme 1: Score review

Test takers were given opportunities for score review. After the test results are released, test takers who have doubts about their test results can consult test administrators. The contact information is made public in the registration notice. TA2 described the score review process as follows:

> *"If test takers want to know their specific scores on each subtest, they can contact me. I will then provide them with the email address of my colleague. Test takers can email her directly to inquire about the specific scores and ask questions such as 'why did I fail the test' or 'which of my skills do I need to improve'."* (TA2)

Nearly half of the test takers reported to be unaware that they could request a score review after receiving their test results. These test takers noted that neither the test specifications nor the registration notice explicitly outlined the procedures for requesting a score review. TT5 and TT7 described the procedures as "non-transparent". One test taker underscored the importance of score review to uphold test fairness:

> *"I think the availability of score review opportunities is essential for ensuring test fairness. Although score review may not result in a change of score, it offers psychological reassurance to me. It also makes me feel that I am treated fairly after receiving my test results."* (TT13)

Sub-theme 2: Stakeholder engagement

Two test takers, three teachers and one test user noted the lack of channels for gathering stakeholder input about the EPT and relevant testing practices. To address this, they recommended developing a platform to collect opinions and views from various stakeholder groups.

Opening the platform for stakeholder input was considered "highly beneficial" (TT8, 15; TD1, 4, 6; TU3). First, collecting feedback from stakeholders can "facilitate communication among different stakeholder groups" (TD4). Second, stakeholder engagement can serve as a catalyst for "continuous improvement in testing practices" (TD6). Lastly, creating opportunities for different stakeholders to voice their concerns and expectations of test fairness would contribute to a "democratic policy-making process" (TD1; TU3).

Sub-theme 3: Fairness evaluation

The following two paragraphs provide an overview of stakeholders' perceptions regarding the evaluation of test fairness. Two key issues emerged from the interviews: (1) who should be involved in fairness evaluation and (2) what aspects should be prioritized.

There was a divergence of views among the stakeholders regarding who should be involved in evaluating the fairness of the EPT. A test administrator believed that

the professors and graduate students from the language testing research team at the university should evaluate the fairness of the EPT from a professional perspective. The test administrator added that test takers, with "firsthand test-taking experience", could contribute to the evaluation of test fairness (TA2). However, a test user expressed concerns about the test-takers' ability to evaluate the fairness of the EPT. TU3 argued that "most test takers possess limited language assessment literacy". He believed that test takers might not be qualified to evaluate test fairness.

Regarding the question of "what aspects should be prioritized" in fairness evaluation, two test users with expertise in language testing assessment emphasized that "validity" should be the primary focus (TU1, 2). If there were validity issues related to score interpretation and use, evaluating the fairness of a test would be "futile" (TU1). The university's language testing research team conducted validation studies on the speaking and writing subtests of the EPT. According to TU1, these ongoing validation efforts were "essential for ensuring the fairness of the EPT and maintaining the professionalism of relevant testing practices".

Sub-theme 4: Responsibility for test fairness

Four test takers, two test administrators, five teachers, and three test users believed that ensuring test fairness required the joint efforts of all stakeholder groups involved in every aspect and assessment stage of the EPT.

Test users played a pivotal role in ensuring consistency and uniformity in testing practices. They were supposed to "develop the test specifications" (TD6) and "formulate the administration plan" (TD1) for the EPT.

Test developers were expected to fully leverage their expertise in language testing and assessment to develop a fair test:

> *"Test developers should draw on their expertise in language testing and assessment to ensure that the test results accurately reflect test-takers' English proficiency levels and are fair to all test takers."* (TU3)

Test content and test difficulty should be taken into consideration during test development. Specifically, the test content should be "free from bias" (TD6) toward

any group of test takers Meanwhile, the overall difficulty of different test forms should be balanced by referring to "the item difficulty parameters" (TD4) and conducting "expert reviews" (TA2).

Test takers also played a role in ensuring the fairness of the EPT. Prior to the test, they were expected to proactively seek test-related information through multiple sources (TT12, 15, 16, 18). The information can be accessed from: (1) official documents (e.g., test syllabus), (2) test registration notice, and (3) the university's online forum. Relevant test information includes, but is not limited to, test format, test content, rating criteria, registration deadline, and test procedures. During the test, test takers should "fully demonstrate their English proficiency levels" (TU3) and "adhere to the rules and regulations set by the test organizer" (TA2). If test takers had been treated unfairly during test-taking, they should approach test administrators to "express their grievances" (TA2).

Apart from test users, test developers, and test takers, other stakeholders also share the responsibility for ensuring test fairness. Technical support staff are expected to "update the hardware in the test rooms and optimize the test delivery system on a regular basis" (TA2). Proctors should be responsible for "verifying test-takers' identities, supervising discipline, and maintaining order in the test rooms" (TA2). Moreover, raters must ensure the consistency of scoring (TD2, 3, 6).

4.2.2.2 Factors influencing stakeholders' perceptions

This section aims to delineate four factors influencing stakeholders' perceptions regarding the fairness of the EPT: sociocultural factors, educational factors, institutional factors, and personal factors.

I. Theme 1: Sociocultural factors

"Sociocultural factors" refer to the sociocultural influences that shape stakeholders' perceptions of the fairness of the EPT and its associated testing practices. Analysis of

the interview data revealed two factors: (1) the sociocultural context regarding the importance of English proficiency in China and (2) the societal norms around English testing in China.

Sub-theme 1: Importance of English proficiency in China

English, as an essential tool for international communication, plays an important role in academia and professional settings in China. All the three test users deemed the development and administration of the EPT as "highly necessary" due to the importance of English in China (TU1, 2, 3). A test user stated:

> *"As an international lingua franca, English is undoubtedly one of the most important languages. We must ensure that the undergraduates have adequate general English proficiency upon graduation to meet the basic language requirements in workplace or academic settings."* (TU1)

The above excerpt indicates that the instrumental value of English leads the university to place great importance on its undergraduates' English proficiency. The EPT-as-exit-test policy is indented to motivate undergraduates to "continuously improve their general English proficiency during their years of study at the university" (TU3) and to "provide a solid foundation for their future academic pursuits and professional endeavors" (TU2).

Test takers acknowledged the importance of English proficiency for their future academic pursuits. More than half of the test takers interviewed planned to pursue further studies, either domestically or abroad, after completing their undergraduate studies. They highlighted the extensive use of English in various academic contexts, such as "reading literature" (TT3, 6, 10, 11, 13, 18, 19), "writing papers" (TT3, 6, 12, 15, 19), "programming" (TT20), and "delivering presentations at international conferences" (TT6, 15).

Test takers also considered English to be crucial for their future career aspirations. For instance, TT2 was preparing to join the research and development department of a company specializing in electronic displays. However, due to a self-

reported unsatisfactory general English proficiency, TT2 encountered language barriers during his internship in that company:

> *"I think I have mastered the specialized terminology in the field of electronics. However, I still struggle to understand technical reports in English in either written or oral form. I believe that my limited general English proficiency in reading and listening is to blame. I'm worried about whether I can secure an official position in this company after graduation."* (TT2)

Sub-theme 2: Societal norms around English testing in China

English testing practices in China could influence test-takers' perceptions of the fairness of the EPT. Four test takers argued that the formulation of the EPT-as-exit-test policy should take into consideration the testing practices of other Chinese universities with rankings comparable to the university. If other universities also used in-house English proficiency test results as a criterion for graduation eligibility, the test takers were more likely to accept and support the use the of EPT. For instance:

> *"I think it is reasonable for the university to require undergraduates to pass the English Proficiency Test before graduation. As far as I know, most universities in China impose specific English proficiency requirements for their undergraduates. For instance, some universities require undergraduates to pass the CET-4 before graduation."* (TT18)

Although national language tests (e.g., the CET-4) have been used to determine undergraduate students' eligibility for graduation in many universities in China, all the teachers, test administrators, and test users preferred the EPT over national English tests for making graduation decisions at the university. TU1 explained:

> *"The national English tests do not align well with the university's requirements for its undergraduates. After all, these nationwide tests are intended for undergraduate students in universities across China and are not appropriate in our local context."* (TU1)

II. Theme 2: Educational factors

"Educational factors" refer to the aspects of the educational system and learning resources, both before and after the students enter the university, that influence

stakeholders' perceptions of the fairness of the EPT. These factors include: (1) disparities in pre-university English education and (2) adequacy and effectiveness of the learning resources at the university.

Sub-theme 1: Disparities in pre-university English education

The quality of pre-university English education received by the undergraduates in the university exhibit regional disparities. According to four test takers, the regions of Shanghai, Jiangsu, and Zhejiang are renowned for the "high quality of their English education" (TT7, 14, 15, 20). Three test takers from Zhejiang province expressed confidence in passing the EPT and credited their "advantage" in English proficiency to the province's "high-quality English teaching resources" (TT14, 15, 20).

High-quality English education resources, however, are relatively scarce in certain regions of China. According to six test takers, the disparity in educational resources is evident in two aspects: (1) the limited availability of quality teaching resources in central and western regions compared to eastern regions and (2) the limited access to quality teaching resources in township-level administrative regions compared to their prefecture-level counterparts. The unequal distribution of educational resources influences undergraduates' English language proficiency upon entering university. Undergraduates from Shanxi and Henan provinces, for example, struggle with English listening comprehension, potentially due to an insufficient emphasis on listening skills in the high-school teaching syllabus (TT9, 12, 20).

Sub-theme 2: Adequacy and effectiveness of university learning resources

The perceived fairness of the EPT is influenced by the adequacy and effectiveness of on-campus English learning resources. All the teachers and test users believed that the learning resources were adequate for the undergraduates to improve their English proficiency and successfully pass the EPT. The university offers College English courses (Levels II-VI) for the undergraduates. These courses are designed to improve the undergraduates' general English proficiency in listening, speaking, reading, and writing. It should be noted that while the College English courses are not specifically

designed for test preparation purposes, the teaching and learning objectives "align closely with the university's established general English proficiency benchmark for its undergraduates" (TU1, 3). The university also offers a broad spectrum of EFL courses, which enable the undergraduates to "tailor their language learning trajectories to their individual needs and interests" (TD4).

Test takers believed that the College English courses were effective in enhancing their general English proficiency. TT18 reflected on his experience:

> *"College English courses have made me take English seriously. To pass the final exams of these courses, I focused on the improvement of my listening, speaking, reading, and writing skills. I also spend extra time studying English outside class."* (TT18)

Three test takers reported that the effectiveness of English language learning at university was influenced by undergraduates' English proficiency levels prior to entering university. For instance, TT9 faced challenges in fully understanding the lectures delivered by College English teachers upon entering the university:

> *"At the beginning of my first semester, my listening proficiency was rather limited. During the first lecture of the College English course (Level III), I found it difficult to understand the teacher's instructions during class."* (TT9)

The College English courses are delivered in English, which can indeed place a certain demand on undergraduates' English listening proficiency.

The university acknowledges the importance of English listening proficiency for freshmen and prioritizes the development of their listening skills. To address any weaknesses in the freshmen's listening skills, the university offers a "non-credit-bearing supportive English listening course" (TU2).

To meet the individual learning needs of the undergraduates, the university has developed an "AI-powered English learning platform" (TU2). A diagnostic test is administered quarterly through the platform to assess the undergraduates' listening, reading, and writing skills. Following the test, the undergraduates receive individualized feedback through the platform. Based on the feedback, the platform

regularly recommends tailored learning resources to the undergraduates. The undergraduates can also schedule one-on-one writing tutorials via the platform. Additionally, a speech-based chatbot embedded within the platform allows the undergraduates to practice oral English in a flexible and interactive manner (TU2).

III. Theme 3: Institutional factors

"Institutional factors" refer to the aspects of policies, personnel, resources, and services within the university that influence the stakeholders' perceptions of the fairness of the EPT. These factors cover three aspects: (1) the institutional policy, (2) the expertise of local personnel in language testing and assessment, and (3) the infrastructure for administering the EPT.

Sub-theme 1: Institutional policy

The policy support from the university leadership is fundamental in ensuring the development and use of the EPT. In 2013, amidst a nationwide reduction in class hours and credits for College English courses, maintaining the quality of undergraduate English education emerged as a pressing issue. The then dean of the School of International Studies proposed using a proficiency test to assess whether the undergraduates' English proficiency met the requirements set by the university. Since this proposal touched upon the undergraduates' eligibility for graduation, it gained "considerable attention from university leadership" (TU1). Following multiple rounds of discussions and deliberations, the university decided to develop and administer the EPT. In this collaborative effort, the Academic Affairs Office assumed responsibility for providing necessary "funding and technical support" (TA1). The School of International Studies was responsible for "test design, development, administration, scoring, and score reporting" (TU1). It is worth mentioning that without the university's policy support, the development and administration of the EPT would not be possible due to a lack of financial, technical, and human resources.

The transparency of test-related policies plays a pivotal role in shaping the test-takers' perceptions of the fairness of the EPT. The undergraduates are informed about the relevant test policies from the outset of their university journey:

> *"The university's Undergraduate School and the Academic Affairs Office have provided policy support for the English Proficiency Test, incorporating relevant policies into the Handbook for Undergraduate Students* (the *Handbook*)*. By reading relevant chapters of the Handbook upon enrollment, the undergraduates can learn about the English proficiency requirements set upon them and how to meet those requirements."* (TU1)

The *Handbook* serves as a comprehensive guide, covering the code of conduct, degree awarding criteria, disciplinary measures for violations of regulations, and so forth. During the orientation week, all freshmen are asked to take a test intended to assess their "understanding of the *Handbook*" (TT8). The test represents the university's commitment to ensuring that the undergraduates are adequately acquainted with the policies guiding their university journey.

Sub-theme 2: Local expertise in language testing and assessment

The expertise of the university's College English teachers in language testing and assessment plays a key role in ensuring the fairness of the EPT. The test development team is composed of College English teachers at the university and is led by "a group of language testing professionals with years of research and test development experience" (TU1). All core members hold doctoral degrees in linguistics and applied linguistics and have a research focus on language testing and assessment. A test user expressed confidence in the team's ability to develop an in-house English proficiency test:

> *"The language testing team at the university has conducted extensive research over the past years. I believe, with accumulated expertise, this team has the ability to design and develop a high-quality in-house English proficiency test."* (TU1)

The other members in the test development team "have experience in designing and developing English tests at the university, provincial, and even national levels" (TU2). The team operates with a clear division of responsibilities. According to TU1, the

core members are responsible for designing the structure of the EPT, training College English teachers involved in item writing, and evaluating the quality of test items. More importantly, they are responsible for conducting validation research to ensure the quality of the EPT and to promote fair decision-making. The remaining team members mainly take on the role of items writers.

The College English teachers at the university assume multiple roles in various testing practices. They are familiar with the university's English teaching syllabus and the undergraduates' English language profiles. As test developers, the teachers are able to minimize the influence of construct-irrelevant factors, such as test topics, on test-takers' performance. In addition, the teachers serve as composition raters and oral examiners. They have had "extensive experience in evaluating compositions and oral responses for national English tests" (TD2). One teacher also serves as a psychometrician. She possesses specialized knowledge in educational measurement. She is responsible for post-test score equating which ensures the comparability of the test results across test forms and administrations. Statistical analyses of test performance data are expected to offer valuable insights for "future item development and English instruction" (TD3; TU2).

Sub-theme 3: Infrastructure for administering the EPT

The infrastructure for administering the EPT plays a crucial role in shaping the test-takers' perceptions of test fairness. The standardized setup of the test rooms ensures a uniform and consistent test environment for all test takers. For instance, the layout is the same for each test room. A divider is placed between two neighboring seats. In addition, each test room is equipped with the same hardware, including computer monitors, audio equipment, keyboards, mice, and monitoring cameras. Besides, the same test delivery system is installed on all computers in the test rooms by the technical staff. As can be seen from the above, the university strives to mitigate test-takers' feelings of unfairness stemming from "the differences in test environment" (TU2).

The test-takers' perceived fairness of the EPT was influenced by the functioning of test equipment during the test. As detailed in Section 4.2.2.1, the majority of test takers reported no hardware malfunctioning during the test. Therefore, their performance remained unaffected by technical issues. Moreover, several test takers appreciated the stability and user-friendly interface design of the test delivery system. To safeguard against potential equipment failures, the university provides "two backup test rooms" (TA2). This contingency plan demonstrates the university's commitment to providing a level playing field for all test takers to demonstrate their English proficiency during the test.

The test takers expressed appreciation for the university's provision of two support services. The first is the service of registration notification. Prior to test administration, test administrators post the registration notice on the school's official website. This information is then forwarded by administrative staff from the Academic Affairs Office to the undergraduates' information service platform. The Student Union also plays an active role in disseminating the registration notice. By utilizing social media platforms popular among the test takers, the Student Union provides an additional and often "more accessible" channel for test takers to receive registration notice (TT3, 6, 7). The registration notification service proves to be effective in keeping the test takers well-informed about the registration details. The second is the inter-campus shuttle service. The university provides a shuttle bus service to facilitate the transportation of test takers from their respective campuses to the test location (which is in the main campus). The inter-campus shuttle service exemplifies the university's "care and respect for every test taker" (TT6, 11, 12; TA2). Moreover, this service demonstrates the university's commitment to providing "equitable test-taking opportunities" for all test takers (TA1, 2).

IV. Theme 4: Personal factors

"Personal factors" refer to individual characteristics, beliefs, experiences, and circumstances that influence test-takers' perceptions of test fairness. These factors include test-takers' language proficiency levels and their beliefs about the EPT.

Sub-theme 1: Language proficiency

Four test takers perceived the difficulty level of the EPT as appropriate. For these test takers, the EPT was not a barrier for graduation. For example, TT17 passed the written test of the EPT on her first attempt. She regarded the EPT as an attainable goal for the undergraduates:

> *"I believe that the goal of passing the English Proficiency Test is achievable. For me, the test exhibits a moderate level of difficulty. While test takers with lower English proficiency may face some challenges in passing the test, they can still pass the test with continuous learning."* (TT17)

Passing the listening and reading subtests of the EPT presents a challenge for some test takers. As mentioned in Section 3.2.1, listening and reading subtests have equal weighting, each contributing 30% to the final score. The test takers must achieve a minimum score of 36 out of 60 to pass the two subtests. Even if a test taker scored a perfect 30 in the reading subtest, scoring 0 in the listening subtest would still result in test failure. If separate passing thresholds were established for the listening and reading subtests, it would likely increase the difficulty of passing the listening subtest for many test takers:

> *"The absence of a specific passing score for the listening subtest makes passing the English Proficiency Test less difficult for students with weak listening skills, including myself. If such a requirement were in place, I might fail both the listening and reading subtests due to my limited English listening proficiency."* (TT2)

The university's decision of not setting separate passing scores for the listening and reading subtests was made in acknowledgment of the fact that the undergraduates from certain provinces might "have limited exposure to systematic listening training before entering the university" (TU1). The decision took into consideration the

disparities in the quality of English education that the undergraduates had received prior to university admission.

Sub-theme 2: Beliefs about the EPT

Test-takers' beliefs about the EPT can influence their perceptions of its fairness. The fairness of the EPT, a criterion-referenced test, is widely recognized among its test takers. The threefold reasons for this recognition are outlined as follows. First, there is no pass rate restriction:

> *"I think the English Proficiency Test is fair. There is no predetermined pass rate, and whether I can pass the test depends solely on my test performance."* (TT15)

Second, the EPT does not create a competitive atmosphere among its test takers. This is because test takers are only informed about their pass or fail status for each subtest rather than the specific scores. If specific scores were disclosed and if results of the test contributed to the Grade Point Average (GPA), undergraduates may "retake the test multiple times to achieve a high score" (TT15), which would further "intensify the competition among undergraduates" (TT16).

