

ScholarWorks@GSU

Towards Including Second Language Varieties of English on High-stakes International Tests of English Proficiency: A Perceptual Adaptation Study

Authors	Tan, Yi
Citation	Tan, Yi. "Towards Including Second Language Varieties of English on High-stakes International Tests of English Proficiency: A Perceptual Adaptation Study." PhD diss., Georgia State University, 2023. https://doi.org/35520963 .
DOI	https://doi.org/10.57709/35520963
Download date	2026-04-12 00:35:11
Link to Item	https://hdl.handle.net/20.500.14694/446

Towards Including Second Language Varieties of English on High-stakes International Tests of
English Proficiency: A Perceptual Adaptation Study

by

Yi (Laura) Tan

Under the Direction of Sara Cushing, PhD

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

The present dissertation reports on a second language (L2) perceptual adaptation study, investigating whether or not the assistance of a 60-second exposure to an L2 speaker's accent improves comprehension of texts delivered by that speaker in the context of a simulated TOEFL iBT listening test. Two L2 speakers (one Turkish and one Ukrainian) recorded four TOEFL iBT listening scripts. A total of 317 Chinese undergraduate and graduate students each listened to four passages (two with the original TOEFL iBT voice actors and two with one of the two L2 speakers) and answered comprehension questions, alongside ratings of accent familiarity and speaker attitudes (across all conditions), and ratings of speaker comprehensibility and exposure efficacy (under conditions with an exposure). Participants were randomly assigned to one of three conditions for the passages featuring L2 speakers: audio-with-script exposure, audio-only exposure, and no exposure.

Many-Facet Rasch Measurement (MFRM) was used to compare passage-level listening comprehension scores and item-level response across four conditions (i.e., audio-with-script, audio-only, no exposure, and baseline), among L2 listeners at three proficiency levels (i.e., high, medium, and low), and between two passage types (i.e., monologic lectures and conversations)/three item types (i.e., main idea, explicit detail, and implicit detail). MFRM analyses revealed that the baseline condition was significantly easier than all experimental conditions across all listening passages, although bias detected between the baseline and the experimental conditions was particularly pronounced on lectures, Lecture 2 in particular, and test items associated with Lecture 2. In addition, a 60-second exposure was most useful for low-proficiency L2 listeners and also necessary for high-proficiency L2 listeners. Factorial ANOVA was employed to analyze L2 listeners' perceived efficacy of the 60-second exposure and their

attitudes towards speakers. Results showed that the 60-second exposure was perceived to be more useful for listening to L2 speakers participating in conversations than delivering lectures, and L2 listeners' attitudes were overwhelmingly negative towards L2 speakers, whose average favorability ratings were only half of those received by speakers with General American accents. Implications of the findings for test developers of high-stakes English proficiency tests and TESOL practitioners in a multilingual society were discussed.

INDEX WORDS: second language perceptual adaptation, second language listening assessment, high-stakes English proficiency test, fairness, authenticity, construct validity, English as a lingua franca, language attitudes

Copyright by
Yi Tan
2023

Towards Including Second Language Varieties of English on High-stakes International Tests of
English Proficiency: A Perceptual Adaptation Study

by

Yi Tan

Committee Chair: Sara Cushing

Committee: Stephanie Lindemann

Youjin Kim

Luke Harding

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

May 2023

DEDICATION

This dissertation is dedicated to my mom and dad, who first worked very hard to provide me with a life of comfort and then allowed me to give up that comfortable life so that I could find my own way in this world and grow into the woman I wanted to be, rather than the daughter they expected me to be.

ACKNOWLEDGEMENTS

One fun fact of my life is that nothing of significance has ever gone as planned. My Ph.D. is no exception, and naturally, so is my dissertation. Another fun fact of my life is that everything of significance has always turned out to be much better than planned, although the process is often times mind-crushing and nerve-wracking, if not soul-destroying. And this is exactly how my Ph.D. and dissertation have been playing out. I have been able to sit in my cozy and comfortable room and write down this acknowledgment with gratitude that is difficult to convey in words, only because of the tireless and unwavering support and kindness that I have received and felt from my committee, my friends, and my family.