Third, the EPT can exert a positive washback. Overall, test takers endorsed the EPT-as-exit-test policy, describing the test as "highly beneficial" for enhancing their general English proficiency (TT16, 19, 20). TT16 noted that if the EPT were not an exit requirement, she would not have prioritized English learning throughout her university studies. TT19 further added that, in the absence of the EPT, many undergraduates might "terminate the English learning journey" after completing the College English courses. Additionally, TT20 reported an improvement in his speaking proficiency during test preparation:

> *"I struggled with my spoken English for a long time. I practiced speaking intensively every day before the test. On the test day, I could respond to the examiners' questions very fluently. Thus, I strongly believe that without the English Proficiency Test, I would not be able to improve my spoken English."* (TT20)

4.3 Chapter summary

This chapter presents the results of fairness evaluation of the EPT from the perspectives of psychometrics (Section 4.1) and stakeholders (Section 4.2) across four dimensions: comparability, accessibility, consistency, and accountability.

Regarding comparability, statistical analyses revealed a relatively low proportion of items and testlets exhibiting DIF and DBF across the four listening and reading test forms. Two test forms exhibited DTF. Content analysis is needed to determine whether the identified DIF, DBF, and DTF introduce bias to the listening and reading subtests. While the characteristics of the listening input materials were found to be comparable in terms of lexical complexity, syntactic complexity, discourse complexity, and speed of delivery across the four test forms, the reading input materials did not show comparability in syntactic complexity, discourse complexity, and readability. This incomparability could lead to variations in reading task difficulty across test forms. The questionnaire survey revealed that test takers generally believed that they had comparable opportunities to fully demonstrate their English proficiency levels during the test. Qualitative findings indicate that the teachers (also test developers) and test users believed that: (1) the construct measured by the EPT was comparable across test-taker groups and (2) the test scores were comparable across test forms.

As for accessibility, results from the questionnaire survey and interviews indicate that the EPT ensured transparency in test-related information. Additionally, test takers were provided with sufficient opportunities to take the EPT and adequate learning resources to effectively enhance their general English proficiency. Moreover, the university provided shuttle service to address a logistical challenge arising from the multi-campus arrangement of the university. Lastly, a few test takers suggested the development of a mock delivery system that would help first-time test takers to become familiar with the delivery system before the test day.

With regard to consistency, test takers who participated in the questionnaire survey believed that the EPT was administered in a consistent manner. However, the interviewees offered various perspectives on the consistency of testing practices across different stages of the test, such as test development, test administration, scoring, and score interpretation.

In terms of the dimension of accountability, results from the questionnaire survey indicate that test takers can request score reviews and report instances of unfair treatment experienced during the test. The interview results revealed a mix of compliments, concerns, and suggestions regarding score review, stakeholder engagement, fairness evaluation, and the responsibilities of different stakeholders in ensuring test fairness.

This chapter also presents findings on the sociocultural, educational, institutional, and personal factors that influenced the stakeholders' perceived fairness of the EPT (see Section 4.2.2.2). These factors represent the reasoning or justification for the stakeholders' perceptions of the fairness of the EPT.

Chapter 5 Discussion

In this chapter, evaluation results of the EPT's fairness from both psychometric and stakeholders' perspectives will be discussed in Section 5.1. Informed by the research findings, Section 5.2 will propose a model of test fairness evaluation. A summary of this chapter will be provided in Section 5.3.

5.1 Evaluation of test fairness

5.1.1 Evaluation from the psychometric perspective

Test performance data and input materials were examined to investigate whether the EPT is fair from the perspective of psychometrics (RQ2.1). Specifically, item-, testlet-, and test-level score comparability across humanities and science groups was examined by conducting DIF, DBF, and DTF analyses. Difficulty comparability of input materials across multiple forms of the listening and reading subtests was examined by analyzing their characteristics. Overall, the findings suggest that the EPT was generally fair from the psychometric perspective. The following sections will discuss relevant findings revolving around the dimension of comparability.

5.1.1.1 Test score comparability across test-taker groups

This study employed the SIBTEST procedure and MG-CFA to analyze DIF/DBF and DTF across academic background in the listening and reading subtests of the EPT. Overall, the statistical analyses indicate the presence of DIF, DBF, and DTF in both subtests, suggesting that item-level DIF and testlet-level DBF can manifest themselves at the test level and lead to DTF. Nevertheless, analyses of the problematic items and testlets indicate no systematic bias favoring either the humanities or the science group.

I. DIF and DBF investigations of the listening subtest

The science group demonstrated a slightly better performance than the humanities group across all listening test forms, except for Form 3. No significant differences in test performance between the humanities and science groups were observed for Forms 1, 3, and 4. However, for Form 2, the Mann-Whitney *U* test indicates that the science group performed significantly better than the humanities group (U = 39,580.00, p = .04), with a small effect size (r = .052). These findings seem to align with previous studies suggesting that sciences test takers are generally better at listening than humanities counterparts in the context of China (He, 2022; Yang et al., 2022). Nevertheless, given that the English tests examined in these two studies differ from the EPT, this observed alignment is only coincidental.

Among the 40 items associated with the Short Conversation (SC) section of the listening subtest, only two items (Item 6 in Form 2 and Item 10 in Form 3) exhibited DIF, accounting for 5% of all stand-alone SC items. The DIF ratio is relatively low compared to what has been reported in previous DIF studies on listening comprehension between test-taker groups in humanities and science (Aryadoust et al., 2024; Chen, 2013; He, 2022; Pae, 2004; Semiyari & Ahangari, 2022; Xiao, 2013; Zhang & Jin, 2012). Specifically, two items displayed uniform DIF, both favoring the humanities group. One of these items also displayed nonuniform DIF, suggesting that the cross-group difference in the probability of answering it correctly varies across listening ability levels. As the presence of DIF does not necessarily imply that the individual items are biased (Abbott, 2007), in the following two paragraphs, the researcher will try to explain DIF occurrence by analyzing the topics of the listening input materials, the listening subskills assessed by the two DIF items, and the characteristics of the input materials. In accordance with the confidentiality protocols established between the test provider and the researcher, the original input materials and items will not be displayed in the following discussion.

In the SC for Item 6 (Form 2), a student tells her professor that some students in the class did not receive the reading assignment. The professor explains that he brought copies only for those on the class list, and the student suggests sharing the assignment among classmates. According to schema theory (Carrell & Eisterhold, 1983), test takers with background knowledge of classroom settings should be able to comprehend this input material. Since all test takers are college students, regardless of their academic background, they are expected to be familiar with the content of the dialogue. As stated by the test specifications of the EPT and the overall scale of language ability of the *CSE*, test takers with listening ability at the level of CSE-5[1] should be able to understand spoken language on general topics. Apparently, the content of this input material aligns with the specifications and the *CSE*. Moreover, Item 6 assesses test-takers' ability to identify the main idea or important details from a classroom-related context. According to the CSE, students at the level of CSE-5 should "obtain main ideas and supporting details" (Ministry of Education of the People's Republic of China & National Language Commission of the People's Republic of China, 2018, p. 6). Thus, the listening ability assessed by this item aligns with the language ability description for CSE-5. Additionally, the characteristics of the SC do not appear to introduce academic background-related bias. To answer the item correctly, test-takers must understand key phrases (i.e., *reading assignment*, *class list*, and *spare copies*)—expressions frequently encountered in College English courses. A lexical frequency analysis of the script using VocabProfiler (Web VP Classic v.4 version; Cobb, n.d.) showed that 93.18% of the words belong to the K1 and K2 frequency bands, while only one word—*assignment*—is drawn from the AWL. The results suggest that performance on Item 6 depends on test-takers' listening ability rather than their specialized academic knowledge.

In the SC associated with Item 10 (Form 3), the woman, noticing that the man appears upset, asks about the cause of his distress. When the man refuses to discuss

[1] As mentioned in Section 3.2.1, the test takers have to reach CSE-5 to pass the listening subtest.

his feelings, the woman encourages him to express his emotions, emphasizing the negative consequences of suppressing feelings. This dialogue focuses on emotions—a general topic familiar to all test takers of the EPT. Meanwhile, the content of the input material aligns well with both the test specifications and the listening ability descriptors for CSE-5. From the perspective of schema theory (Carrell & Eisterhold, 1983), the content of the input material is unlikely to advantage either the humanities or science group. Additionally, Item 10 assesses test-takers' ability to understand the key part of the woman's suggestion—encouraging the man to talk about his feelings rather than suppress it. The listening skill assessed by Item 10 aligns with the descriptor for CSE-5, which states that students at this level should be able to "obtain main ideas and supporting details" when listening to scripts on general topics (Ministry of Education of the People's Republic of China & National Language Commission of the People's Republic of China, 2018, p. 8). Moreover, analysis of the listening script's characteristics reveals no systematic causes contributing to nonuniform academic background DIF in Item 10. The script contains no AWL vocabulary and consists of 97% K1 and K2 words, indicating its use of everyday language rather than specialized terminology. To correctly respond to Item 10, test takers must understand both the literal and figurative meanings of idiomatic expressions in the SC such as *let off some steam* and *keep one's feelings pent up*. Obviously, successful response does not depend on specialized academic knowledge. The above discussion helps explain the absence of systematic performance advantage for either the humanities or science group. Despite this, the content analysis suggests that Item 10 does not systematically favor or disfavor test takers based on their academic background.

Among the 10 passage-based testlets used in the four forms of the listening subtest, four (40%) were identified as having DBF. This proportion is lower than the findings reported by Yang et al. (2022), where two out of three testlets (66.67%) demonstrated DBF. Of the four flagged testlets in the current study, the second

listening passage (LP2) in Form 1 displayed uniform DBF which functioned in favor of the humanities group, while the long conversations (LCs) in Form 3 and Form 4 and the third listening passage (LP3) in Form 4 favored the science group. The four testlets also exhibited nonuniform DBF, indicating that the cross-group differences in the probabilities of answering these testlets correctly varied across test-takers' listening ability continuum. The following four paragraphs will be devoted to explaining the presence of DBF in each of the four testlets.

The second listening passage (LP2) in Form 1 deals with the global increase in overweight children. LP2 talks about a variety of contributing factors of this issue (including unhealthy food consumption and sedentary urban lifestyle) and potential solutions on the part of individuals and governments, including maintaining a healthy diet, increasing physical activity, breastfeeding for infants, and implementing government interventions. As test takers of the EPT are frequently exposed to health-related expressions in College English courses, they are expected to be familiar with this topic. Given that obesity is a widely discussed societal issue in China, no systematic relationship appears to exist between the content of this passage and the directions of DBF. Moreover, the five items associated with LP2 showed no evidence of DIF. Item 1 assesses test-takers' ability to recall factual details, specifically regarding the statistics about overweight children under the age of five. Item 2 assesses the subskill of retrieving detailed information by requiring test takers to recall the three criteria used by doctors to determine an individual's weight status. To answer Item 3 correctly, test takers are required to identify factual information not mentioned in the passage. Item 4 assesses the ability to recognize solutions mentioned in the passage. Item 5 requires test takers to synthesize key points in the passage, particularly the shared responsibilities of individuals and governments in addressing the issue. The targeted listening ability aligns with the test specifications, which require test takers to grasp important details and summarize key information presented in the listening material. The ability assessed also aligns with a CSE-5

descriptor from the overall listening comprehension scale, which states that individuals should be able to "obtain main ideas and supporting details" (Ministry of Education of the People's Republic of China & National Language Commission of the People's Republic of China, 2018, p. 8). Furthermore, the speech rate of the recording (130 words/minute) complies with both test specifications and *CSE* requirements. The lexical analysis shows that 88.66% of the words in the passage fall into the K1 and K2 frequency bands, indicating that the passage is generally accessible to most test takers. However, the presence of 4.26% AWL words and 7.09% off-list words could pose challenges for some test takers to fully understand the passage. According to schema theory (Carrell & Eisterhold, 1983), test takers with prior knowledge of obesity-related terminology might have an advantage in this testlet. That being said, LP2 does not introduce systematic bias in favor of either the humanities or the science group.

In the long conversation (LC) in Form 3, a customer places an order for a large half-and-half pizza with different toppings (pepperoni and mushrooms on one half, and Italian sausage and green peppers on the other). The pizzeria has a special promotion, which includes free breadsticks and a $3 coupon with the purchase of any large pizza and drink, and promises delivery within 30 minutes. Food ordering and delivery is a familiar topic for the test takers of the EPT for two reasons. First, western cuisine, including pizza, is consistently featured in English textbooks for Chinese secondary and college students to learn about Western culture. Second, there are many pizzerias in the city where the EPT is administered. As such, it is reasonable to assume that test takers, regardless of their academic background, possess sufficient topical knowledge to aid their understanding of this dialogue. Third, the five individual items associated with LC did not exhibit DIF. Item 1 assesses test-takers' ability to recognize the order details, including the specific pizza type and toppings. Item 2 examines the ability to make inferences based on contextual clues and menu descriptions. Item 3 requires test takers to recall the details of the promotion offered

by the pizzeria. Item 4 assesses the ability to recall and identify specific personal information (e.g., the customer's name, address, and phone number) provided in the conversation. Lastly, Item 5 asks test takers to evaluate four statements based on the content of the conversation and identify the one that is false. Correctly answering this item requires a comprehensive understanding of the conversation. The listening abilities assessed by these items align well with the test specifications and the ability descriptors for CSE-5. A lexical frequency analysis shows that the script consists of 82.5% K1 and K2 words and 1.07% words from the AWL, indicating that everyday language is used in the dialogue rather than specialized terminology. However, the script contains 16.43% off-list words, including seafood-related vocabulary (e.g., *clams*, *shrimp*, and *squid*) and pizza-related vocabulary (e.g., *pepperoni*, *toppings*, *mushrooms*, and *peppers*). It should be pointed out that the stem and options of each item do not include these words. Although familiarity with seafood- and pizza-related vocabulary might provide some advantages in comprehending the conversation, it can be claimed that LC demonstrates no systematic bias favoring either the humanities or the science group.

The long conversation (LC) in Form 4 is between a woman and her husband, who is mentally and physically exhausted despite having just returned from a vacation in Florida. The man, a 60-year-old employee, has to commute from New England area to New York for work. In the conversation, he describes a near accident caused by his absent-mindedness during a drive to New York. In response, his wife suggests that he discuss the possibility of working in New York with his employer to reduce the need for frequent travel. The man then reflects on his past work experience with the company and expresses frustration with his current situation. The conversation concludes with the husband deciding to have some milk and take a rest. The topic of the conversation is generally familiar to college students in China, as the themes of stress, fatigue, career challenges, and family support are universal and relatable. Test takers from both the humanities and science groups are able to engage

with and understand this dialogue. Furthermore, Item 2 and Item 4 in this testlet exhibit DIF, potentially leading to the DBF in the LC. Item 2 exhibits only nonuniform DIF, while Item 4 demonstrates both uniform DIF (favoring the science group) and nonuniform DIF. The testlet comprises five items assessing a variety of listening comprehension subskills. Item 1 requires test takers to infer the relationship between the two speakers. Item 2 assesses the ability to synthesize information and make inferences about the man's vacation. Item 3 asks test takers to infer the man's mental state based on his account of the driving experience. Item 4 requires test takers to draw conclusions from multiple statements about the man's job and his feelings toward it. Finally, Item 5 assesses the ability to infer what the man is likely to do next based on contextual clues presented toward the end of the conversation. The listening subskills targeted by this testlet mirror the test specifications and ability descriptors for CSE-5. Lexical analysis reveals that the script consists of 91.56% K1 and K2 words, 0.62% AWL words, and 7.81% off-list words (primarily names for people and cities). The high proportion of high-frequency words ensures that the conversation is accessible to most test takers from both the humanities and science groups. As can be seen from the above, performance on LC depends primarily on test-takers' listening ability rather than their specialized academic knowledge. In other words, LC does not appear to introduce systematic bias at the testlet level that would advantage either the humanities or the science group.

The third listening passage (LP3) in Form 4 describes the advances in face transplant surgery. The surgery, unimaginable just a few years ago, now provides hope for those with severe facial injuries, such as accident victims and wounded soldiers. While some recipients have regained the ability to eat, speak, and appear in public without attracting attention, experts caution that face transplants will remain rare due to the high risks involved and the need for lifelong medication to prevent rejection. Considering the topic of this passage, LP3 is presumably more familiar to the science group. Nonetheless, the presence of nonuniform DIF in this testlet

indicates an interaction between students' abilities and their group membership. Turning to the items associated with LP3, Item 1 assesses students' ability to make inferences about a surgeon's statement, while the remaining four items assess the ability to identify detailed information. The listening subskills targeted by this testlet align with the test specifications and ability descriptors for CSE-4 and CSE-5. Among the five items, Item 4 is the only one that displays both uniform and nonuniform DIF. The presence of DIF in Item 4 may cause LP3 to display DBF. Furthermore, lexical analysis reveals that 82.95% of the words in the passage belong to the K1 and K2 frequency bands, suggesting that the passage is generally accessible to most test takers. However, the presence of 3.79% AWL words and 13.26% off-list words may present challenges for some test takers to fully understand the passage. As can be seen from above, students' performance on this testlet is determined by their listening ability rather than their group membership. Thus, LP3 does not introduce systematic bias in favor of either the humanities or the science group.

This study also revealed that testlets exhibiting DBF in one listening test form did not show DBF in the anchored test forms. For example, while LP2 and LC showed no evidence of DBF in Form 2, they displayed DBF in Form 1 and Form 3, respectively[1]. Similarly, LP3 showed DBF in Form 4 but not in Form 3[2]. The findings indicate that the detection of DBF is dependent on the test performance of a specific sample of test takers. In other words, DBF detection within a single test form examines whether the testlets function well across test-taker subgroups only for that particular sample. This finding aligns with Huggins's (2014) caution that items demonstrating DIF in one test form may not necessarily exhibit DIF at the linked test level. Therefore, before making decisions about removing DBF testlets from the item bank, further research is warranted to determine whether these testlets exhibit DBF at the linked test level.

---

[1] Form 1 and 2 share five common items associated with a listening passage (LP2). Form 2 and Form 3 share five common items associated with a long conversation (LC).

[2] Form 3 and 4 share five common items associated with a listening passage (LP3).

II. DIF and DBF investigations of the reading subtest

Overall, based on the mean scores and the results of the Mann-Whitney *U* tests, the humanities and science groups demonstrated comparable reading performance across the four test forms (see Table 4.1)—a finding consistent with Xiao's (2013) study. However, the finding diverges from earlier studies, which reported that science students outperformed their humanities counterparts (Alavi et al., 2011; Brati et al., 2006; Semiyari & Ahangari, 2022; Yang, 2022) or, conversely, that humanities students performed better than their science counterparts (Song et al., 2015). These contrasting results are understandable given that the English tests examined by these studies differ in task types and difficulty levels.

Two out of the nine passage-based testlets (22.22%) used in the four test forms were identified as displaying DBF. This percentage falls within the range of previously reported DBF ratios. Compared with earlier studies (see Appendix 2), this percentage is lower than that reported by Chen and Zeng (2021) and Yang et al. (2022), but slightly higher than that that by Song et al. (2015). Specifically, the second reading passage (RP2) and the Banked Cloze task (BC) in Form 2 displayed both uniform and nonuniform DBF. While the uniform DBF suggests an advantage for the science group in both testlets, the nonuniform DBF indicates that this advantage is not consistent across all ability levels of test takers.

The second reading passage (RP2) in Form 2 is about driverless cars. It presents both promises and challenges of autonomous vehicle technology discussed by industry experts. The content area of this passage seems to align more closely with the interests of the science group. According to schema theory (Carrell & Eisterhold, 1983), it is reasonable to regard passage topic effects as a potential source of DBF in RP2. However, all test takers, regardless of their academic background, encountered technology-related topics in College English textbooks before test-taking. Meanwhile, news about self-driving cars is prevalent on social media platforms in China.