I would like to start by thanking my advisor, Dr. Sara Cushing. Sara took me in, knowing that my main research interest lies in second language listening assessment, not writing assessment, the area she knows a great deal about, and when she only vaguely knew me as a person. Put simply, this was a highly risky decision on her part. But, just like that, she has been my rock on this roller-coaster ride. She helped with my conference proposals, grant proposals, and job application materials, in addition to everything involved in my dissertation. She also listened to my complaints and comforted me when COVID hit my hometown of Wuhan, when my data collection plan in China was likely ruined, and when I was deeply concerned if I would be able to finish my dissertation. Every step of the way, through my highs and lows, she was always there for me, making this arduous journey bearable. For these actions and support, I am tremendously grateful to Sara.

A massive thank you also goes to Dr. Stephanie Lindemann for being my honorary dissertation “co-chair”, for opening up the world of sociolinguistics to me, and for being my spiritual mentor since the beginning of this journey. I still remember the first time we met. It was

on my graduate program orientation day. Towards the end of the orientation, she shook hands with each new Ph.D. student, in part because she would be our general advisor before we would work closely with our academic advisor two years later. I remember thinking, “Good; she looks nice and approachable, though small in stature”. But then I constantly found myself not able to keep up with her rate of speech, witticism, and logic, feeling 1,000 steps behind her. Fortunately, I grew out of that phase long ago and I now feel connected with Stephanie in a way that I could with almost no others. I do not need to say much, and she would understand. Also, we would discuss things that most people might find hard to relate to. For example, how many of you have had the experience of finding a person who you know very well sound significantly more attractive when you only hear them without looking at them? I never thought I would talk about this with anyone, and yet Stephanie brought this up while we were discussing something not that relevant.

Due to her research interests, Dr. You Jin Kim was the only reason I applied for and came to this Ph.D. program, though I ‘betrayed’ her two years later when I got much more attracted by language testing. Yet, she has always been super understanding and supportive. Two occurrences with You Jin will always live in my mind. The first one happened during my first year in the program. We had a meeting about my final coursework paper, during which I told her that knowing she was usually in the office gave me a sense of security, as I could turn to her for help. A couple hours later, well after the typical work hours, I discovered that she was still in her office, so I asked her why and she responded, “because this makes you feel safe.” The second occurrence was actually very embarrassing for me, but I do not mind sharing this. It was during the pandemic and before I defended my proposal. The trigger was me being underpaid for two months in a row. You Jin was in the email thread about my paycheck because she was the

director of graduate studies. I clearly sounded rather infuriated and frustrated in my email. It was just too much stress building up for an extended period of time. She texted me right away, asking if I wanted to talk. During our virtual meeting, I was whining and crying, and she patiently listened and calmed me down. Regarding my dissertation, she is probably the only one whose expertise does not align with my chosen topic, and yet she helped me improve the study design with her abundant experience with quantitative research.

I never got a chance to talk with Dr. Luke Harding when I was a master's student at Lancaster University in England, because he was on sabbatical leave, and because I was not working on assessment at the time. Therefore, it is fair to say that his becoming a member of my dissertation committee almost a decade later was not even in my wildest dreams. It made my day when he responded to Sara's invitation to be on my committee. It fed my ego when he said my dissertation project was the kind of project he wished he could have done for his dissertation. It warmed my heart when he comforted me during my proposal defense because he was worried that the committee had asked too many questions, so that I might be left with the impression that my proposal was not strong enough.

In addition to my committee, I would like to thank all my participants, including raters, speakers, and listeners. Without them, this dissertation simply would not have come together. The participants were all very generous to take the time to work on the task(s) and to give me feedback in great detail, the listeners in particular. Some of my listeners were working on the tasks while their life was turned upside down because of the COVID lockdown policy in Shanghai. Very importantly, I must thank Sharon Cavusgil. We have gone from colleagues, to friends, and to adopted mom-and-daughter. She would call me immediately, and then leave cards and/or succulents on my porch, when I was in deep sorrow. I would also like to thank Dr. Ute

Römer for being very kind and patient with me, and sending positive vibes all the time. Many thanks to Dr. Daniel Dixon for helping me resolving the three-way ANOVA interaction issue in my data analysis that bothered me for months.

I would also like to thank my brilliant students in the two classes that I am teaching: IEP 0740: Oral Communication for Academic Purposes IV and AL 3031 Language in Society. Teaching two new courses (and having to go to the campus from Monday to Thursday) while trying to graduate and land a job is certainly bonkers and I will not recommend this to anyone. However, these two courses fit quite well into the theme of my dissertation. And spending time with them is a great break from my dissertation. I cannot tell you how happy I was to see my lovely students after being buried in my dissertation during the whole spring break. The discussions and activities I have had with my students have given me a lot to think as an applied linguist and a TESOL practitioner at the same time, the two groups who often time do not communicate with each other. I took a lot of inspiration from their perspectives and thoughts while drafting the discussion and conclusion chapters.