Therefore, students from the humanities group are unlikely to be entirely unfamiliar with this topic. In this regard, autonomous driving technology represents a general topic and does not inherently favor or disfavor either academic group. Among the five items associated with this passage, Item 3, which requires test takers to understand the contextualized meaning of the phrase "driver engagement", displayed a large uniform and nonuniform DIF. Consistent with the findings from previous studies (Abbott, 2007; Song et al., 2015), the presence of DIF may contribute to DBF at the testlet level. In contrast, the remaining items showed no evidence of DIF. These items assess various reading subskills. Item 1 assesses the ability to infer the underlying meaning of a statement made by an expert. Item 2 asks test takers to identify the stance of experts regarding the development and feasibility of driverless cars. Item 4 assesses the ability to infer the meaning of an unfamiliar word (i.e., *underaroused*) based on contextual clues. Finally, Item 5 assesses the subskill of understanding the main ideas of an expert's perspective. The reading subskills assessed by these items align with the requirements outlined in both the specifications and the *CSE*. Additionally, a lexical analysis reveals that 80.94% of the words in the passage fall within the K1 and K2 frequency bands, suggesting that the majority of the vocabulary is accessible to test takers from both academic groups. However, the passage also includes 7.33% AWL words (e.g., *vehicle*) and 11.73% off-list words (e.g., *autonomous*, *battery*, *prototype*, and *sensors*). While familiarity with these technical terms may facilitate reading comprehension, such familiarity is not inherently linked to test-takers' academic background. Thus, RP2 does not introduce systematic testlet-level bias that advantages either the humanities or the science group.

The other testlet that was identified as displaying uniform and nonuniform DBF consists of 10 items associated with the Banked Cloze task (BC) in Form 2. This task features a passage with 10 blanks and requires test takers to select the most appropriate word for each blank from a word bank provided after the passage. The passage discusses the perceptions of boarding schools in Britain. Its focuses on the

educational systems and historical changes place it within the humanities domain. In line with schema theory (Carrell & Eisterhold, 1983), it was hypothesized that the humanities group would perform better on tasks related to education and history compared to their science counterparts. However, the presence of nonuniform DBF in the BC task does not support this hypothesis, suggesting the existence of an interaction between reading ability and group membership. The rejection of this hypothesis may be attributable to the unfamiliarity of test takers from both groups with British culture and education systems. Moreover, of the 10 items in this task, Item 10 was flagged as having uniform DIF, while the other items showed no evidence of DIF. This finding echoes previous studies (Abbott, 2007; Song et al., 2015) which demonstrated that item-level DIF can lead to DBF at the testlet level. Item 10 assesses the ability to correctly use adverbs to construct a grammatically correct and contextually appropriate sentence. This item does not appear to systematically favor or disfavor either the humanities or the science group, as success in answering it depends more on reading ability than on academic knowledge. Successful completion of the BC task requires test takers to demonstrate both organizational competence (e.g., vocabulary and grammar knowledge) and reading subskills such as understanding detailed information, making inferences, and identifying main ideas. The abilities targeted by this task align well with the assessment requirements outlined in the test specifications and the CSE-5 ability descriptions. Furthermore, lexical frequency analysis indicates that the passage comprises 88.62% K1 and K2 words and only 2.07% AWL words. To facilitate comprehension, Chinese translations are provided for off-the-list words such as *spartan*, *sentient*, and *sadist*. Although the BC task demonstrates no academic background-based bias, the topic of the passage may introduce cultural bias against all intended test takers of the EPT.

Another observation is that the shared testlet between Form 1 and Form 2—RP2—exhibited DBF in Form 2 but not in Form 1. DBF analysis conducted on a

single test form only examines how well a testlet functions for a specific test-taker sample. Therefore, the detection results may vary across different samples. As previous studies indicate, testlets displaying DBF in one test form may not necessarily exhibit DBF at the linked test level (Aryadoust et al., 2024; Huggins, 2014). Further investigation is needed to determine whether this testlet demonstrates DBF at the linked test level. Moreover, the performance data analyzed in this study represents only a small subset of the performance data collected over past administrations. To enhance the robustness and generalizability of the findings, future studies should cross-validate the DBF results reported in this study using a larger sample size.

III. DTF investigation of the listening and reading test forms

At the test level, measurement invariance was established at the test level for Forms 1 and 3, but not for Forms 2 and 4 across the humanities and science groups. The results indicate the presence of DTF in Forms 2 and 4. In other words, these two test forms did not measure the intended construct equivalently between the two groups. The existence of DTF can pose a threat to both score interpretation and score-based decision-making between the humanities and science groups. However, the accuracy of DTF detection warrants further investigation, as the statistical analyses were conducted using real test performance data and was likely to be limited by substantially unbalanced (though representative) sample sizes for the humanities and science groups (see Table 4.1).

This study yielded mixed findings regarding the relationships among DIF, DBF, and DTF. For Form 1, the presence of DBF in the second listening passage (LP2) did not result in test-level DTF. Similarly, for Form 3, DIF in Item 10 and DBF in the long conversation (LC) did not translate into DTF at the test level. These results corroborate the findings from previous studies (e.g., Elosua, 2024; Min & He, 2020; Pae, 2004). In contrast, a DIF item (Item 6) led to DTF in Form 2. Pae and Park (2006) also observed that DIF can lead to DTF. For Form 4, two DBF testlets (LC & LP3)

contributed to the test-level DTF. The intricate DIF–DTF and DBF–DTF relationships require further empirical inquiry.

5.1.1.2 Input material comparability across test forms

To address RQ1.2 (i.e., "To what extent are the input materials comparable in terms of difficulty across different test forms, as indicated by their input characteristics?"), the characteristics of the listening and reading input materials from four listening and reading test forms were extracted using tailor-made schemes (see Section 3.6.1.2 for details). To determine whether significant differences existed in the characteristics of the listening input materials across the four test forms, a multivariate Kruskal-Wallis test, a permutation test, and separate Kruskal-Wallis tests for each measure were conducted. In contrast, the characteristics of the reading input materials were analyzed descriptively rather than using inferential statistical methods due to the small sample size for each test form (n = 3).

Results from both descriptive and statistical analyses revealed that the linguistic complexity of the listening input materials was generally comparable across the four test forms in terms of lexical complexity, syntactic complexity, discourse complexity, and speed of delivery. Previous studies have suggested that these characteristics of the listening input can contribute to the difficulty levels of listening tasks and the cognitive processing load placed on test-takers (e.g., Brunfaut, 2016; Brunfaut & Révész, 2015; Gui, 2000; He et al., 2018; Pan, 2021; Pan & Fan, 2021). Accordingly, the observed comparability of input characteristics can contribute to ensuring comparable task difficulty across different test forms. In other words, the listening input materials are less likely to impose incomparable cognitive processing demands on test-takers across different test forms. In this regard, the listening subtest can be considered fair for test-takers who registered for different test sessions.

The reading materials of the EPT exhibited variations in input characteristics (including syntactic complexity, discourse complexity, and readability) across the

four test forms, with lexical complexity being the only exception. This finding aligns with Liao (2020) who also reported partial comparability in the syntactic and readability characteristics of reading passages across IELTS test forms. Previous studies have shown that the characteristics of reading input materials can influence the overall difficulty of reading tasks (Liao, 2020; Skehan, 1998; Zeng, 2022). Consequently, the observed incomparability in input characteristics may lead to variations in the difficulty levels of reading tasks across test forms, raising serious fairness concerns for test-takers who registered for different test sessions. Repeat test-takers who did not pass the EPT on their first attempt may be particularly sensitive to these variations in reading task difficulty and may perceive the test as unfair. It should be noted that the scaled scores (as opposed to raw scores) are used to counterbalance potential variations in the overall difficulty of test forms and determine the pass or fail status of the test-takers. However, as is often the case, information regarding the linking design and post-test equating procedures may not be disclosed to test takers. Even if such information were made available, test-takers might struggle to fully understand these technical details due to limited levels of language assessment literacy. As a result, the perceived variations in text difficulty levels across reading test forms could undermine the credibility of the EPT, particularly among repeat test-takers. That said, since the input materials from only four test forms were examined for comparability, further research is needed to validate the conclusions drawn here.

5.1.2 Evaluation from the stakeholders' perspective

RQ2.1 examines stakeholders' perceptions of the fairness of the EPT. To address this research question, a mixed-methods approach was employed, comprising a questionnaire survey administered to 1,646 test takers and semi-structured interviews with the representatives of four stakeholder groups: test takers (n = 20), teachers (n = 6), test administrators (n = 2), and test users (n = 3). The findings suggest that the four stakeholder groups generally accepted the use of the EPT and perceived it to be

fair overall. The following sections will discuss relevant findings in relation to the four dimensions of test fairness (i.e., comparability, accessibility, consistency, and accountability), as outlined in Section 2.3.

5.1.2.1 Stakeholders' perceived fairness of the EPT

Findings from the questionnaire survey and interviews provided insights into two key aspects of comparability: (1) opportunities for test-taker subgroups to demonstrate English proficiency and (2) test results across different test forms. Regarding opportunities to demonstrate proficiency, test takers who participated in the questionnaire survey believed that they had equal opportunities to demonstrate their English proficiency during the test. This belief, according to teachers and test users involved in test development, could be attributed to two reasons. First, the test content of the EPT features a balanced representation of diverse topics that are relevant to university context. This stands in contrast to the findings from previous studies which highlight concerns about biased test content that does not align well with university context (Yao, 2023) or may favor test takers with specific academic background (Jang, 2002). Second, the EPT is a full-skill test that does not allow test-takers' strength in one skill area to compensate for their weaknesses in others. In comparison with the GSEEE which assesses only reading and writing skills (see Song, 2018), the EPT does not unfairly favor test takers who excel in reading and writing but have weaker listening and speaking skills.

With regard to the comparability of test results across different test forms, teachers and test users maintained that test takers with the same proficiency levels would achieve comparable scores, regardless of which test session they attended. With an anchor-item design during test assembly and post-test equating procedures, the test results would not be influenced by potential variations in the difficulty levels of different test forms. As can be seen from the above, ensuring comparability of test results across multiple test forms is a top priority for the teachers and test users

involved in designing and developing the EPT. According to Fan and Jin (2013), test equating constitutes the most serious concern for test fairness and warrants serious attention from test developers. It can be concluded that the EPT is fair from a psychometric perspective in that the test results are comparable across test forms, test administrations, and even academic years.

Results from the questionnaire survey administered to test takers, along with semi-structured interviews conducted with relevant stakeholders, provide insights into five aspects of accessibility issues. First, consistent with previous studies (Choi, 2016; Fan & Ji, 2014; Lu et al., 2023; Moghadam & Nasirzadeh, 2020; Yao, 2023), stakeholders in this study unanimously agreed that the test-related information was transparent and could be accessed through multiple channels. Second, the test takers were provided with sufficient opportunities to take the EPT. Specifically, test takers can choose to sit for the test during any academic year or test session they prefer from their sophomore year onward. These findings are consistent with those of previous studies which reported that test takers tend to perceive themselves as being fairly treated when provided with sufficient opportunities to take a test (Tofighi & Safa, 2023) and the flexibility to decide when to take it (Yao, 2023). Third, test takers had access to various on-campus learning resources before taking the EPT. These resources were considered to be effective in improving their general English proficiency. Moreover, in line with Moghadam and Nasirzadeh's (2020) findings, the test location of the EPT did not present a geographical barrier for the test takers. Lastly, first-time test takers of the EPT reported challenges in adapting to the delivery system of the computer-based written test. They advocated for the development and use of a mock delivery system to familiarize themselves with the interface of the actual delivery system.

Findings from the analysis of questionnaire and interview data shed light on consistency issues in testing practices across four areas: (1) test development, (2) test administration, (3) scoring, and (4) score interpretation. Regarding test development,

efforts have been made to ensure item quality. Specifically, all item writers receive professional training before writing items. And all items are rigorously reviewed by testing experts, pilot tested, statistically analyzed, and revised before being included in the item bank. These procedures are consistent with Wainer's (1989) recommendation that test items should be reviewed and statistically evaluated after they are written.

Regarding consistency in administration, the questionnaire results showed that test takers generally perceived the administration of the EPT to be consistent. In the interviews, stakeholders shared their views on three aspects of test administration: (1) equipment functionality, (2) test environment, and (3) test security measures. First, no equipment malfunctions were reported by the test takers during the test. Second, concerning the test environment, test takers perceived the test rooms to be standardized in the layout and expressed satisfaction with the seating arrangements. However, several test takers complained about the noise generated by test equipment, a concern also noted by TOEFL test takers in Jang's (2002) study. These findings highlight that while standardized facilities and seating arrangements are valued, noise management remains an issue that requires test administrators' attention. Lastly, stakeholders expressed appreciation for the measures in place to ensure test security. Specifically, each test room was staffed with two proctors; test-takers' IDs were verified before entering test rooms; and monitoring cameras were in operation to oversee test discipline. These measures play an instrumental role in upholding test security (Moghadam & Nasirzadeh, 2020; Song, 2018).

In terms of rating consistency, stakeholders generally deemed the rating quality to be acceptable and consistent. This perceived acceptability and consistency could be attributed to: (1) the introduction of an AWE system in assisting human raters and (2) the measures implemented throughout the scoring procedures to ensure the reliability of rating outcomes. EPT stakeholders endorsed the integration of human and machine scoring methods as this integration could ensure consistency in the

rating process. As for rating procedures, both teachers and test users reported that all raters received formal training before scoring written and oral responses. During the interviews, they also provided detailed explanations of how scores from the AWE system and human raters were integrated to produce a final score. The discussion here underscores the importance of rating procedures in ensuring "acceptable scoring" (Dorans et al., 2022, p. 93).

With regard to consistency in score interpretation, the EPT was aligned with the *CSE*, a national English proficiency standard in China, as reported by a test user. To pass the EPT, test takers must achieve CSE-5. EPT-CSE alignment facilitates accurate and consistent score interpretation among stakeholder groups (Min et al., 2022). For test takers, the performance descriptors available in the overall scales, subscales, and self-assessment scales of the *CSE* provide a comprehensive understanding of their English proficiency levels. For those who fail to pass the test, the performance descriptors corresponding to CSE-5 serve as a benchmark for setting learning goals. For test developers, the performance descriptors in the *CSE* offer a valuable reference point for developing items with appropriate difficulty levels. Furthermore, for test users, the alignment between the EPT and the *CSE* allows for an evidence-based interpretation and use of test results.

The findings from the questionnaire survey and interviews also provide valuable insights into stakeholders' perceptions of accountability issues, including: (1) score review procedures, (2) stakeholder engagement, (3) fairness evaluation, and (4) the shared responsibilities among various stakeholders in ensuring the fairness of the EPT. First, test takers were given the opportunity to request score reviews after receiving their test results. However, specific procedures for score review application were unknown to nearly half of the test takers during the interviews. A review of the test syllabus and registration notices revealed no information about how to apply for score reviews. It is therefore understandable why test takers in the interviews expressed a need for greater transparency in score review procedures.

Second, despite the importance of stakeholder engagement, little effort has been directed toward collecting stakeholders' concerns, criticisms, compliments, or suggestions regarding the fairness of the EPT and the associated testing practices. Consistent with the findings of previous studies (Barrance & Elwood, 2018; Deygers, 2019), the test takers interviewed in this study reported feeling excluded from contributing their opinions both in the formulation of test policies and after taking the test. As emphasized by Young et al. (2013), feedback from a variety of stakeholders should be gathered before, during, and after the test is put into operational use. This greater stakeholder engagement is expected to improve stakeholders' perceived fairness of the test (Bøggild, 2016; De Cremer et al., 2008; Nisbet & Shaw, 2020), foster communication among stakeholders (Taylor, 2023), and improve assessment practices (Sonnleitner & Kovacs, 2020). The discussion underscores the need for greater attention to the stakeholders' perceptions throughout the lifecycle of the EPT.

Third, while stakeholders acknowledged the importance of fairness evaluation, they did not reach a consensus on who should be involved in fairness evaluation. Some argued that test takers could contribute to fairness evaluation as they had first-hand test-taking experience. Others argued that the researchers from the language testing research team at the university were qualified to conduct fairness evaluation. This perspective aligns with Standard 4.8 in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) which emphasizes the importance of using expert judges in the test review process.

Lastly, stakeholders acknowledged that ensuring the fairness of the EPT is a collective responsibility shared by all stakeholder groups involved in every stage of assessment. This finding aligns with Sabbaghan and Fazel's (2023) claim that "all stakeholders...must be involved and work in harmony towards more equitable and inclusive language assessment" (p. 182). In this study, test developers were identified as pivotal in maintaining the consistency of testing practices. Their responsibilities were reported to include formulating test specifications, developing administration

plans, and establishing test management mechanisms to uphold standardized testing practices. They were also reported to bear direct responsibility for test quality. It should be noted that the test developers in this study also have teaching responsibilities for the intended test takers of the EPT. This dual role as both test developers and College English teachers epitomizes the valuable contributions of local teaching staff to the development of local language tests (Dimova et al., 2020). Test users, as decision-makers, must be held accountable for the consequences of test use on test takers and society (Fan, 2014). In this study, test users are also experienced researchers in language testing and assessment, with expertise in developing and validating real-world language tests. Given their high level of language assessment literacy, it is unlikely that they would engage in unethical use of the EPT. Test takers also have a role to play in upholding test fairness. Consistent with Spaan's (2000) arguments, test takers are responsible for proactively seek test-related information prior to test-taking. They are also expected to fully demonstrate their English proficiency and adhere to test room regulations.

5.1.2.2 Factors influencing stakeholders' perceptions

Thematic analysis of the interview transcripts identified four factors influencing stakeholders' perceptions of the fairness of the EPT: sociocultural factors, educational factors, institutional factors, and personal factors.

The sociocultural factors identified are: (1) stakeholders' perceived importance of English proficiency and (2) societal norms around English testing practices in China. First, stakeholders' perceived importance of English shaped their value judgments on the fairness of the EPT. Specifically, stakeholders in this study considered English to be a lingua franca and a gateway to test-takers' future academic pursuits and professional prospects. This finding aligns with the well-documented instrumental value of English in China (e.g., Fang, 2018; Jin & Cortazzi, 2002). Given this perceived importance, it is not surprising that most stakeholders in this

study supported the development and use of the EPT. Second, some test takers were more inclined to endorse the use of EPT if peer universities also use in-house English proficiency tests to make graduation decisions. This finding illustrates the "relational" sense of fairness, reflecting test-takers' expectation that similar testing practices be implemented across peer universities. Furthermore, this finding suggests that alignment with peer universities in testing practices could enhance test-takers' perceived legitimacy of the EPT.

Educational factors have been identified in this study as influencing stakeholders' perceived fairness of the EPT. First, disparities in test-takers' pre-university English education have raised fairness concerns about the use of the EPT. Specifically, students who had had access to high-quality and comprehensive pre-university English education covering all skill domains—listening, speaking, reading, and writing—perceived themselves as having an advantage in English proficiency and found it easy to pass the EPT. In contrast, those who lacked access to such educational resources were at a relative disadvantage in terms of English proficiency upon entering university. This finding aligns with that of Fan et al. (2022) who reported that many low-achieving test takers of the FET came from provinces with lower performance levels in the NMET, such as Qinghai, Sichuan, and Guizhou provinces. Second, most stakeholders supported the use of the EPT as a graduation benchmark, largely because the university provided various learning resources to meet the undergraduates' individualized learning needs. Consistent with the findings from existing literature (Butler et al., 2021; Tofighi & Safa, 2023; Tsai & Tsou, 2009), stakeholders are more likely to endorse the use of language tests when there is an alignment between test purposes and teaching syllabi. Cole and Zieky (2001) cautioned against interpreting low test scores as evidence of an inability to learn if test takers do not have equal opportunities to learn the target construct. In the case of the EPT, the test construct closely aligned with the instructional goals of College English courses. Test takers have the flexibility to enroll in courses that match their

current language proficiency levels and learning needs. Beyond College English courses, additional learning resources are available to support test takers, including various EFL courses, remedial courses, one-on-one speaking and writing tutorials, and AI-powered English learning platforms, among others. These on-campus resources align with Standard 12.8 of the *Standards for Educational and Psychological Testing* (AERA et al., 2014), which states that "[w]hen test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test" (p. 197). Moreover, stakeholders' positive attitudes toward the EPT-as-exit-test policy suggest that locally developed language tests, as opposed to external ones, may better represent local contexts by taking into consideration local language teaching and learning situations (Dimova et al., 2020; Sireci & Randall, 2021; Tsai & Tsou, 2009).