Turning now to my colleagues/friends, Eunice, Bineta, Sasha, Sanghee, Julie, Annalyn, Selahattin, Qian, Taylor, and Terry. Notice that I have used a slash, simply because they are colleagues *and* friends, a group of brilliant, kind, and funny souls whom I have met here. They are the only people in this whole universe who know exactly what I have gone through in the past six years. To me, this kind of friendship is no different than brotherhood among soldiers. This Ph.D. has never been easy for any of us, but I am very fortunate to have them as my company during ups and downs, to offer comfort and to be comforted. Thank you to my dear friends Yi (Theresa) Tao for being a big sister taking care of me even before I arrived in Atlanta,

and Gefei for exploring with me the natural beauty of Atlanta and unique groceries from Trader Joes.

Finally, and most importantly, I extend my deepest and most heartfelt gratitude to my mom, sister, and aunt. My mom is indisputably the most beautiful and toughest feminist mom in this world. I have inherited her intelligence and perseverance (not beauty, though). We rarely see each other eye to eye, and yet she has always respected and supported my choices. My sister is one of my best friends. Our long-distance video chats over the years as I have pursued my Ph.D. and my dissertation were a big part of maintaining my sanity. I am also very thankful to her for taking care of our mom while I am thousands of miles away. My aunt deserves a huge thanks. Growing up, I was always close to her, especially so after my dad passed away in the most unimaginably way possible six years ago. She seemed to take on the paternal role, checking on me regularly to see if I was doing well (including if I had enough money), and helping me with my data collection (which saved a great deal of money). My utmost thanks go to my dad, who was always genuinely so proud of any tweeny weeny achievements that my sister and I made. I have always felt that it was his wish for me to do my Ph.D. in the United States. And he has always been looking down on me from heaven, sorting me right out whenever I am in trouble. To give up or walk away from my Ph.D. was never an option, simply because I am doing this also for my dad, and I am willing to do whatever it takes to become a Dr. and to make him proud.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS		V
LIST OF TABLES		XV
LIST OF FIGURES		XVI
1 INTRODUCTION		1
2 LITERATURE REVIEW		4
2.1 The Ongoing Debate on the Inclusion of Second Language (L2) Varieties		4
2.1.1 Arguments from the Perspective of ELF		4
2.1.2 Arguments from the Perspectives of Test Authenticity, Construct Validity, and Communicative Competence		6
2.1.3 Arguments from the Perspective of Washback		8
2.1.4 Legitimate Concerns		9
2.2 Prior Endeavors		14
2.2.1 Ockey and French (2016) – The Strength of Accent Threshold		15
2.2.2 Kang et al., (2018) – The Intelligibility Threshold		18
2.2.3 Two Critical Questions Concerning the Two Thresholds		21
2.3 Insights from and Limitations in L1 Perceptual Adaptation Research		23
2.4 Insights from and Limitations in L2 Perceptual Adaptation Research		28
3 THE CURRENT STUDY		34
3.1 L2 Speakers/Accents		35
3.2 Length of Exposure		37

3.3	Exposure Conditions.....	38
3.4	Listener Proficiency	39
3.5	Passage Type.....	40
3.6	Item Type.....	42
4	METHODS	43
4.1	Recruiting and Initial Screening of L2 Speakers	43
4.2	Candidate Speakers and Audition Recordings	44
4.3	Raters, Rating Scale, and Speaker Audition Survey	45
4.3.1	<i>Raters</i>	<i>45</i>
4.3.2	<i>Revised Rating Scale</i>	<i>46</i>
4.3.3	<i>Speaker Audition Survey</i>	<i>47</i>
4.4	Selecting the Final Speakers	47
4.5	Listeners.....	48
4.6	Instruments.....	49
4.6.1	<i>Pre-test.....</i>	<i>50</i>
4.6.2	<i>Experimental Test</i>	<i>52</i>
4.7	Testing Procedure	60
4.8	Scoring, Coding and Analysis	61
4.8.1	<i>Test Item Scoring</i>	<i>61</i>
4.8.2	<i>Item Type Coding</i>	<i>61</i>