In addition to sociocultural and educational factors, this study also identified a few institutional factors that influence the stakeholders' perceived fairness of the EPT. First, the university leadership provided policy support, without which, the development and administration of the EPT would not have been possible. As pointed out by Nguyen and Gu (2020), the formulation of policies concerning the use of an exit test requires consideration of a variety of issues, including test-takers' characteristics, practicality, and local context. The decision to develop, administer, and use the EPT is no exception. The EPT-as-exit-test policy was established over a decade ago following multiple rounds of discussions and deliberations among the university leadership. With this policy in place, various departments within the university were assigned specific responsibilities, such as test design and development, test administration, scoring, and delivery system design. Such inter-department collaboration would not have been possible without policy support. Second, stakeholders' perceived fairness of the EPT can be influenced by the expertise of test developers in language testing and assessment. The EPT was

developed by College English teachers, some of whom hold doctoral degrees in language testing and assessment and have professional expertise in test development and validation. Their extensive teaching experience, specialized knowledge in language testing and assessment, and data analysis skills were essential in supporting various testing practices, including test design, item development, scoring, and post-test equating. As emphasized by Dimova et al. (2020), local expertise is indispensable for the development of local language tests. Third, stakeholders perceived the EPT as fair due to the standardized setup of the test rooms, the proper functioning of test equipment, and relevant services provided by the university. The uniformity of test environment ensures that all test takers of the EPT have an equal opportunity to demonstrate their English proficiency during the test. In contrast, inconsistencies in administration conditions may "inadvertently influence the performance of some test takers relative to others" (AERA et al., 2014, p. 51). Therefore, for a computerized test such as the EPT, it is essential to minimize variability in administration conditions that may arise from faulty test equipment. Two institutional services—test registration notification and inter-campus transportation—contributed to the test-takers' perceived fairness of the EPT. The registration notification service keeps test takers informed of the registration procedures. The inter-campus transportation service demonstrates the university's commitment to accommodating the undergraduates' test-taking needs. Together, these two services facilitate test-takers' access to the test. However, some test takers reported experiencing anxiety and stress stemming from the EPT-as-exit-test policy. It is important for test users to consider the potential psychological impact associated with using the EPT as a graduation requirement. When institutional resources allow, post-test support services should be provided to protect test-takers' psychological well-being (Wang, 2016).

Thematic analysis of the interview transcripts also identified several personal factors influencing test-takers' perceptions of the fairness of the EPT. First, consistent with the claim of Iwashita and Elder (1997), test-takers' English proficiency levels

were found to influence their perceived difficulty level of the EPT and their acceptance of the passing standard. Test takers who met the proficiency requirements generally perceived the EPT's difficulty level as appropriate and were more likely to accept the EPT-as-exit-test policy. In contrast, repeat test takers whose English proficiency fell short the required threshold faced substantial challenges in passing the EPT. For these individuals, the EPT represented a barrier to graduation, as they perceived the test to be excessively difficult. This finding aligns with Tsai and Tsou's (2009) observation that test takers who self-identified as having satisfactory English proficiency exhibited more positive attitudes toward the use of high-stakes proficiency tests as graduation benchmarks. However, it is important to note that test-takers' perceptions of the EPT's difficulty level are inherently subjective and cannot serve as the sole basis for determining the EPT's actual level of difficulty. Second, test takers generally perceived the EPT as fair, largely because it is a criterion-referenced test. This finding aligns with the results of a questionnaire survey conducted by Wallace and Ng (2023), which revealed that test takers considered criterion-referenced tests to be the fairest when compared with individual-referenced and norm-referenced tests. Test-takers' perceived fairness of the EPT can be attributed to: (1) the absence of a predetermined pass rate, (2) the absence of peer competition among test takers, and (3) the positive washback of the EPT. Criterion-referenced tests assess test-takers' performance against predefined criteria or standards rather than comparing their performance with that of others. As such, the reasons cited by test takers for the perceived fairness of the EPT are closely associated with the features of criterion-referenced tests. In the context of the EPT, test takers were reported to engage in test preparation activities aimed at improving their English proficiency prior to taking the test. This finding is consistent with the observations of Fan and Ji (2014) and is further supported by Yao (2024), who found that test takers who perceive a test as fair are more likely to engage in test preparation.

5.1.3 Convergence and divergence of evaluation results from quantitative and qualitative inquiries

This study identified both convergence and divergence in the fairness evaluation results derived from the quantitative and qualitative inquiries. Quantitatively, test performance data and input materials were analyzed to address RQ1. Specifically, the test-takers' performance in the listening and reading subtests was analyzed using DIF, DBF, and DTF techniques to examine the comparability of test scores between the humanities and science test-taker groups. Additionally, the characteristics of the input materials in the listening and reading subtests were analyzed to examine their comparability across multiple test forms. Meanwhile, a questionnaire survey was administered to test takers to partially address RQ2. Qualitatively, one-on-one semi-structured interviews were conducted with a sample of test takers, teachers, test administrators, and test users to further address RQ2. The subsequent discussion is organized around the four dimensions of test fairness (i.e., comparability, accessibility, consistency, and accountability).

Comparability. Converging evidence was identified regarding test-takers' opportunities to demonstrate their English proficiency. Content analysis of DIF items and DBF testlets revealed no systematic item- or testlet-level bias that favored either the humanities or the science group. Similarly, test takers who participated in the questionnaire survey generally considered that they had comparable opportunities to demonstrate their proficiency during the test ($M = 4.32$, $SD = .96$). These quantitative findings were corroborated by the results of the interviews. Specifically, stakeholders reported that the test content featured a balanced representation of diverse topics that were familiar to test takers from various academic background. Taken together, these findings suggest that the EPT provides equal opportunities for test takers to demonstrate their English proficiency.

Accessibility. Findings from the questionnaire survey and interviews revealed unanimous agreement among test takers that test-related information was transparent

and accessible through various channels. Specifically, the analysis of the questionnaire data indicates test-takers' perceived transparency in test-related information, including test procedures, delivery modes, and rating criteria (M = 5.08, SD = .84). This finding was supported by the interview results, which highlighted test-takers' acknowledgment of the diverse channels available for accessing test-related information.

Consistency. Mixed findings were obtained from the questionnaire survey and interviews regarding the consistency of test administration. The questionnaire results indicate that test takers generally perceived the administration of the EPT to be consistent (M = 5.79, SD = .38). However, thematic analysis of the interview data revealed that test takers did not agree with one aspect of test administration—test environment (see Section 4.2.2.1 for details). Negative perceptions arise from test-takers' dissatisfaction with noise levels in test rooms. Nevertheless, most test takers in the interviews acknowledged and appreciated the efforts made to ensure the security of the EPT, a finding that aligns with the questionnaire results.

Accountability. Inconsistent findings emerged from the questionnaire survey and interviews regarding accountability issues. The questionnaire survey suggests that test takers can request score reviews and report instances of unfair treatment experienced during the test (M = 5.27, SD = .72). However, specific procedures for score reviews were unknown to nearly half of the test-taker interviewees. It should be pointed out that the interview results should be interpreted with caution, as they may not reflect the perceptions of the overall test-taker population.

5.2 Proposing a model of test fairness evaluation

Guided by the tentative conceptual model of test fairness evaluation (see Section 2.3), this study seeks to investigate whether the high-stakes EPT is fair from the perspectives of psychometrics (RQ1) and stakeholders (RQ2). To address RQ2, a questionnaire (see Section 3.4.1 for details), developed based on the tentative

conceptual model, was administered to 1,646 test takers. EFA and CFA were performed on the questionnaire data. Meanwhile, one-on-one semi-structured interviews were conducted with 20 test takers, six teachers, two test administrators, and three test users. A thematic analysis was performed to examine the stakeholders' perceived fairness of the EPT and the underlying reasoning behind their perceptions. The findings of this study: (1) serve as empirical evidence supporting the four dimensions of test fairness outlined in the tentative conceptual model and (2) shed light on the factors influencing the stakeholders' perceptions of test fairness.

Regarding the dimensions of test fairness, results from the EFA and CFA demonstrated a congruence between the empirically derived factors and the initially proposed dimensions outlined in the conceptual model (see Table 4.11 in Section 4.2.1.1). Each of the tentative dimensions—comparability, consistency, and accountability—corresponded to a standalone factor in the questionnaire. The thematic analysis identified four themes and 15 sub-themes that aligned well with the four dimensions of test fairness (see Section 4.2.2.1).

Based on the tentative conceptual model and the aforementioned findings, this study proposes a model of test fairness evaluation (see Figure 5.1). Compared with the tentative conceptual model, the newly proposed model incorporates an additional layer of the identified factors that influence stakeholders' perceptions of test fairness—sociocultural, educational, institutional, and personal factors. This layer reflects the contextually-situated nature of test fairness. Stakeholders' perceptions and interpretations of test fairness are not universal or fixed; instead, they are shaped by and embedded within specific contexts. What is considered "fair" in one context might be viewed differently in another due to these varying contextualized factors.

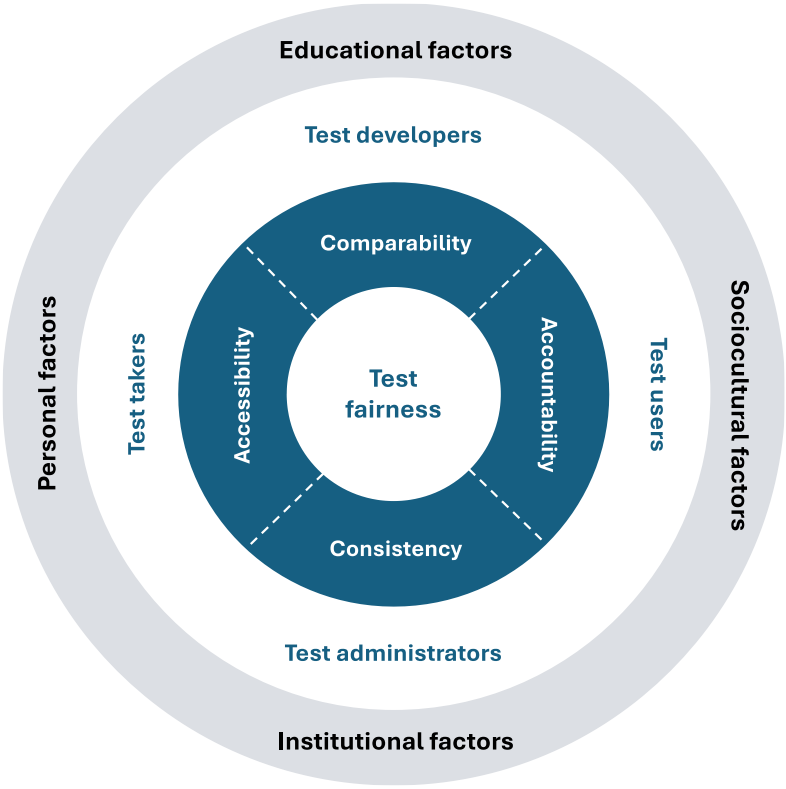

Figure 5.1 A model of test fairness evaluation.

This model of test fairness evaluation contributes to existing frameworks of fairness evaluation in the field of language testing. While previous frameworks (see Section 2.2.1 for a review) have been invaluable in unveiling a broad spectrum of fairness concerns (e.g., the psychometric qualities of language tests, consistency in test administration, and ethical issues inherent in test use), they tend to have limited applicability for evaluating the measurement and value judgement attributes of test fairness. In comparison with the first type of frameworks, which focuses predominantly on the psychometric qualities of language tests and address fairness concerns within validation frameworks (e.g., Xi, 2010), the model proposed in this study does justice to fairness by establishing it as the central focus of a dedicated evaluation model. While the second type of frameworks positions testing professionals as central to articulating fairness claims and navigating fairness evaluation (e.g., Kunnan, 2004, 2018), the proposed model acknowledges the socially-constructed nature of test fairness and underscores the importance of engaging stakeholders in fairness evaluation. The third type of frameworks (Huggins-

Manley et al., 2022), though promising in promoting greater stakeholder engagement in fairness evaluation, cannot be effectively applied in a specific context, particularly when the question of *what constitutes test fairness in a specific context* remains unresolved. Given the contextually-situated nature of test fairness (Jang, 2002; Nisbet & Shaw, 2020), an endless list of potential arguments and claims regarding the fairness of a local language test may arise if the values and priorities of test fairness held by stakeholders in a specific context are not identified. The fairness dimensions in the model (see Figure 5.1) were empirically validated based on feedback from various stakeholder groups, including test takers, teachers, test administrators, and test users. Therefore, the model is expected to be a bedrock for the proposal of fairness arguments and claims tailored specifically to the EPT.

The model proposed in this study carries several implications. First, it enhances our understanding of the diverse perspectives on test fairness held by various stakeholder groups in a local context in China. Second, the model offers valuable insights into constructing context-specific fairness arguments, which can serve as a basis for systematic fairness evaluation. Third, by outlining the dimensions of test fairness and the factors influencing the stakeholders' perceived fairness of the EPT, the model contributes to the fair design, development, administration, and use of language tests. Fourth, the model underscores that ensuring test fairness is a collective responsibility shared by all stakeholder groups involved in the entire lifecycle of local language tests. Lastly, the model reflects the multifaceted (Camilli & Newton, 2022; Opesemowo et al., 2023), socially-constructed (Camilli, 2006; Huggins-Manley et al., 2022; Jonson & Geisinger, 2022; Moss et al., 2005; Stobart, 2005; Tierney, 2014), and contextually-situated (Jang, 2002; Nisbet & Shaw, 2020) nature of test fairness. It is expected to guide the fairness evaluation of: (1) national- or international-level language tests, (2) language tests used in other cultural contexts, and (3) language tests used for admission, placement, promotion, and naturalization purposes. It should be noted that each dimension of test fairness outlined in the model may carry different

weight for different stakeholder groups and across different contexts. What is prioritized in the context of the EPT might differ from what stakeholders consider important in other test contexts.

5.3 Chapter summary

Chapter 5 provides a discussion of the evaluation results of the EPT's fairness from both psychometric and stakeholders' perspectives (Section 5.1). From the psychometric perspective, Section 5.1.1.1 discusses the findings regarding the comparability of test scores across humanities and science test-taker groups. First, potential sources of the identified DIF, DBF, and DTF are explored through a detailed analysis of the input materials' content, the subskills assessed by each DIF item or DBF testlet, and the characteristics of the listening and reading input materials. Second, the section examines whether psychometrically problematic items, testlets, and subtests exhibit academic background biases that favor or disfavor either the humanities or the science group. Third, the section also examines the relationships among DIF, DBF, and DTF in both the listening and reading subtests of the EPT. Section 5.1.1.2 discusses the comparability of the characteristics of the input materials across multiple test forms for the listening and reading subtests. The section also presents explanations for the observed incomparability of input characteristics across test forms and discusses the implications for test development.

Section 5.1.2 discusses the findings related to stakeholders' perceived fairness of the EPT and the factors influencing their perceptions. The findings are discussed in relation to those of previous studies, with consistencies or inconsistencies explained through an analysis of the contextual features of the EPT, the characteristics of the test takers, and relevant think pieces on test fairness.

Section 5.1.3 synthesizes the fairness evaluation results from both psychometric and stakeholders' perspectives. Explanations are provided for the convergence and

divergence of these results, highlighting the complementary nature of the two lines of inquiry.

Building on the empirical findings from a questionnaire survey and semi-structured interviews, Section 5.2 proposes a model of test fairness evaluation. In addition to the components initially outlined and empirically validated in the tentative conceptual model, this model incorporates an additional layer representing the factors influencing stakeholders' perceptions of test fairness. The section ends with a discussion of the implications of this model.

Chapter 6 Conclusion

This concluding chapter begins with a summary of key findings that address the research questions. Following this, Section 6.2 presents the theoretical, practical, and methodological implications of this study. The chapter then discusses the limitations of the current study and outlines potential avenues for future research. The chapter ends with concluding remarks that highlight the insights gained from the findings of this study.

6.1 Summary of key findings

Informed by a tentative conceptual model of test fairness evaluation (see Section 2.3), an attempt was made in this study to evaluate the fairness of the EPT from the perspectives of psychometrics and stakeholders, using both quantitative and qualitative approaches.

Overall, from the psychometric and stakeholders' perspectives, quantitative and qualitative results indicate that the EPT is generally fair. The results also provide empirical support for the tentative conceptual model of test fairness evaluation. The quantitative results are summarized as follows:

- *Comparability of the test results across humanities and science test-taker groups.* Among the four listening and reading test forms, 5% of the listening items showed DIF, 40% of the listening testlets and 22.22% of the reading testlets showed DBF, and two test forms showed DTF. Nevertheless, content analysis indicates that these statistically identified DIF, DBF, and DTF did not result in bias in the listening and reading subtests. Thus, the listening and reading test scores were comparable across the two groups.
- *Comparability of the input materials across different test forms.* Across the four test forms, the listening input materials exhibited comparable levels of lexical complexity, syntactic complexity, discourse complexity, and speed of

delivery. The reading input materials, while showing comparability in lexical complexity, varied slightly in characteristics such as syntactic complexity, discourse complexity, and readability across the four test forms.

- *Questionnaire survey*. The factor structure identified by the EFA and validated by the CFA aligned well with the dimensions of test fairness presented in the tentative conceptual model of test fairness evaluation. Test takers agreed that: (1) information about the test was transparent and can be accessed through different channels, (2) the EPT was administered in a consistent manner, (3) they were given sufficient opportunities to demonstrate their English proficiency during the test, and (4) they could request a score review or raise concerns regarding unfair practices after tanking the EPT.

The qualitative results are summarized as follows:

- *Stakeholders' perceptions of the fairness of the EPT*. Thematic analysis identified four themes that closely aligned with the four dimensions of test fairness outlined in the tentative conceptual model. Interview results showed that the four stakeholder groups generally accepted the use of the EPT and perceived it to be fair overall.
- *Factors influencing the stakeholders' perceived fairness of the EPT*. Four themes emerged from the interview, representing sociocultural, educational, institutional, and personal factors that influence the stakeholders' perceptions of the fairness of the EPT. Sociocultural factors include the importance of English proficiency in China, and societal norms around English testing in China. Educational factors include disparities in pre-university English education, and adequacy and effectiveness of university learning resources. Institutional factors include institutional policy, local expertise in language testing and assessment, and infrastructure for administering the EPT. Personal factors include test-takers' English proficiency, and test-takers' personal beliefs about the EPT.

Building on the tentative conceptual model and the empirical findings, a model of test fairness evaluation is proposed (see Section 5.2 for details). This model adds an additional layer to the tentative conceptual model, addressing the sociocultural, educational, institutional, and personal factors that were found to influence stakeholders' perceptions of the fairness of the EPT.

6.2 Implications

6.2.1 Theoretical implications

A three-layered model of test fairness evaluation was proposed in this study. The model is structured in concentric circles in response to the multifaceted, socially-constructed, and contextually-situated nature of test fairness. The inner layer encompasses four dimensions—comparability, accessibility, consistency, and accountability—that serve as key focuses and essential criteria for test fairness evaluation. Surrounding these dimensions is middle layer comprising four key stakeholder groups: test takers, test developers, test administrators, and test users. The outer layer encompasses the factors that influence stakeholders' perception of test fairness.

The model of test fairness evaluation holds important implications. First, the model captures the objective and subjective aspects of test fairness. From the perspective of psychometrics, the objective aspects of test fairness are measurable and can be evaluated through statistical procedures. Key concerns of the objective aspects of test fairness include the comparability of test results across test-taker subgroups and across different test forms. To evaluate the comparability of test results across test-taker subgroups, DIF, DBF, and DTF procedures can be used to detect potential bias in language tests. To ensure the comparability of test scores across different test forms, an anchor-item design can be employed in the assembly of test forms to facilitate post-test equating procedures. From the perspective of stakeholders, test fairness, as a value-laden concept, goes beyond a psychometric concern. What is

statistically fair may not be perceived as fair. Therefore, the evaluation of test fairness should include stakeholders' voices and perceptions.

Second, the model provides insights for developing context-specific fairness arguments that can guide fairness evaluation. As mentioned in Section 2.2.1, several studies have examined the fairness of language tests using an argument-based approach. As is often the case, the researchers themselves, who are usually also language testers, articulate claims that require substantiation. They subsequently gather empirical evidence to support, challenge, or even refute those claims. The argument-based approach presents three challenges. For one thing, researchers might need to "imagine the different challenges" associated with a claim, which complicates the identification of various components (including warrants, backing, qualifiers, and rebuttals) in a fairness argument. For another, claims made by testing professionals might be inaccessible to other stakeholder groups. Previous argument-based frameworks have overlooked the possibility of co-constructing fairness arguments with stakeholders outside the testing community. The model of test fairness evaluation, together with empirical evidence about the fairness of the EPT, can be used to develop fairness arguments specifically tailored to the EPT. For example, claims about the four dimensions of test fairness can be articulated with references to different stakeholders' accounts on and expectations of the fairness of the EPT. The quantitative and qualitative evidence collected in this study can inform the identification of warrants, backing, and rebuttals—essential components in Toulmin's argumentation model (Toulmin, 1958/2003). Once the fairness arguments are in place, subsequent empirical investigations into the fairness of the EPT can further substantiate each claim.

Third, by positioning "test fairness" in the center of the concentric circles (see Figure 5.1), the model underscores that ensuring test fairness is a collective responsibility shared among all stakeholder groups involved in every stage of the lifecycle of local language tests. Specifically, test developers should anticipate

fairness concerns likely to be raised by other stakeholder groups. Responsible test developers are those who prevent test bias throughout the test development cycle, publish annual reports on the psychometric properties of test scores, and keep other stakeholders informed about any measures undertaken to maintain the quality of test scores. Accountable test users should ensure test-takers' access to sufficient learning opportunities, examine potential consequences of score-based decisions, and provide other stakeholders with opportunities to voice concerns or objections regarding these decisions. Test administrators should strive to maintain consistency in administration procedures and environments. They should also ensure that every test taker is treated with respect and is able to perform at their full potential throughout test administration. Test takers are expected to be proactively engaged in three phases. Before taking a test, they should access test-related information and seek learning opportunities. During the test-taking process, they are expected to demonstrate their true language ability and adhere to the rules and regulations set forth by the testing organization. Following the test administration and decision-making process, test-takers should communicate any fairness issues encountered with test providers.

6.2.2 Practical implications

The present study holds practical implications too. First, it is feasible to evaluate the fairness of a high-stakes in-house English proficiency test. Test-takers' item-level performance data are relatively accessible for such an in-house test and can be used to evaluate test fairness from a psychometric perspective. In addition, different stakeholder groups are accessible in a local context, an ideal hub of understanding their "felt fairness" of a language test.

Second, evaluating test fairness in a local context holds several promises for real-world testing practices. For one thing, the findings from this study can be used to inform adjustments to the EPT or its associated testing practices. In the context of a local language test, research implications gained from test fairness investigations

can be translated into practical guidelines for local testing professionals to address fairness concerns. For another, by hearing voices from different stakeholder groups, this study holds broader implications for the testing practices of both national and international language tests. The expectations, concerns, and value judgments of stakeholders regarding the fairness of the EPT can, to some extent, represent those associated with nationwide language tests. Research of this kind thereby provides insights into the development, administration, and use of national-level language tests. Meanwhile, China hosts many international language tests, such as the TOEFL® iBT, the IELTS, the Test Deutsch als Fremdsprache (TestDaF), and the Japanese-Language Proficiency Test (JLPT). Findings from this study can help international test providers understand the perceptions of test fairness held by stakeholders in China, along with the sociocultural values underlying their perceptions. This study therefore contributes to the development, administration, and use of culturally-responsive and ethically-accountable international language tests.

Findings of this study also serve as a reference for the drafting of local guidelines for developing and using fair tests. While professional standards and guidelines exist to guide fair testing practices for educational and psychological tests (see AERA et al., 2014), as well as international commercialized language tests (see ETS, 2014, 2022), many of them remain unsubstantiated within the context of China. The criteria of a fair test established in US-centric professional documents may not be entirely applicable and acceptable in the context of local language tests in China and idiosyncrasies of local tests are often overlooked by these documents. To address these gaps and the lack of local standards and guidelines, this study has contributed to identifying context-specific qualities, criteria, and requirements for a fair test and fair testing practices in the context of the EPT. Drawing on the perceptions of multiple stakeholders regarding test fairness, as well as the factors influencing their perceptions, guidelines for local-level fairness review can be formulated in the future.

6.2.3 Methodological implications

This study has demonstrated that a mixed-methods approach can facilitate a relatively comprehensive and objective evaluation of the fairness of a test. As mentioned in Section 1.1, test fairness encompasses both objective and subjective aspects. Accordingly, it is important to recognize the dual nature of test fairness and distinguish between factual judgments and value judgements when evaluating the fairness of a language test.

In this study, factual judgments on the fairness of the EPT were made based on quantitative analyses of score comparability across test-taker groups and input material comparability across test forms. The results are expected to identify potential flaws in the design and development of the EPT. Specifically, evidence of score comparability, obtained through DIF, DBF, and DTF analyses, helps identify potential item-, testlet-, and test-level bias in the EPT. Moreover, analyzing the characteristics of the input materials helps uncover differences in the complexity of the materials used in different test forms of the EPT. Such factual evidence serves as indicators of the objective aspects of test fairness. As opposed to the stakeholders' perceptions of test fairness, evidence from statistical analyses of test performance and input materials tends to be relatively objective. After all, "felt fairness" of a test might be different from how fair the test actually is.

In addressing the subjective aspects of test fairness, stakeholders' perceptions of the fairness of the EPT were examined through a questionnaire survey administered to test takers and one-on-one interviews with representatives of key stakeholder groups, including test takers, teachers (also test developers), test administrators, and test users. As pointed out by Zieky (2006), it is impossible for test developers to prove that their tests are fair. Sen (2010) further argued that an individual's perspective can never be impartial due to the existence of objective illusions. That is why many scholars endorse a multi-stakeholder approach for fairness evaluation (e.g., Huggins-Manley et al., 2022; Nagel, 1989; Sen, 2009), which holds the promise of forming a

"synergistic" and "co-operative communication cycle" among key stakeholder groups. In contrast to the exclusive reliance on statistical techniques mentioned earlier, the multi-stakeholder approach enables a more comprehensive evaluation of test fairness. The engagement of stakeholder groups in test fairness evaluation could address fairness concerns: (1) at the level of individual test takers as opposed to test-taker groups; (2) among all key stakeholder groups, not just test takers; and (3) across a variety of key stages within the testing cycle rather than restricting the evaluation focus solely to test quality.

6.3 Limitations and suggestions for future research

This section discusses the limitations of this study that need to be addressed in future research. First, DIF, DBF, and DTF analyses were conducted using the performance data from four listening and reading test forms used in a single test administration. The dataset may not be fully representative of the performance of the whole test-taker population of the EPT. Replication of these analyses using data from multiple administrations would enhance the generalizability of the findings.

Second, while this study investigated DIF across two test-taker groups with different academic backgrounds, each academic group consists of subgroups defined by variables such as gender, grade, socioeconomic status, and sociocultural background. This within-group heterogeneity could reduce accurate DIF detection rates (Aryadoust et al., 2024; Grover & Ercikan, 2017; Oliveri et al., 2014). Thus, it would be worthwhile for future research to examine: (1) the effects of within-group heterogeneity on DIF detection accuracy, using both simulated and real test performance data; and (2) the extent to which the accuracy of DIF detection influences that of DBF and DTF detection.

Third, this study is cross-sectional in nature, as the fairness evaluation was conducted at a single point in time. Specifically, score comparability across test-taker groups was analyzed using performance data from a single test administration.

Additionally, only the input materials from the test forms used in that administration were used for comparability analysis. Furthermore, only test takers from that single administration were invited to participate in a questionnaire survey about the fairness of the EPT. Future research would benefit from adopting a longitudinal research design. A systematic evaluation of the fairness of the EPT can be achieved by analyzing performance data collected from multiple administrations, examining the input materials used over time, and engaging test takers in a questionnaire survey following each test administration. As test fairness is a perennial concern, it demands comprehensive and continuous evaluation on a long-term basis.

6.4 Concluding remarks

This mixed-methods study, guided by a tentative conceptual model of test fairness evaluation, has evaluated the fairness of an in-house English proficiency test from the perspectives of psychometrics and stakeholders. The findings, first and foremost, demonstrate the multifaceted nature of test fairness. Different stakeholders have varying understandings of test fairness and different perceptions of the fairness of the EPT. As Hamp-Lyons (2000) argued, there was no single perspective from which a test could be determined as 'fair' or 'unfair'. Test fairness is therefore best understood and approached through a lens of pluralism by incorporating diverse stakeholders' voices throughout the evaluation process. Test developers alone are unable to claim or prove that a test is fair. Furthermore, test fairness is not a black-and-white concept. The findings of this study indicate the difficulty in making definitive judgments about the fairness of a language test. Aligned with Nisbet and Shaw's (2020) viewpoint, test fairness exists along a continuum. From a pragmatic standpoint, tests can be made fairer by identifying and ruling out potential sources of unfairness in testing instruments and practices. Admittedly, the pursuit of absolute fairness is akin to the quest for the holy grail (Davies, 2010). On the optimistic side, Zieky (2006) noted that "because fairness and validity are so closely intertwined, the procedures required

to make a fair test overlap considerably with the procedures required to make a good test" (p. 359). Our goal is not to create a perfectly fair test but rather to develop a good or a fairer one that maintains public trust.

A fairer test can be realized through commitment to openness, communication, and collaboration. A prerequisite for test fairness is transparency. To be held accountable for a test, test developers and users should provide the public with sufficient information about the fairness of a test and test-related practices. This move can facilitate evidence-based public debates on fairness issues throughout the entire testing cycle, including test design, development, administration, scoring, and test use. Moreover, achieving a fairer test requires effective stakeholder communication. Stakeholder groups with greater authority (e.g., policymakers, test developers, and test users) should respect those with less power (e.g., test-takers) not by ignoring their voices, but by attending to them. While those in authority or with testing expertise may challenge the perspectives of the powerless stakeholders, the former can also gain valuable insights from the latter. Communication among stakeholders is expected to: (1) identify fairness concerns within both the testing community and public square; (2) construct fairness arguments; and (3) prioritize research efforts toward addressing the most critical fairness issues identified in a particular context. Lastly, all stakeholder groups, whether powerful or powerless, play a role in achieving test fairness. They must collaborate to strive toward a fairer test.

References

Abbott, M. L. (2006). ESL reading strategies: Differences in Arabic and Mandarin speaker test performance. *Language Learning*, *56*(4), 633–670. https://doi.org/10.1111/j.1467-9922.2006.00391.x

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*(1), 7–36. https://doi.org/10.1177/0265532207071510

Ahmadi, A., & Bazvand, A. D. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research*, *4*(1), 63–82.

Alavi, S. M., Rezaee, A. A., & Amirian, S. M. (2011). Academic discipline DIF in an English language proficiency test. *Journal of English Language Teaching and Learning*, *7*, 39–65.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Amirian, S. M. R. (2020). Investigating fairness of reading comprehension section of INUEE: Learner's attitudes towards DIF sources. *International Journal of Language Testing*, *10*(2), 88–100.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.). *Differential item*

functioning (pp. 3–23). Routledge. https://doi.org/10.4324/978020335-7811

Aryadoust, V. (2015). Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing*, *15*(3), 216–238. https://doi.org/10.1080/15305058.2015.1004409

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. *Studies in Educational Evaluation*, *56*, 197–204. https://doi.org/10.1016/j.stueduc.-2018.01.003

Aryadoust, V., Min, S., & Chen, X. (2024). Investigating differential item functioning across interaction variables in listening comprehension assessment. *Studies in Educational Evaluation*, *80*, Article 101322. https://doi.org/10.1016/j.stueduc.2024.101322

Babyak, M. A., & Green, S. B. (2010). Confirmatory factor analysis: An introduction for psychosomatic medicine researchers. *Psychosomatic Medicine*, *72*(6), 587–597. https://doi.org/10.1097/PSY.0b013e3181-de3f8a

Bachman, L., & Palmer, A. (2010/2016). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Foreign Language Teaching and Research Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barrance, R., & Elwood, J. (2018). Young people's views on choice and fairness through their experiences of curriculum as examination specifications at GCSE. *Oxford Review of Education*, *44*(1), 19–36. https://doi.org/10.-1080/03054985.2018.1409964

Bazvand, A. D., & Rasooli, A. (2022). Students' experiences of fairness in summative assessment: A study in a higher education context. *Studies in Educational Evaluation*, *72*, Article 101118. https://doi.org/10.1016/j.stueduc.2021.101118

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of large-scale educational assessments: Issues and recommendations*. National Academy of Education. https://doi.org/10.31094/2020/1

Bestgen, Y. (2024). Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, *74*(3), 638–671. https://doi.org/10.1111/lang.12630

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). *What makes listening difficult? Factors affecting second language listening comprehension* (Technical Report TTO 81434 E.3.1). University of Maryland Center for Advanced Study of Language. https://doi.org/10.21236/ADA550176

Bøggild, T. (2016). *Procedural fairness and public opinion*. Politica.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bowles, M. A. (2022). Using instructor judgment, learner corpora, and DIF to develop a placement test for Spanish L2 and heritage learners. *Language Testing*, *39*(3), 355–376. https://doi.org/10.1177/02655322221076033

Boyd, K., & Davies, A. (2002). Doctors' orders for language testers: the origin and purpose of ethical codes. *Language Testing*, *19*(3), 296–322. https://doi.org/10.1191/0265532202lt231oa

Brati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high stakes tests: The effect of field of study. *Iranian Journal of Applied Linguistics*, *9*(2), 27–49.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. Sage Publications.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369–394. https://doi.org/10.1191/0265532202lt236oa

Brown, J. D. (2008). Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly*, *5*(4), 275–312. https://doi.org/10.1080/15434300802457455

Brunfaut, T. (2016). Assessing listening. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 97–112). Mouton de Gruyter. https://doi.org/10.1515/9781614513827-009

Brunfaut, T. (2021). Assessing reading. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 254–267), Routledge.

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, *49*(1), 141–168. https://doi.org/10.1002/tesq.168

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done?. *Qualitative Research*, *6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Burchett, W. W., Ellis, A. R., Harrar, S. W., & Bathke, A. C. (2017). Nonparametric inference for multivariate data: The R package npmv.

Journal of Statistical Software, *76*(4), 1–18. https://doi.org/10.18637/-jss.v076.i04

Burstein. J. (2023). *Responsible AI standards*. Duolingo.

Butler, Y. G., Peng, X., & Lee, J. (2021). Young learners' voices: Towards a learner-centered approach to understanding language assessment literacy. *Language Testing*, *38*(3), 429–455. https://doi.org/10.1177/02655322219-92274

Caines, J., Bridglall, B. L., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education*, *22*(1), 5–18. https://doi.org/10.1108/QAE-12-2013-0054

Cambridge University Press. (n.d.). Fair. In *Cambridge English Dictionary*. Retrieved August 10, 2023, from https://dictionary.cambridge.org/-dictionary/english/fair

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Praeger.

Camilli, G. (2016). Ongoing issues in test fairness. In H. Karami (Ed.), *Fairness issues in educational assessment* (pp. 16–32). Routledge.

Camilli, G., & Newton, P. E. (2022). Doing justice to fairness. In J. L. Jonson & L. F. Geisinger (Eds.), *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 Standards* (pp. 111–130). American Educational Research Association.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications.

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2023). *Duolingo English Test: Technical manual* [Duolingo research report, May 1, 2023].

Duolingo English Test. http://duolingo-papers.s3.amazonaws.com/other/-technical_manual.pdf

Carlsen, C. H., & Rocca, L. (2022). Language test activism. *Language Policy*, *21*(1), 597–616. https://doi.org/10.1007/s10993-022-09614-7

Carrell, P. L., & Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *TESOL Quarterly*, *17*(4), 553–573. https://doi.org/10.2307/-3586613

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate *Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906-mbr0102_10

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*™. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203937891

Chen, J., & Zeng, Y. (2021). A confirmatory approach to item bias across different academic backgrounds. *Modern Foreign Languages*, *44*(6), 815–826. [陈锦, 曾用强. (2021). 验证性分析框架下学科背景偏差研究. 《现代外语》, *44*(6): 815–826.]

Chen, M. (2013). *DIF detection for examinees with different academic backgrounds and in different economic regions in a standard English listening subtest* [Unpublished master's dissertation]. Guangdong University of Foreign Studies. [陈敏. (2013). 《有关考生不同学术背景和所属经济区域对考生听力成绩的影响——一份标准英语听力试题的项目功能差异研究》. 广东外语外贸大学.]

Cheng, L., & Sultana, N. (2022). Washback: Looking backward and forward. In G. Fulcher & L. Harding (Eds.). *The Routledge handbook of language testing* (pp. 136–152). Routledge.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/s15328-007sem0902_5

China Disabled Persons' Federation, Ministry of Education of the People's Republic of China. (2017, April 7). Notice on Issuing the *Regulations on the Administration of the Nationwide Unified Examination for Admissions to General Universities and Colleges for Disabled Persons*. https://www.-gov.cn/zhengce/zhengceku/2017-04/12/content_5650060.htm

Choi, I. (2016). Test-takers' perceptions of test fairness: A washback study on test information given prior to a high-stakes writing test. *Foreign Language Education Research*, *19*, 1–18.

Christensen, L. L., Shyyan, V. V., & MacMillan, F. (2023). Toward a systematic accessibility review process for English language proficiency tests for young learners. *Language Testing*, *40*(4), 856–876. https://doi.org/10.1177/02655322231168386

Cobb, T. (n.d.). VocabProfiler (Web VP Classic v.4 version). Retrieved from https://www.lextutor.ca/vp/eng/

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355–386). American Council on Education/Praeger.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). American Council on Education & National Council on Measurement in Education.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, *38*(4), 369–382. https://doi.org/10.1111/j.174-5-3984.2001.tb01132.x

Collins English Dictionary. (n.d.). Fair. In *Collins English Dictionary*. Retrieved August 25, 2023, from https://www.collinsdictionary.com/dictionary/english/fair

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage Publications.

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). Sage Publications.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing*, *14*(3), 328–339. https://doi.org/10.1177/02655322970-140030

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176. https://doi.org/10.1177/0265532209349466

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://journals.sagepub.com/doi/10.1177/0265532215582282

De Cremer, D., Cornelis, I., & Van Hiel, A. (2008). To whom does voice in groups matter? Effects of voice on affect and procedural fairness judgments as a function of social dominance orientation. *The Journal of Social Psychology*, *148*(1), 61–76. https://doi.org/10.3200/SOCP.148.1.-61-76

DeVellis, R. F., & Thorpe, C. T. (2022). *Scale development: Theory and applications* (5th ed.). Sage Publications.

Deygers, B. (2017). Just testing: Applying theories of justice to high-stakes language tests. *International Journal of Applied Linguistics*, *168*(2), 143–162. https://doi.org/10.1075/itl.00001.dey

Deygers, B. (2019). Fairness and social justice in English language assessment. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 541–569). Springer. https://doi.org/10.1007/978-3-030-02899-2_30

Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. Routledge. https://doi.org/10.4324/97-80429492242

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, *31*(4), 20–37. https://doi.org/10.1111/j.1745-3992.2012.00250.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates.

Dorans, N. J., Syp, M. J., & Walker, M. E. (2022). Aligning fairness assessment with ensuring fair contexts for motivated test takers. In. J. Jonson & K. F. Geisinger (Eds.), *Fairness in testing in the testing standards* (pp. 89–110). American Educational Research Association.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465–484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Drackert, A., & Timukova, A. (2020). What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing*, *37*(1), 107–132. https://doi.org/10.1177/0265532219861042

Duolingo. (2021). *Analysis of the fairness of the Duolingo English test*. Duolingo.