4.8.3	<i>Ratings on Accent Familiarity, Efficacy of Exposure Clip, and Attitudes Towards Speakers</i>	63
4.8.4	<i>Statistical Analysis</i>	63
5	RESULTS	74
5.1	Familiarity Ratings	74
5.2	Coefficient alpha	74
5.3	RQ1	75
5.3.1	<i>RQ 1.1</i>	75
5.3.2	<i>RQ 1.2</i>	95
5.4	RQ2	113
5.4.1	<i>Helped Adapt to Speaking Style</i>	114
5.4.2	<i>Helped Adapt to Accent</i>	116
5.4.3	<i>Made Me Less Anxious</i>	116
5.4.4	<i>Was Long Enough</i>	117
5.5	RQ3	117
5.5.1	<i>Attitudes Towards Two Accents</i>	118
5.5.2	<i>Attitudes Towards L2 Speakers</i>	120
6	DISCUSSION	123
6.1	RQ1	123
6.1.1	<i>Relative Difficulty of Exposure Conditions at Passage and Item Level</i>	123
6.1.2	<i>Relative Difficulty of Listening passage and Passage Type</i>	128
6.1.3	<i>Relative Difficulty of Test Item and Item Type</i>	129

6.1.4	<i>Rasch Interaction/Bias Analyses at Passage level</i>	131
6.1.5	<i>Rasch Interaction/Bias Analyses at Item Level</i>	133
6.2	RQ2.....	135
6.3	RQ3.....	136
7	CONCLUSIONS	138
7.1	Summary and Implications	138
7.2	Limitations and Future Research	144
	REFERENCES.....	147
	APPENDICES.....	155
	Appendix A	155
	<i>Appendix A.1 Speaker Audition Survey</i>	155
	<i>Appendix A.2 Pre-test</i>	155
	<i>Appendix A.3 Experimental Test</i>	155
	Appendix B	156
	<i>Appendix B.1 Pairwise Bias Report for Exposure Condition and Listening Passage</i>	156
	<i>Appendix B.2 Pairwise Bias Report for Exposure Condition and Proficiency Level at Passage Level</i> 158	
	<i>Appendix B.3 Pairwise Bias Report for Exposure Condition, Listening Passage, and Proficiency Level</i>	160
	<i>Appendix B.4 Pairwise Bias Report for Exposure Condition and Passage Type</i>	165
	<i>Appendix B.5 Pairwise Bias Report for Exposure Condition, Passage Type, and Proficiency Level</i> 166	

<i>Appendix B.6 Pairwise Bias Report for Exposure Condition and Test Item.....</i>	<i>169</i>
<i>Appendix B.7 Pairwise Bias Report for Exposure Condition and Listening Proficiency at Item Level</i>	<i>178</i>
<i>Appendix B.8 Pairwise Bias Report for Exposure Condition, Test Item, and Listening Proficiency.....</i>	<i>180</i>
<i>Appendix B.9 Pairwise Bias Report for Exposure Condition and Item Type.....</i>	<i>207</i>
<i>Appendix B.10 Pairwise Bias Report for Exposure Condition, Item Type, and Listening Proficiency.....</i>	<i>209</i>

LIST OF TABLES

<i>Table 2.1 Strength of Accent Scale (Ockey & French, 2016, p. 712)</i>	15
<i>Table 4.1 Revised Strength of Accent Scale</i>	46
<i>Table 4.2 Accent Rating of Candidate Speakers</i>	47
<i>Table 4.3 The Structure of MET Listening Section</i>	50
<i>Table 4.4 Details of Each Listening Comprehension Passage on the Experimental Test</i>	52
<i>Table 4.5 Structure of Experimental Test Under One Exposure Condition</i>	56
<i>Table 4.6 L2 Listeners by Groups Across Exposure Conditions</i>	56
<i>Table 4.7 L2 Listeners by Proficiency Levels and Exposure Conditions</i>	57
<i>Table 4.8 Test Items by Item Types and Listen Times</i>	62
<i>Table 4.9 Facets for MFRM Analysis on the Experimental Test Data</i>	68
<i>Table 5.1 L2 Listeners' Familiarity with Speakers' Accent (out of 7)</i>	74
<i>Table 5.2 Coefficient Alpha by Test Version</i>	75
<i>Table 5.3 Pairwise Bias Report for Exposure Condition and Listening Passage</i>	85
<i>Table 5.4 Pairwise Bias Report for Exposure Condition and Proficiency Level</i>	87
<i>Table 5.5 Pairwise Bias Report for Exposure Condition, Listening Passage, and Proficiency Level</i>	91
<i>Table 5.6 Pairwise Bias Report for Exposure Condition, Passage Type, and Listening Proficiency</i>	94
<i>Table 5.7 Pairwise Bias Report for Exposure Condition and Test Item</i>	103
<i>Table 5.8 Pairwise Bias Report for Exposure Condition and Listening Proficiency at Item Level</i>	105
<i>Table 5.9 Pairwise Bias Report for Exposure Condition, Test Item, and Listening Proficiency</i>	107
<i>Table 5.10 Pairwise Bias Report for Exposure Condition, Item Type, and Listening Proficiency</i>	112
<i>Table 5.11 Descriptive Statistics for L2 Listeners' Perception of the Efficacy of 60-second Exposure</i>	113
<i>Table 5.12 ANOVA Descriptive Statistics for L2 Listeners' Perception of the Efficacy of 60-second Exposure</i>	115
<i>Table 5.13 Descriptive Statistics for L2 Listeners' Attitudes Towards Six Speakers on the Experimental Test</i>	118
<i>Table 5.14 Descriptive Statistics for L2 listeners' Attitudes Towards Two Accents on the Experimental Test</i>	119
<i>Table 5.15 ANOVA Descriptive Statistics for L2 listeners' Attitudes Towards Two Accents on the Experimental Test</i>	119
<i>Table 5.16 Descriptive Statistics for L2 Listeners' Attitudes Towards L2 Speakers on the Experimental Test</i>	120
<i>Table 5.17 ANOVA Descriptive Statistics for L2 Listeners' Attitudes Towards L2 Speakers on the Experimental Test</i>	122