Educational Testing Service. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments*. Educational Testing Service.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Educational Testing Service.

Educational Testing Service. (2016). *ETS international principles for the fairness of assessments: A manual for developing locally appropriate fairness guidelines for various countries*. Educational Testing Service.

Educational Testing Service. (2022). *ETS guidelines for developing fair tests and communications*. Educational Testing Service.

Elosua, P. (2024). A three-step DIF analysis of a reading comprehension test across regional dialects to improve test score validity. *Language Assessment Quarterly, 21*(2), 141–158. https://doi.org/10.1080/15434303.2024.2307621

Ercikan, K. (2006). Developments in assessment of student learning and achievement. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 929–953). Lawrence Erlbaum Associates.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education* (pp. 545–569). American Psychological Association.

Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). National Academy of Education. https://doi.org/10.31094/2020/1

Eslami, H. (2014). The effect of syntactic simplicity and complexity on the readability of the text. *Journal of Language Teaching and Research*, *5*(5), 1185–1191. https://doi.org/10.4304/jltr.5.5.1185-1191

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fan, J. (2014). Research on the fairness of language tests: Conceptualizations, theories, and responsibilities. *Foreign Language Testing and Teaching*, (2), 11–19. [范劲松. (2014). 语言测试的公平性研究: 概念、理论与责任. 《外语测试与教学》, (2): 11–19.]

Fan, J. (2018). A survey of English language testing practices in China: Students' and teachers' perspectives. In D. Xerri & P. V. Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 283–300). Springer.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL*, *9*, 1–35.

Fan, J., Frost, K., & Jin, Y. (2022). Local English testing in China's tertiary education: Contexts, policies, and practices. *Language Testing*, *39*(3), 453–473. https://doi.org/10.1177/02655322211070839

Fang, F. (2018). Ideology and identity debate of English in China: Past, present and future. *Asian Englishes*, *20*(1), 15–26. https://doi.org/10.1080/13488-678.2017.1415516

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, *58*(3), 840–852. https://doi.org/10.10-44/2015_JSLHR-L-14-0280

Field, A. (2009). *Discovering statistics using SPSS*. Sage Publications.

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478–488.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/-10.1093/applin/21.3.354

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education, 14*(1), 9–26. http://dx.doi.org/10.10-80/09695940701272773

Geisinger, K. F. (2015). Test evaluation. In S.Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 624–638). Routledge.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the Certificate in Advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222. https://doi.org/10.1080/15434300-701375758

Ghaemi, H., & Khorami, M. (2024). Item response theory and Mantel-Haenszel procedures in detecting academic discipline differential item functioning and differential skill functioning with English proficiency test. *Applications of Language Studies*, *2*(2), 20–58. https://doi.org/10.22034/-jals.2024.2022383.1009

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 105–118). Springer.

Good, P. I. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (3rd ed.). Springer Science & Business Media.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3-758/BF03195564

Green, R. (2017). *Designing listening tests: A practical approach*. Palgrave Macmillan.

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly, 26*(2), 385–390. https://doi.org/10.-2307/3587015

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education*, *30*(3), 178–195. https://doi.org/10.1080/08957347.2017.1316276

Gui, S. (2000). *Psycholinguistics (New Edition)*. Shanghai Foreign Language Education Press. [桂诗春. (2000). 《新编心理语言学》. 上海外语教育出版社.]

Gujord, A. K. H. (2023). Who succeeds and who fails? Exploring the role of background variables in explaining the outcomes of L2 language tests. *Language Testing*, *40*(2), 227–248. https://doi.org/10.1177/0265532222-1100115

Guzman-Orth, D., Steinberg, J., & Albee, T. (2023). English learners who are blind or visually impaired: A participatory design approach to enhancing fairness and validity for language testing accommodations. *Language Testing*, *40*(4), 933–959. https://doi.org/10.1177/02655322231159143

Hair, J., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate data analysis* (4th ed.). Prentice Hall.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language Testing in Asia*, *9*(16), 1–20. https://doi.org/10.1186/s40468-019-0092-9

Hamp-Lyons, L. (2000). Social, professional, and individual responsibility in language testing. *System*, *28*(4), 579–98. https://doi.org/10.1016/S0346-251X(00)00039-7.

Harper, D. (n.d.). Fair. In *Online Etymology Dictionary*. Retrieved August 25, 2023, from https://www.etymonline.com/word/fair

Hawkey, R. (2008). An impact study of a high-stakes test (IELTS): Lessons for test validation and linguistic diversity. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 215–228). UCLES/Cambridge University Press.

He, L., Chen, D., & Min, S. (2018). Investigating the effects of test method on task difficulty in English listening comprehension test. *Modern Foreign Languages*, *41*(1), 43–54. [何莲珍, 陈大建, 闵尚超. (2018). 英语听力测试中测试方法对任务难度的影响研究. 《现代外语》, *41*(1): 43–54.]

He, M. (2015). An investigation into fairness of college achievement tests—Survey from students and teachers. *Journal of Hubei Normal University (Philosophy and Social Science)*, *35*(3), 99–101. [贺满足. (2015). 大学学业考试的公平性及影响研究——来自学生的反馈. 《湖北师范学院学报(哲学社会科学版)》, *35*(3): 99–101.]

He, M. (2018). Exploring of the fairness of college English achievement tests: DIF analysis of reading tests across academic backgrounds. *Education and Examinations*, (5): 51–57. [贺满足. (2018). 大学英语成就测试的公平性探究——阅读测试的专业背景 DIF 检验. 《教育与考试》, (5): 51–57.]

He, N. (2022). *A study on the bias of listening comprehension test of CET-6: A differential item functioning approach* [Unpublished master's dissertation]. Sichuan International Studies University. [何念念. (2022). 《大学英语六级听力测试偏差研究》. 四川外国语大学.]

Heeren, J., Speelman, D., & De Wachter, L. (2021). Post-entry language assessment in higher education: Ensuring fairness by including the voice of the test-taker. *Collated Papers for the ALTE 7th International Conference*, 45–48.

Henning, G. (1990). National issues in individual assessment: The consideration of specialization bias in university language screening tests. In J. H. A. L. de Jong & D. K. Stevenson (Eds.)., *Individualizing the assessment of language abilities* (pp. 38–50). Multilingual Matters.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. https://doi.org/10.1007/BF022894-47

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.). *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage Publications.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, D., & Garner, M. (2009). A case of test impact: Cheating on the College English Test in China. In L. Taylor & C. Weir (Eds.), *Language testing matters: Investigating the wider social and educational impact of assessment* (pp. 59–76). Cambridge University Press.

Huggins-Manley, A. C., Booth, B. M., & D'Mello, S. K. (2022). Toward argument-based fairness with an application to AI-enhanced educational assessments. *Journal of Educational Measurement, 59*(3), 362–388. https://doi.org/10.1111/jedm.12334

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating. *Educational and Psychological Measurement*, *74*(4), 627–658. https://doi.org/10.1177/0013164413506222

IBM Corp. (2024). *IBM SPSS statistics for macOS* (Version 29.0) [Computer software]. IBM Corp.

Iwashita, N., & Elder, C. (1997). Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing*, *6*(1), 53–67.

Jafaripour, S., Tabatabaei, O., Salehi, H., & Dastjerdi, H. V. (2024). Applying IRT model to determine gender- and discipline-based DIF and DDF: A study of the IAU English proficiency test. *International Journal of*

Language Testing, 14(1), 56–74. https://doi.org/10.22034/IJLT.2023.407-117.1268

Jang, E. E. (2002). *In search of folk fairness in language testing* [Unpublished master's dissertation]. University of Illinois at Urbana-Champaign.

Jin, L., & Cortazzi, M. (2002). English language teaching in China: A bridge to the future. *Asia Pacific Journal of Education*, *22*(2), 53–64. https://doi.org/10.1080/0218879020220206

Jin, Y., & Wu, E. (2017). An argument-based approach to test fairness: The case of multiple-form equating in the College English Test. *International Journal of Computer-Assisted Language Learning and Teaching*, *7*(3), 58–72. https://doi.org/10.4018/IJCALLT.2017070104

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, *14*(2), 101–119. https://doi.org/10.1080/15434303.2016.126-1293

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*. Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Joint Committee on Testing Practices.

Jonson, J. L., & Geisinger, K. F. (2022). Looking forward: Cross-cutting themes for the future of fairness in testing. In J. L. Jonson & K. F. Geisinger (Eds.), *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 Standards* (pp. 399–416). American Educational Research Association. https://doi.org/-10.3102/9780935302967_17

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. http://dx.doi.org/10.1177/001316446002000116

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/00-1316447403400115

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. (2013). Validity and fairness in the testing of individuals. In M. Chatterji, (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 17–53). Emerald Group Publishing Limited.

Kang, O., Yan, X., Kostromitina, M., Thomson, R., & Isaacs, T. (2024). Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English test. *Language Testing*, *41*(2), 263–289. https://doi.org/10.1177/02655322231179134

Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation*, 19(2–3), 158–169. https://doi.org/10.1080/13-803611.2013.767618

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests* (Version 0.7.2). https://CRAN.R-project.org/package=rstatiz

Kim, Y. H., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*(4), 825–865. https://doi.org/10.1111/j.1467-9922.2009.00527.x

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Report No. RR-79). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.-tb01938.x

Kunnan, A. J. (2000). *Fairness and justice for all*. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic, C. Weir, & S. Bolton (Eds.), *Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27–48). Cambridge University Press.

Kunnan, A. J. (2013). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1098–1114). Wiley Blackwell. https://doi.org/10.1002/9781118411360.wbcla-144

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge. https://doi.org/10.4324/9780203803554

Kunnan, A. J. (2020). A case for an ethics-based approach to evaluate language assessments. In G. J. Ockey & B. A. Green (Eds.). *Another generation of fundamental considerations in language assessment: A Festschrift in Honor of Lyle F. Bachman* (pp. 77–93). Springer. https://doi.org/10.100-7/978-981-15-8952-2_6

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Lakin, J. M., Elliott, D. C., & Liu, O. L. (2012). Investigating ESL students' performance on outcomes assessments in higher education. *Educational and Psychological Measurement*, *72*(5), 734–753. https://doi.org/10.117-7/0013164412442376

Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement*, *39*(3), 24–27. https://doi.org/-10.1111/emip.12359

Lee, Y.-S., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review*, *10*(3), 365–375. https://doi.org/10.1007/s12564-009-9039-7

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). Springer. https://doi.org/-10.1007/978-1-4613-3087-5_2

Li C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s-13428-015-0619-7

Li, D. (2021). Development and validation of a scale for evaluating the fairness of language tests. *Foreign Language World*, (1), 88–95. [李迪. (2021). 语言测试公平性检验量表研制与效度验证. 《外语界》, (1): 88–95.]

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677. https://doi.org/10.1007/BF02294041

Li, H., Hunter, C. V., & Bialo, J. A. (2022). A revisit of Zumbo's third generation DIF: How are we doing in language testing?. *Language Assessment Quarterly*, *19*(1), 27–53. https://doi.org/10.1080/1543430-3.2021.1963253

Liao, L. (2020). A comparability study of text difficulty and task characteristics of parallel academic IELTS reading tests. *English Language Teaching*, *13*(1), 31–42. https://doi.org/10.5539/elt.v13n1p31

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in general English proficiency test-kids listening. *Frontiers in Psychology*, *12*, Article 767244. https://doi.org/10.3389/fpsyg.2021.767244

Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL® iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education*, *24*(3), 235–255. https://doi.org/10.1080/08957347.2011.580645

Liu, Y., & Zheng, B. (2022). Comparability of difficulty levels of translation tasks in CET-6 parallel test forms: evidence from product and process-based data. *The Interpreter and Translator Trainer*, *16*(4), 428–447. https://doi.org/10.1080/1750399X.2022.2036938

Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A demonstration of the impact of outliers on the decisions about the number of factors in exploratory factor analysis. *Educational and Psychological Measurement*, *72*(2), 181–199. https://doi.org/10.1177/0013164411410878

Lu, W., Zeng, Y., & Yang, M. (2023). Examining the fairness of tests from the perspective of learning opportunities—A case of the VETS. *Foreign Language Testing and Teaching*, (4), 20–28. [卢伟烈, 曾用强, 杨敏迎. (2023). 从学习机会角度检验考试的公平性——以 VETS 考试为例. 《外语测试与教学》, (4): 20–28.]

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.201-5.06.003

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters?. *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Lv, Z. (2018). *Validating an in-house oral proficiency test: Prompt Effects* [Unpublished doctoral dissertation]. Zhejiang University.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, *2*, 49–55.

Malone M. E., Montee M. (2014). *Stakeholders' beliefs about the TOEFL iBT test as a measure of academic language ability*. ETS Research Report No. RR-4-42. Educational Testing Service. https://doi.org/10.1002/ets2.12039

Mardia, K. V. (1971). The effect of nonnormality onsome multivariate tests and robustness to non-normality in the linear model. *Biometrika*, *58*(1), 105–121. https://doi.org/10.1093/biomet/58.1.105

McArthur, J. (2018). *Assessment for social justice: Perspectives and practices within higher education*. Bloomsbury.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Wiley-Blackwell.

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.-2011.565438.

Merriam-Webster. (n.d.). Fairness. In *Merriam-Webster.com dictionary*. Retrieved August 25, 2023, from https://www.merriam-webster.com/dictionary/fairness

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). Macmillan Publishing Company & American Council on Education.

Min, H. (2011). Differential item functioning of an ESL placement exam for the examinees with different academic disciplines. *Korean Journal of Applied Linguistics*, *27*(4), 105–124.

Min, S., & He, L. (2020). Test fairness: Examining differential functioning of the reading comprehension section of the GSEEE in China. *Studies in*

Educational Evaluation, *64*, Article 100811. https://doi.org/10.1016/j.stueduc.2019.100811

Min, S., Zhang, J., Li, Y., & He, L. (2022). Bridging local needs and national standards: Use of standards-based individualized feedback of an in-house EFL listening test in China. *Language Testing*, *39*(3), 425–452. https://doi.org/10.1177/02655322211070990

Ministry of Education of the People's Republic of China, & National Language Commission of the People's Republic of China. (2018). *China's Standards of English Language Ability*. Higher Education Press & Shanghai Foreign Language Education Press.

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*, Article 7. https://doi.org/10.1186/s40468-020-0010-52

Mohamadi, Z. (2013). Determining the difficulty level of listening tasks. *Theory and Practice in Language Studies*, *3*(6), 987–994. https://doi.org/-10.4304/tpls.3.6.987-994

Moss, P., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement*, *3*(2), 63–83. https://doi.org/10.1207/s15366359mea0302_1

Motteram, J., Spiby, R., Bellhouse, G., & Sroka, K. (2023). Implementation of an accommodations policy for candidates with diverse needs in a large-scale testing system. *Language Testing*, *40*(4), 904–932. https://doi.org/-10.1177/02655322231166587

Nagel, T. (1989). *The view from nowhere*. Oxford University Press.

Neittaanmäki, R., & Lamprianou, I. (2024). Communal factors in rater severity and consistency over time in high-stakes oral assessment. *Language Testing*, *41*(3), 584–605. https://doi.org/10.1177/02655322241239363

Nguyen, H., & Gu, Y. (2020). Impact of TOEIC listening and reading as a university exit test in Vietnam. *Language Assessment Quarterly*, *17*(2), 147–167. https://doi.org/10.1080/15434303.2020.1722672

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.158-6643

Nisbet, I., & Shaw, S. (2020). *Is assessment fair?*. Sage Publications. https://doi.org/10.4135/9781529739480

Noble, T., Wells, C. S., & Rosebery, A. S. (2023). English learners and constructed-response science test items challenges and opportunities. *Educational Assessment*, *28*(4), 246–272. https://doi.org/10.1080/10627-197.2023.2226387

Noori, M. (2022). "The road not taken" in language testing: Sociocultural implications of test and teaching contents. *TESOL Quarterly, 56*(4), 1486–1503. https://doi.org/10.1002/tesq.3189

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, *27*(4), 286–300. https://doi.org/10.1080/08957347.2014.94-4305

Opesemowo, O. A., Ayanwale, M. A., Opesemowo, T. R., & Afolabi, E. R. I. (2023). Differential bundle functioning of National Examinations Council Mathematics Test items: An exploratory structural equation modelling approach. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 1–18. https://doi.org/10.21031/epod.1142713

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, *21*(1), 53–73. https://doi.org/10.1191/0265532204lt2-74oa

Pae, T.-I., & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*(4), 475–496. https://doi.org/10.1191/0265532206lt338oa

Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554. https://doi.org/10.1177/0265532211434027

Pan, Z. (2021). A study on the factors of test items affecting the difficulty of listening comprehension tasks. *Foreign Language Testing and Teaching*, (2), 28–37. [潘之欣. (2021). 影响听力理解任务难度的试题因素研究. 《外语测试与教学》, (2): 28–37.]

Pan, Z., & Fan, X. (2021). Discourse features that influence the perception of difficulty in listening comprehension tasks, *Contemporary Foreign Language Studies*, (6), 119–131. [潘之欣, 樊雪. (2021). 影响听力理解任务难度感知的语篇特征. 《当代外语研究》, (6): 119–131.]

Papageorgiou, S., & Manna, V. F. (2021). Maintaining access to a large-scale test of academic language proficiency during the pandemic: The launch of TOEFL iBT Home Edition. *Language Assessment Quarterly*, *18*(1), 36–41. https://doi.org/10.1080/15434303.2020.1864376

Petour, M. T. F. (2015). Validity and equity in educational measurement: The case of SIMCE. *Psicoperspectivas*, *14*(3), 31–44. https://doi.org/10.5027/-psicoperspectivas-Vol14-Issue3-fulltext-618

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Randez, R. A., & Cornell, C. (2023). Advancing equity in language assessment for learners with disabilities. *Language Testing*, *40*(4), 984–999. https://doi.org/10.1177/02655322231169442

Raquel, M. (2019). The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment research. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment, Vol. I: Fundamental techniques* (pp. 103–131). Routledge. https://doi.org/10.4324/9781315187815-6

Rasooli, A. (2021). *Fairness in classroom assessment: Conceptual and empirical investigations* [Unpublished doctoral dissertation]. Queen's University.

Rasooli, A., Zandi, H., & DeLuca, C. (2018). Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond. *Studies in Educational Evaluation*, *56*, 164–181. https://doi.org/10.1016/j.stueduc.2017.12.008

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing*, *14*(3), 304–314. https://doi.org/10.1177/026553229701400307

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31–65. https://doi.org/10.1017/S02722631-12000678

Rezai, A. (2022). Fairness in classroom assessment: Development and validation of a questionnaire. *Language Testing in Asia*, *12*. https://doi.org/10.1186/s40468-022-00162-9

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.-v048.i02

Roussos, L. A., & Stout, W. F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371. https://doi.org/10.1177/014662169602000404

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, *33*(2), 215–230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Roussos, L., & Stout, W. F. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 107–115). Sage Publications. https://doi.org/10.4135/-9781412986311.n6

Sabbaghan, S., & Fazel, I. (2023). None of the above: Integrity concerns of standardized english proficiency tests. In S. E. Eaton, J. J. Carmichael, & H. Pethrick (Eds.), *Fake degrees and fraudulent credentials in higher education* (pp. 169–185). Springer. https://doi.org/10.1007/978-3-031-21796-8_8

Safari, P., & Rashidi, N. (2018). Democratic assessment as scales of justice: The case of three Iranian high-stakes tests. *Policy Studies*, *39*(2), 127–144. https://doi.org/10.1080/01442872.2018.1435042

Saito, Y., & Joachims, T. (2022). Fair ranking as fair division: Impact-based individual fairness in ranking. *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1514–1524. https://doi.org/10.1145/3534678.3539353

Semiyari, S. R., & Ahangari, S. (2022). Examining differential item functioning (DIF) for Iranian EFL test takers with different fields of study. *Research in English Language Pedagogy*, *10*(1), 169–190. https://doi.org/10.304-86/relp.2021.1935588.1295

Sen, A. (2009). *The idea of justice*. Belknap Press of Harvard University Press.

Shaw, S. D., & Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based

examinations. *Language Assessment Quarterly*, *10*(4), 452–475. https://doi.org/10.1080/15434303.2013.866117

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Shepard, L. A. (1987). The case for bias in tests of achievement and scholastic aptitude. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 177–190). The Falmer Press.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317–375. https://doi.org/10.3102/10769986006004317

Shi, Y. (2019). *Investigating comparability of writing tasks with picture prompts in an in-house English proficiency test* [Unpublished doctoral dissertation]. Zhejiang University.

Shimizu, Y., & Zumbo, B. D. (2005). A logistic regression for differential item functioning primer. *Japan Language Testing Association Journal*, *7*, 110–124. https://doi.org/10.20622/jltaj.7.0_110

Sireci, S. G. & Gándara, M. F. (2016). Testing in educational and developmental settings. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger & D. Iliescu (Eds.), *International test commission handbook of testing and assessment* (pp. 187–202). Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0013

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–135). Routledge. https://doi.org/10.4324/9780367815318-6

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, *4*, Article 43. https://doi.org/10.3389/feduc.2019.00043

Song, X. (2014). *Test fairness in a large-scale high-stakes language test* [Unpublished doctoral dissertation]. Queen's University.

Song, X. (2018). The fairness of a graduate school admission test in China: Voices from administrators, teachers, and test-takers. *The Asia-Pacific Education Researcher*, *27*(2), 79–89. https://doi.org/10.1007/s40299-018-0367-4

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Papers in Language Testing and Assessment*, *4*(1), 97–124. https://doi.org/10.58379/rshg8366

Sonnleitner, P., & Kovacs, C. (2020). Differences between students' and teachers' fairness perceptions: Exploring the potential of a self-administered questionnaire to improve teachers' assessment practices. *Frontiers in Education*, *5*, 1–14. https://doi.org/10.3389/feduc.2020.00017

Spaan, M. (2000). Enhancing fairness through a social contract. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 35–37). Cambridge University Press.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, *12*(3), 275–287. https://doi.org/-10.1080/09695940500337249

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, *51*(2), 161–179. https://doi.org/10.1080/0013188-0902891305

Sung, H., Cho, S., & Kyle, K. (2024). An empirical evaluation of lexical diversity indices in L2 Korean writing assessment. *Language Assessment Quarterly, 21*(2), 159–180. https://doi.org/10.1080/15434303.2024.231-1728

Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics* (6th ed.). Pearson Education.

Tajeddin, Z., Khatib, M., & Mahdavi, M. (2022). Critical language assessment literacy of EFL teachers: Scale construction and validation. *Language Testing*, *39*(4), 649–678. https://doi.org/10.1177/02655322211057040

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340. https://doi.org/10.-1177/026553220001700303

Taylor, L. (2023). Reframing the discourse and rhetoric of language testing and assessment for the public square. *Language Testing*, *40*(1), 47–53. https://doi.org/10.1177/02655322221127421

Tierney, R. D. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation*, *43*, 55–69. https://doi.org/-10.1016/j.stueduc.2013.12.003

Tierney, R. D. (2017). Fairness in educational assessment. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 793–798). Springer. https://doi.org/10.1007/978-981-287-588-4_400

Tofighi, S., & Safa, M. A. (2023). Fairness in classroom language assessment from EFL Teachers' Perspective. *Teaching English as a Second Language Quarterly*, *42*(2), 81–110. https://doi.org/10.22099/tesl.2023.46825.3173

Toulmin, S. (1958/2003). *The uses of argument*. Cambridge University Press.

Tsai, Y., & Tsou, C. H. (2009). A standardised English language proficiency test as the graduation benchmark: Student perspectives on its application

in higher education. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 319–330. https://doi.org/10.1080/09695940903319711

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211–234. https://doi.org/10.1191/0265532205lt301oa

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, *26*(2), 191–208. https://doi.org/10.1111/j.1745-3984.1989.-tb00328.x

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized-adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wallace, M. P., & Ng, J. S. W. (2023). Fairness of classroom assessment approach: Perceptions from EFL students and teachers. *English Teaching & Learning*, *47*(4), 529–548. https://doi.org/10.1007/s42321-022-00127-4

Wallace, M. P., & Qin, C. Y. (2021). Language classroom assessment fairness: Perceptions from students. *Language Education and Acquisition Research Network*, *14*(1), 492–521.

Walters, F. S. (2022). Ethics and fairness. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 563–577). Routledge. https://doi.org/10.4324/9781003220756-44

Wang, L. C. (2016). The effect of high-stakes testing on suicidal ideation of teenagers with reference-dependent preferences. *Journal of Population Economics*, *29*, 345–364. https://doi.org/10.1007/s00148-015-0575-7

Weideman, A. (2017). Does responsibility encompass ethicality and accountability in language assessment?. *Language & Communication*, *57*, 5–13. http://dx.doi.org/10.1016/j.langcom.2016.12.004

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weir C. (2019). Global, local, or “glocal”: Alternative pathways in English language test provision. In Su L. I., Weir C., Wu J. R. W. (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 193–225). Routledge. https://doi.org/10.4324/978135-1254021-8

Weir, C. J., & Wu, J. R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, *23*(2), 167–197. https://doi.org/10.1191/0265532206lt326oa

West, M. (1953). *A general service list of English words*. Longman.

Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (Ed.), *Assessment in higher education: Issues in access, quality, student development, and public policy* (pp. 213–242). Lawrence Erlbaum Associates.

Willingham, W. W., & Cole, N. (1997). *Gender and fair assessment*. Lawrence Erlbaum Associates.

Wollack, J. A., & Case, S. M. (2016). Maintaining fairness through test administration. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 33–53). Routledge.

Worrell, F. C. (2016). Commentary on perspectives on fair assessment. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 283–293). Routledge. https://doi.org/10.4324/9781315-774527-26

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

Xiao, Y. (2013). *Detection of DIF across different academic backgrounds and genders in CET-4* [Unpublished master’s dissertation]. Guangdong University of Foreign Studies. [肖园园. (2013). 《大学英语四级考试对

不同学术背景和不同性别学生的项目功能差异研究》. 广东外语外贸大学.]

Xu, W. (2019). Ensuring test fairness—Taking the NMET (Shanghai) as an example. *Foreign Language Testing and Teaching*, (2): 9–16. [徐雯. (2019). 落实考试公平性——以上海英语高考为例. 《外语测试与教学》, (2): 9–16.]

Yang, H., & Gui, S. (2007). The sociology of language testing. *Modern Foreign Languages*, *30*(4), 368–374. [杨惠中, 桂诗春. (2007). 语言测试的社会学思考. 《现代外语》, *30*(4): 368–374.]

Yang, Z., Zeng, Y., & Chen, G. (2022). Research on the fairness of the Practical English Test for Colleges. *Foreign Language Testing and Teaching*, (4): 18–27. [杨志强, 曾用强, 陈刚. (2022). 高等学校英语应用能力考试公平性研究. 《外语测试与教学》, (4): 18–27.]

Yao, D. (2023). Examining the subjective fairness of at-home and online tests: Taking Duolingo English Test as an example. *Plos One*, *18*(9), Article e0291629. https://doi.org/10.1371/journal.pone.0291629

Yao, D. (2024). Does perceived test fairness affect test preparation?—A case study of Duolingo English Test. *Heliyon*, *10*(23), Article e40579. https://doi.org/10.1016/j.heliyon.2024.e40579

Yoo, H., & Manna, V. (2017). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing*, *34*(1), 101–126. https://doi.org/10.1177-/0265532215618987

Yoon, S.-Y., Cho, Y., & Napolitano, D. (2016). Spoken text difficulty estimation using linguistic features. *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 267–276). Association for Computational Linguistics. https://doi.org/10.18653/v1/-w16-0531

Young J. W., So Y., Ockey G. J. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments*. Educational Testing Service.

Zeng, Y. (2022). A study of text features within the framework of China's Standards of English Language Ability. *Language Testing and Assessment*, (1), 58–69. [曾用强. (2022). 基于《中国英语能力等级量表》的文本特征研究. 《语言测试与评价》, 2022, (1): 58–69.]

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, Article 100505. https://doi.org/-10.1016/j.asw.2020.100505

Zhang, K., & Jin, L. (2012). Study on the differential item functioning of the listening test of the CET-4. *Journal of Inner Mongolia Normal University* (Educational Science), *25*(7), 107–111. [张琨, 金力. (2012). CET-4 听力测试的项目偏差研究. 《内蒙古师范大学学报(教育科学版)》, *25*(7): 107–111.]

Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied Linguistics*, *18*(1), 49–68. https://doi.-org/10.1093/applin/18.1.49

Zhejiang University. (n.d.). *University profile*. https://www.zju.edu.cn/engli-sh/mission_vision/list.htm

Zhu, X., & Aryadoust, V. (2019). Examining test fairness across gender in a computerized reading test: A comparison between the Rasch-based DIF technique and MIMIC. *Papers in Language Testing and Assessment*, *8*(2), 65–90. https://doi.org/10.58379/nvft3338

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203874776.ch16

Zieky, M. J. (2015). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.

Zumbo B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*(2), 136–147. https://doi.org/10.1191/0265532203lt248oa

Appendices

Appendix 1

Summary of DIF, DBF, and DTF studies on listening comprehension across test-taker groups with humanities and science backgrounds

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
Henning (1990)	English as a Second Language Placement Exam (ESLPE)	Angoff item difficulty by group plotting, regression residual analyses & MH	DIF	- 1/30 (3.33%) item showed uniform DIF, favoring the science group	/
Pae (2004)	English subtest of the Korean National Entrance Exam for Colleges and Universities	MH & IRT-LRT	DIF	- 5/17 (29.41%) items showed DIF, with 2 items displaying uniform DIF, and 3 items displaying both uniform and nonuniform DIF (all with small effect size) - The 2 uniform DIF items favored the humanities group - All the nonuniform DIF items were more discriminating for the humanities group	- Topical knowledge - Item discrimination
Zhang & Jin (2012)	Listening section of the CET-4	Independent samples-*t* tests	DIF	- 3/36 (8.33%) items showed uniform DIF, with 2 items favoring the humanities group, and one favoring the science group	- Strategy use - Topic familiarity - Vocabulary familiarity

(Continued)

(Continued).

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
Chen (2013)	Listening subtest of a standard English test	MH & LR	DIF	- 5/15 (33.33%) items showed DIF, with 3 displaying uniform DIF, and 2 displaying nonuniform DIF - All the 3 uniform DIF items (2 at a moderate level, and one at a large level) favored the natural science group	- Test content - Exposure to English
Xiao (2013)	CET-4	MH & LR	DIF	- 5/25 (20%) items showed DIF, incl. 3 moderate/large level uniform DIF items and 2 nonuniform DIF items - Among the 3 uniform DIF items, one item favored the science group, and 2 favored the humanities group - Both nonuniform DIF items favored the higher ability members in the humanities group and lower ability members in the science group	- Sensitivity to culture-related information - Exposure to discip-line-related topics - Sensitivity to culture-related information - Exposure to discip-line-specific topics
He (2022)	CET-6	LR	DIF	- 2/25 (8%) items showed large DIF, with one item favoring the science group, and one favoring the humanities group	- Background knowledge - Personal interest - Item discrimination
Semiyari & Ahangari (2022)	MSRT proficiency test	IRT-LRT	DIF	- 6/30 (20%) items showed DIF, with 2 items displaying uniform DIF, and 4 displaying both uniform and nonuniform DIF	- /

(*Continued*)

(Continued).

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
Yang et al. (2022)	Practical English Test for Colleges (PRETCO)	MIMIC & MG-CFA	DBF & DTF	- 2/3 (66.67%) testlets showed negligible DBF, both favoring the science group - Listening comprehension section showed negligible DTF and favored the science group	Strategy use
Aryadoust et al. (2024)	Listening subtest of an in-house high-stakes exit test	Rasch & MH	DIF	- 6/50 (12%) items showed uniform DIF (2 at a negligible level, 2 at a moderate level, and 2 at a large level)	/

Notes. Studies included in this table are presented in chronological order. DIF = Differential item functioning. DBF = Differential bundle functioning. DTF = Differential test functioning. MH = Mantel-Haenszel. LR = Logistic regression. IRT-LRT = Item-response-theory-likelihood-ratio test. MIMIC = Multiple indicators multiple causes.

Appendix 2

Summary of DIF, DBF, and DTF studies on reading comprehension across test-taker groups with humanities and science backgrounds

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
Henning (1990)	English as a Second Language Placement Exam (ESLPE)	The Angoff item difficulty by group plotting method, regression residual analyses & MH	DIF	- No items showed DIF	/
Pae (2004)	English subtest of the Korean National Entrance Exam for Colleges and Universities	MH & IRT-LRT	DIF	- 13/38 (34.21%) items showed DIF, with 4 items displaying uniform DIF, 2 displaying nonuniform DIF, and 7 displaying both uniform and nonuniform DIF - One uniform DIF item favored the humanities group, and 3 favored the science group - One uniform and 2 nonuniform DIF items were at a moderate level - All the nonuniform DIF items were more discriminating for the humanities group	- Topical knowledge - Item discrimination

(Continued)

(Continued).

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
Brati et al. (2006)	English subtest of the Iranian National University Entrance Exam	IRT	DIF	- 14/15 (93.33%) items showed uniform DIF - 7 DIF items were identified as easier or the easiest for the humanities group, 8 for the science group, and 7 for the mathematics group - Overall, the reading section were the easiest for the mathematics group and easier for the science than the humanities group	- Subskill assessed by the items - Academic background
Alavi et al. (2011)	University of Tehran English Proficiency Test	Generalized MH & LR	DIF	- 6/26 (23.08%) items showed DIF (5 items displayed uniform DIF, and one displayed nonuniform DIF) - An item favored the humanities group (the direction of other DIF items not reported) - The magnitude of DIF was not statistically significant for any of the DIF items	/
Min (2011)	Reading comprehension section of the ESLPE	MH & LR	DIF	- 3/20 (15%) items showed DIF (one uniform DIF and two nonuniform DIF) - The uniform DIF item favored the biological/physical science group	- Topical knowledge - English language proficiency
*Liu (2011)	TOEFL® iBT	A two-stage standardization method	DIF & DBF	- 81/84 (96.43%) items showed uniform DIF (63 at a small level, 13 at a moderate level, and 5 at a large level)	- Topic familiarity - Terminology familiarity

(*Continued*)

(Continued).

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
				- 2 moderate and a large DIF items favored test takers familiar with Eastern European, Western European, and Scandinavian culture; 4 moderate and a large DIF items favored those familiar with East Asian culture; 3 moderate and 3 large DIF items favored those not familiar with East Asian culture; 3 moderate DIF items favored those with physical science majors; a moderate DIF item favored those without physical science majors - No testlets (n = 6) showed DBF - A bundle with moderate DBF favored test takers familiar with Eastern European, Western European, and Scandinavian culture; a bundle with moderate DBF favored test takers familiar with science terminology	- Cultural familiarity - Academic background - English language proficiency
Xiao (2013)	CET-4	MH & LR	DIF	- 7/30 (23.33%) items showed DIF, incl. 5 moderate/large level uniform DIF items and 2 nonuniform DIF items - Among the 5 uniform DIF items, 2 items favored the science group, and 3 favored the humanities group - Among the 2 nonuniform DIF items, one favored the higher ability members of the humanities group and lower ability members of the science group, and one favored the lower ability members of the humanities group and higher ability members of the science group	- Exposure to discipline-specific topics - Response format

(Continued).

Song et al. (2015)	Graduate School Entrance English Examination (GSEEE)	SIBTEST & Poly-SIBTEST	DIF & DBF	- 1/40 (2.5%) item showed moderate DIF, favoring the science group - 1/6 (16.67%) testlet showed moderate DBF, favoring the science group	- Exposure to English vocabulary specific to discipline-related topics - Background knowledge
He (2018)	Reading comprehension section of an in-house achievement test for first-year college students	MH & LR	DIF	- 14/15 (93.33%) items showed DIF - Among the items showing uniform DIF, 12 items were at a small level, and one at a moderate level - An item showed both uniform and nonuniform DIF	- Item discrimination
*Chen & Zeng (2021)	Reading section of an in-house college English test	SIBTEST & IRT-LRT	DBF	- 1/4 testlet (25%) about sport showed DBF, favoring the humanities group - 2/6 (33.33%) bundles assessing scanning and skimming strategies showed DBF (one favored the humanities group, and one favored the science group)	- Background knowledge
Semiyari & Ahangari (2022)	Ministry of Science, research, and technology (MSRT) proficiency test	IRT-LRT	DIF	- 4/40 (10%) items showed DIF (direction of the DIF items not disclosed)	- /

(*Continued*)

(Continued).

Study	Test	Detection method	Detection type	Key findings	DIF/DBF causes
				-	
Yang et al. (2022)	Practical English Test for Colleges (PRETCO)	MIMIC & MG-CFA	DBF & DTF	- 3/4 (75%) testlets showed negligible DBF, favoring the humanities group - Reading comprehension section showed negligible DTF, favoring the humanities group	- English language proficiency - Background knowledge - Subskill assessed by the item
Ghaemi & Khorami (2024)	An English proficiency test	MH & IRT	DIF	- 5/60 items (8.33%) showed uniform DIF, with 3 items (1 at a small level and 2 at a moderate level) favoring the humanities group, and 2 items (both at a small level) favoring the science and engineering group	- Topic of the reading passages - English language proficiency
Jafaripour et al. (2024)	Islamic Azad University English Proficiency Test	Rasch	DIF	- 2/35 (5.71%) items showed uniform DIF, both favoring the science group	- Exposure to English - Cognitive abilities - Learning motivation - Academic background

Appendix 3

“英语水平测试的公平性”调查问卷

亲爱的同学：

您好!我们希望了解您对英语水平测试公平性的看法,以期进一步提升考试服务质量。请花费一点宝贵时间,仔细阅读题目后按实际情况作答。问卷不记名,答案也无对错之分,请放心填写。感谢您的参与!

一、请填写或勾选您的基本信息。

1. 性别：☐ 女 ☐ 男 2. 年龄：____ 3. 专业：_____

4. 年级：☐ 大一 ☐ 大二 ☐ 大三 ☐ 大四 ☐ 大五 ☐ 其他（请说明）：___

5. 参加水平测试机考的次数（含本次考试）：☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ > 5

6. 参加水平测试口试的次数：☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ > 5

二、请阅读下列表述，并根据自己对下列表述的同意/不同意程度在相应位置打“√”。

		同意/不同意程度					
		非常不同意	不同意	基本不同意	基本同意	同意	非常同意
1	考试之前，我了解水平测试的目的。	1	2	3	4	5	6
2	考试之前，我了解水平测试的报考条件。	1	2	3	4	5	6
3	考试之前，我了解水平测试的报名流程。	1	2	3	4	5	6
4	考试之前，我了解水平测试的考试流程。	1	2	3	4	5	6
5	考试之前，我了解水平测试的考试形式（如机考、口试）。	1	2	3	4	5	6
6	考试之前，我了解水平测试考查的语言能力（如听力、阅读、写作、口语）。	1	2	3	4	5	6
7	考试之前，我了解水平测试的题型（如选择题、选词填空）。	1	2	3	4	5	6
8	考试之前，我了解水平测试各部分的分值。	1	2	3	4	5	6
9	考试之前，我了解水平测试的评分标准。	1	2	3	4	5	6
10	考试之前，我了解水平测试的评分方式（如机评、人评）。	1	2	3	4	5	6
11	考试之前，我了解水平测试的成绩报告方式（即合格或不合格）。	1	2	3	4	5	6
12	考试之前，我知道水平测试合格后才能获取相应学分。	1	2	3	4	5	6

		同意/不同意程度					
		非常不同意	不同意	基本不同意	基本同意	同意	非常同意
13	考试之前，我可以通过官方文件（如考试大纲）、校内论坛等途径了解考试内容所涉及的话题（如人文社科类、自然科学类话题）。	1	2	3	4	5	6
14	考试之前，我可以通过修读英语课程等方式提升水平测试所考查的英语能力。	1	2	3	4	5	6
15	考试之前，我可以通过相应的练习熟悉水平测试考查的题型（如选择题、选词填空）。	1	2	3	4	5	6
16	考试之前，我可以通过备考提升水平测试所考查的微技能（如理解主旨、理解细节、进行推断）。	1	2	3	4	5	6
17	我所在的校区距离考试地点很远，如遇不良天气、交通堵塞等状况，无法按时抵达考场参加考试。	1	2	3	4	5	6
18	我所在的校区距离考试地点很近，可以按时抵达考场参加考试。	1	2	3	4	5	6
19	入场时，监考人员对我的身份证件进行了核验。	1	2	3	4	5	6
20	开考前，我需要关闭手机等电子设备。	1	2	3	4	5	6
21	开考前，我需要将与考试无关的物品放至考场内的指定地点。	1	2	3	4	5	6
22	我的考试环境和其他考生相同（如考位间隔、考场的光线、听力音频的音量大小）。	1	2	3	4	5	6
23	考试当天，我经历的考试流程（如证件核验、考试界面登录、试卷提交）和其他考生相同。	1	2	3	4	5	6
24	所有考生各语言技能的考查顺序是相同的（即依次考查听力、阅读、写作）。	1	2	3	4	5	6
25	所有考生需要作答的题目数量是相同的。	1	2	3	4	5	6
26	所有考生需要作答的题型（如选择题、选词填空）是相同的。	1	2	3	4	5	6
27	所有考生的给定作答时间是相同的。	1	2	3	4	5	6
28	水平测试的考试形式（如机考、面对面口试）不利于我展示自己真实的英语水平。	1	2	3	4	5	6
29	水平测试的题型（如选择题、选词填空）不利于我展示自己真实的英语水平。	1	2	3	4	5	6
30	本次考试的内容（如某些话题、观点）令我感到不适，影响了我的正常发挥。	1	2	3	4	5	6
31	我对本次考试涉及的话题不熟悉，这影响了我的正常发挥。	1	2	3	4	5	6

		同意/不同意程度					
		非常不同意	不同意	基本不同意	基本同意	同意	非常同意
32	针对本次考试所涉及的话题，我的背景知识储备不足，影响了我的正常发挥。	1	2	3	4	5	6
33	我对水平测试的考试界面不太熟悉,这影响了我的正常发挥。	1	2	3	4	5	6
34	水平测试的机考界面不那么友好，影响了我的正常发挥。	1	2	3	4	5	6
35	我对考试硬件设备（如键盘）的操作不熟练，影响了我的正常发挥。	1	2	3	4	5	6
36	考试过程中，硬件设备（如显示器、鼠标、键盘）出现故障，影响了我的正常发挥。	1	2	3	4	5	6
37	考试过程中，考场内或周边的噪音影响了我的正常发挥。	1	2	3	4	5	6
38	考试期间，考场内的监控设备实时监控考生的作答情况。	1	2	3	4	5	6
39	考试过程中，监考人员会在教室内巡考。	1	2	3	4	5	6
40	考试作弊行为在考试过程中是被严令禁止的。	1	2	3	4	5	6
41	一旦发现作弊现象，监考人员会及时制止。	1	2	3	4	5	6
42	考生如有替考、作弊等违纪行为，需要承担相应后果。	1	2	3	4	5	6
43	所有考生获取考试结果的渠道是相同的。	1	2	3	4	5	6
44	如果我对考试结果有异议，我可以申请查分。	1	2	3	4	5	6
45	如果我对考试结果有异议，我可以要求重新评卷。	1	2	3	4	5	6
46	针对考试过程中其他考生的违纪行为，我可以举报。	1	2	3	4	5	6
47	我可以对监考人员的不当行为进行投诉。	1	2	3	4	5	6
48	如果我对考试当天的体验不满意,我可以向有关部门反映。	1	2	3	4	5	6

Appendix 4

针对考生的访谈提纲

第一部分：开场

- 向受访者表达感谢
- 介绍研究者以及研究目的
- 介绍访谈目的、流程及预计时长
- 说明受访者的权利
- 征得受访者同意以继续进行访谈并录音
- 受访者介绍自己的年级、专业、英语学习经历、参加英语水平测试的次数以及英语对于自己毕业后生涯规划的重要性

第二部分：对英语水平测试公平性的总体看法

1. 总体而言，您如何看待英语水平测试的公平性？
2. 根据您的经历，您认为该英语水平测试在哪些方面是公平的？您能具体说明为什么您认为英语水平测试在这些方面公平吗？
3. 根据您的经历，您认为该英语水平测试在哪些方面是不公平的？您能具体说明为什么您认为英语水平测试在这些方面不公平吗？

第三部分：对英语水平测试公平性不同维度的看法

可比性

1. 您认为英语水平测试在不同考生群体中测量的构念是否相同？
2. 您如何看待不同场次英语水平测试在难度和测试结果方面的可比性？

可及性

1. 在您看来，考生能否在考前获取有关英语水平测试的信息？如果能，他们通常是怎样获取这些信息的？如果不能，什么原因导致他们难以获取这些信息？
2. 您认为考生是否拥有充足的考试机会？能否详细阐述您的观点？
3. 您认为学生是否拥有充足的学习资源？能否详细阐述您的观点？

4. 您认为考试地点对所有考生而言是否便利？能否详细阐述您的观点？
5. 考生在考试之前是否拥有熟悉考试设备的机会？能否详细阐述您的观点？
6. 考生在考试之前是否拥有熟悉考试系统的机会？能否详细阐述您的观点？

一致性

1. 在您看来，考试环境和考试流程是否在每次施考过程中保持一致？

问责制

1. 对于学校将英语水平测试作为毕业要求之一的做法，您是如何看待的?
2. 考试结束后，考生能否申请成绩复议？
3. 考生是否有机会反映有关英语水平测试及相关测评实践的意见或建议？
4. 您认为谁应该对测试公平性负责？为什么？
5. 目前，是否存在某种机制确保相关人员对英语水平测试的公平性负责？如果有，您能提供一些具体的例子吗？如果没有，您对于问责机制的有效落实有何建议？

第四部分：家庭及教育背景调查问卷

1. 您高中所在的学校位于（填写省份）
2. 您高中所在的学校地处：乡镇/县城/地级市/省会城市/直辖市
3. 您的家庭所在地为：乡镇/县城/地级市/省会城市/直辖市
4. 父亲的最高文化程度是：小学及以下/初中/高中或中专/大专、高职或本科/研究生及以上
5. 母亲的最高文化程度是：小学及以下/初中/高中或中专/大专、高职或本科/研究生及以上
6. 父母对您的学业期待：不高/一般/较高/非常高

访谈尾声

- 向受访者征求有关访谈或本研究的意见和建议
- 再次向受访者表达感谢

Appendix 5

针对教师、施考人员和考试使用者的访谈提纲

第一部分：开场

- 向受访者表达感谢
- 介绍研究者以及研究目的
- 介绍访谈目的、流程及预计时长
- 说明受访者的权利
- 征得受访者同意以继续进行访谈并录音
- 受访者介绍自己的教育背景、工作经历以及在英语水平测试中承担的具体职责

第二部分：对英语水平测试公平性的总体看法

1. 总体而言，您如何看待英语水平测试的公平性？
2. 根据您的经历，您认为该英语水平测试在哪些方面是公平的？您能具体说明为什么您认为英语水平测试在这些方面公平吗？
3. 根据您的经历，您认为该英语水平测试在哪些方面是不公平的？您能具体说明为什么您认为英语水平测试在这些方面不公平吗？

第三部分：对英语水平测试公平性不同维度的看法

可比性

1. 您认为英语水平测试在不同考生群体中测量的构念是否相同？
2. 您如何看待不同场次英语水平测试在难度和测试结果方面的可比性？

可及性

1. 在您看来，考生能否在考前获取有关英语水平测试的信息？如果能，他们通常是怎样获取这些信息的？如果不能，什么原因导致他们难以获取这些信息？
2. 您认为考生是否拥有充足的考试机会？能否详细阐述您的观点？

3. 您认为学生是否拥有充足的学习资源？能否详细阐述您的观点？
4. 您认为考试地点对所有考生而言是否便利？能否详细阐述您的观点？
5. 考生在考试之前是否拥有熟悉考试设备的机会？能否详细阐述您的观点？
6. 考生在考试之前是否拥有熟悉考试系统的机会？能否详细阐述您的观点？

一致性

1. 您认为英语水平测试的设计与开发是否具有一致性和规范性？能否详细阐述您的观点？
2. 在您看来，考试环境和考试流程是否在每次施考过程中保持一致？
3. 在您看来，英语水平测试的评分实践是否在不同评分员之间和不同考试场次之间保持一致？
4. 您认为考试使用者对于考试分数的解释是否一致？如果是，哪些措施确保了分数解释的一致性？如果不是，您认为可通过哪些方式提升分数解释的一致性？

问责制

1. 对于学校将英语水平测试作为毕业要求之一的做法，您是如何看待的?
2. 考试结束后，考生能否申请成绩复议？
3. 考生是否有机会反映有关英语水平测试及相关测评实践的意见或建议？
4. 您认为谁应该对测试公平性负责？为什么？
5. 目前，是否存在某种机制确保相关人员对英语水平测试的公平性负责？如果有，您能提供一些具体的例子吗？如果没有，您对于问责机制的有效落实有何建议？

访谈尾声

- 向受访者征求有关访谈或本研究的意见和建议
- 再次向受访者表达感谢

Appendix 6

研究项目信息表

我们诚挚地邀请您参加“高风险语言测试的公平性研究”项目。本知情同意书将向您介绍该研究的相关信息，以帮助您决定是否参与该研究。请您仔细阅读以下内容，如有任何疑问，请向研究人员提出，研究人员会为您及时解答。如果您决定参与该研究，请签署知情同意书。

[研究内容]

涉考群体对“***大学英语水平测试”公平性的看法

[研究人员]

浙江大学外国语学院 在读博士生张娟

[研究目的]

本研究旨在了解***大学非英语专业本科生、考试开发者、施考人员及考试使用者对“***大学英语水平测试”公平性的看法，以期提升考试服务质量。

[研究过程]

本研究包含三个部分。第一部分，研究人员会就您的个人背景进行提问。第二部分，研究人员将了解您对测试公平性的理解。第三部分，研究人员将询问您对“***大学英语水平测试”公平性的看法。整个过程大约需要花费您 30–45 分钟的时间。

您的参与完全是自愿的。访谈过程中，如果您有疑问，可以随时向研究人员提出。在征得您同意的情况下，研究人员将对访谈进行录音。您可以随时退出本研究。如果拒绝参与或中途退出本研究，您将不会受到任何形式的伤害或惩罚。针对令您感到不适的研究问题，您可以拒绝回答。

[受访者获益及风险]

您在本研究中没有直接获益。但是，您的参与能够帮助研究人员深入了解校本语言考试语境下不同涉考群体对测试公平性的看法。您提供的信息有助于研究人员了解威胁测试公平性的因素，为提升测试科学性、保证测试公平性、促进决策合理性做出贡献。

研究过程中的部分问题可能涉及个人隐私。信息保护条款详见下方，请您仔细阅读后再决定是否参与本研究。

[信息保护]

如果您决定参与本研究，研究人员将对访谈内容和您的个人信息进行严格保密。本研究收集的资料仅用于科学研究和学术发表。受访者信息、访谈转写稿及调查问卷将通过编码加以标识。所有研究资料将无限期存储于设置密码的移动硬盘内，仅供研究人员查阅。

[联系方式]

如果您对本研究有任何疑问，请随时联系研究人员。

姓名：张娟

地址：浙江大学外国语学院东五-206

邮箱：zhangjuan@zju.edu.cn

手机：17816862122

Appendix 7

Consent form

受访知情同意书

感谢您参与本研究。在您签署本同意书之前，研究人员有义务为您解答与本研究相关的任何疑问。因此，如有疑问，请向研究人员提出。研究结束后，您可以通过以下联系方式了解本研究的具体信息和研究进展。

[研究内容]

涉考群体对“***大学英语水平测试”公平性的看法

[研究人员]

姓名：张娟
地址：浙江大学外国语学院东五-206
邮箱：zhangjuan@zju.edu.cn
手机：17816862122

[信息确认]

若您已阅读研究项目信息表，对本研究没有疑问，请确认以下条款内容，并签名。

请您确认：

- ☐ 我已知晓本研究的基本情况
- ☐ 我已知晓我可以随时退出本研究
- ☐ 我已知晓访谈过程将全程录音
- ☐ 我已知晓访谈录音转写稿将被研究人员用于后续的分析和研究
- ☐ 我已知晓研究结果会被用于学术发表，我的个人信息将作匿名处理，并被严格保密
- ☐ 我有充足的时间和机会进行提问，并对研究者给予的答复很满意
- ☐ 我确认参与本研究

受访者姓名：_____

受访者签名：_____

日　　期：_____

Appendix 8

Coding scheme of the stakeholders' perceptions of the EPT's fairness

Code	Subcode (Level 1)	Subcode (Level 2)	Example interview excerpt
Comparability	Comparability of the construct measured by the EPT across test-taker groups		*"The test topics do not show bias toward test takers from different fields of study."* (TU3)
	Comparability of test results across test forms		*"From a technical standpoint, anchor items are embedded in multiple test forms used for each administration. Moreover, since the initial administration of the test, post-test equating method has remained the same. This ensures that test-takers' scores remain comparable across test forms, test sessions, and academic years."* (TU2)
Accessibility	Test information accessibility		*"The staff from the university's Academic Affairs Office issues the registration notice for the English Proficiency Test. There are many attachments to the notice, including the test syllabus, administration details, and a sample test paper."* (TT12)
	Test-taking opportunities		*"Test takers have the flexibility to register for the test according to their individual preferences. They can choose to take the test when they feel fully prepared and confident in their ability to pass the test."* (TD3)
	Learning opportunities		*"Undergraduates are encouraged to attend tutorial sessions, during which teachers of College English courses are available to provide academic support."* (TD2)
	Test location accessibility		*"The university provides a shuttle bus service to facilitate the transportation of test takers from their respective campuses to the test location."* (TA2)
	Delivery system accessibility		*"I don't know how the delivery system works before taking the test. I wish there were a mock system available for me to familiarize myself with it beforehand."* (TT19)

(Continued)

(Continued).

Code	Subcode (Level 1)	Subcode (Level 2)	Example interview excerpt
Consistency	Test administration	Equipment functionality	*"I didn't experience any equipment malfunctions during the test."* (TT5)
		Test environment	*"During the test, test takers are seated separately with dividers placed between the seats."* (TT1)
		Test security	*"Each testing room was staffed with two proctors to ensure discipline."* (TA2)
	Rating	Rating methods	*"The integration of human and machine rating methods ensures the consistency and reliability in the ratings of the test-takers' writing scripts."* (TD5)
		Rating procedures	*"If there is a large discrepancy between the two sets of ratings, a second human rater will review or re-score the compositions to ensure the accuracy of the ratings."* (TD5)
	Score interpretation		*"By referring to the cut scores used for classifying test-takers' performance into corresponding CSE levels, we know whether test takers have attained CSE-5 and determine their pass or fail status on the test."* (TU2)
Accountability	Score review		*"I think the availability of score review opportunities is essential for ensuring test fairness."* (TT13)
	Stakeholder engagement		*"Allowing stakeholders to voice their concerns and expectations of test fairness can contribute to a democratic policy-making process."* (TU3)
	Fairness evaluation		*"Test takers can contribute to the evaluation of test fairness because they have firsthand test-taking experience."* (TA2)
	Responsibility for test fairness		*"Test developers should draw on their expertise in language testing to ensure that the test results accurately reflect test-takers' English proficiency levels."* (TU3)

Notes. Each participant was assigned a unique code for ease of reference in reporting the results. Test takers are coded as TT1–TT20; teachers (also test developers) as TD1–TD6; test administrators as TA1–TA2; and test users as TU1–TU3.

Appendix 9

Coding scheme of the factors influencing the stakeholders' perceived fairness of the EPT

Code	Subcode	Example interview excerpt
Sociocultural factors	Importance of English proficiency in China	*"As an international lingua franca, English is undoubtedly one of the most important languages. We must ensure that the undergraduates have adequate general English proficiency upon graduation to meet the basic language requirements in workplace or academic settings."* (TU1)
	Societal norms around English testing in China	*"I think it is reasonable for the university to require undergraduates to pass the English Proficiency Test before graduation. As far as I know, most universities in China impose specific English proficiency requirements for their undergraduates. For instance, some universities require undergraduates to pass the CET-4 before graduation."* (TT18)
Educational factors	Disparities in pre-university English education	*"Test-takers from eastern provinces, benefiting from relatively high-quality educational resources, may pass the English proficiency test without any preparation. However, test-takers from less developed provinces may struggle to pass the test, despite exerting considerable effort throughout their university studies."* (TT12)
	Adequacy and effectiveness of learning resources	*"College English courses have made me take English seriously. To pass the final exams of these courses, I focused on the improvement of my listening, speaking, reading, and writing skills. I also spend extra time studying English outside class."* (TT18)
Institutional factors	Institutional policy	*"The university's Undergraduate School and the Academic Affairs Office have provided policy support for the English Proficiency Test, incorporating relevant policies into the Handbook for Undergraduate Students. By reading relevant chapters of the Handbook upon enrollment, the undergraduates can learn about the English proficiency requirements set upon them and how to meet those requirements."* (TU1)

(*Continued*)

(Continued).

Code	Subcode	Example interview excerpt
Institutional factors	Local expertise in language testing and assessment	*"The language testing team at the university has conducted extensive research over the past years. I believe, with accumulated expertise, this team has the ability to design and develop a high-quality in-house English proficiency test."* (TU1)
	Infrastructure for administering the EPT	*"Two backup test rooms were provided to safeguard against potential equipment failures during each test administration."* (TA2)
Personal factors	Language proficiency	*"I believe that the goal of passing the English Proficiency Test is achievable. For me, the test exhibits a moderate level of difficulty."* (TT17)
	Beliefs about the EPT	*"I think the English Proficiency Test is fair. There is no predetermined pass rate, and whether I can pass the test depends solely on my test performance."* (TT15)

Notes. Each participant was assigned a unique code for ease of reference in reporting the results. Test takers are coded as TT1–TT20; teachers (also test developers) as TD1–TD6; test administrators as TA1–TA2; and test users as TU1–TU3.

Author Introduction and Research Achievements

Education background

Zhejiang University, Hangzhou, China Sept. 2018–June 2025

- M.A. & Ph.D. in Foreign Language and Literature

University of Bristol, Bristol, UK Oct. 2022–Oct.2023

- Visiting Ph.D. student

Zhejiang University, Hangzhou, China Sept. 2014–June 2018

- B.A. in English Language and Literature

Research achievements

何莲珍、**张娟**. “公平”的理论向度——兼论对语言测试公平性研究的启示[J]. 浙江大学学报(人文社会科学版), 2024, 54(5): 122–130.

何莲珍、**张娟**. 语言测试的公平性：内涵、公平观及研究启示[J]. 外语教学与研究, 2022, 54(1): 79–89.

张娟、何莲珍. 义务教育英语学业质量标准的解读与应用前景展望[J]. 中小学外语教学(中学篇), 2022, (11): 1–7.

Min, S., **Zhang, J.**, Li, Y., & He, L.* (2022). Bridging local needs and national standards: Use of standards-based individualized feedback of an in-house EFL listening test in China. *Language Testing*, *39*(3), 425–452. https://doi.org/10.1177/0265532221107-0990 (SSCI, A&HCI)

张娟、何莲珍. 语言测试的公平性：中国考试文化回眸[J]. 中国考试, 2022, (1): 45–52.

何莲珍、**张娟**. 《中国英语能力等级量表》在补偿式教学与学习中的应用[J]. 外语测试与教学, 2021, (3): 1–11.

何莲珍、**张娟**. 中国语言测试之源与流[J]. 浙江大学学报(人文社会科学版), 2019, 49(6): 29–38.