LIST OF FIGURES

<i>Figure 4.1 Structure of One Listening Passage on the Experimental Test</i>	58
<i>Figure 5.1 Wright Map of ID, Exposure Condition, and Listening Passage</i>	77
<i>Figure 5.2 Exposure Condition Measurement Report (Based on Listening Passages)</i>	80
<i>Figure 5.3 Listening Passage Measurement Report (Individual Listening Passage)</i>	81
<i>Figure 5.4 Listening Passage Measurement Report (Passage Type)</i>	81
<i>Figure 5.5 Bias Report for Exposure Condition and Listening Passage</i>	83
<i>Figure 5.6 Interaction between Exposure Condition and Listening Passage</i>	83
<i>Figure 5.7 Interaction between Exposure Condition and Proficiency Level at Passage Level ...</i>	87
<i>Figure 5.8 Interaction between exposure condition and proficiency level (based on observation average score)</i>	89
<i>Figure 5.9 Wright Map of ID, Exposure Condition and Test Item</i>	97
<i>Figure 5.10 Exposure Condition Measurement Report Based on Test Items</i>	98
<i>Figure 5.11 Test Item Measurement Report (Individual Item)</i>	100
<i>Figure 5.12 Test Item Measurement Report (Item Type)</i>	101

1 INTRODUCTION

The approach of using ‘standard’ accents (e.g., General American, Standard Southern British English) has been the gold standard in second language (L2) listening assessment for decades (Ockey & Wagner, 2018), and had its merit several generations ago when academia in English-speaking contexts was largely dominated by traditionally ‘native speakers’ of English (Hamid, 2014). However, with changing demographics, globalization, and the increasing use of English as a Lingua Franca (ELF), the sociolinguistic landscape of this target language use (TLU) domain has significantly changed in recent years (Jenkins, 2006; Kang et al., 2018), which in turn renders this gold standard increasingly unjustifiable. In response, the approach of using highly intelligible L2 speakers on the listening sections of high-stakes English proficiency assessments has been increasingly embraced by scholars and researchers (e.g., Harding, 2011; Kang et al., 2018), due to its potential to minimize test bias, to broaden the listening construct – multidialectal listening skills, and to accommodate practicality related issues.

However, the exclusion of less intelligible speakers on large-scale English listening proficiency tests may fall short of a full representation of accent strengths in the TLU domain and may not necessarily help to achieve a broader construct of listening comprehension, if the listening stimuli “have an L2 ‘flavor’ only” (Harding, 2012, p. 177). Going further, the notion of including only L2 speakers with high intelligibility and/or comprehensibility may not only neglect the fact that intelligibility has an interactional dimension with both speakers and listeners playing their roles (e.g., Smith & Nelson, 1985), but also the fact that intelligibility is dynamic, and first language (L1) listeners and L2 listeners alike, despite initial difficulties understanding unfamiliar L2 accents, are eventually able to adapt to these accents (e.g., Bradlow & Bent, 2008; Harding, 2018).

To make a case for including accents with lower intelligibility on examinations, however, more research is needed on how quickly adaptation to an unfamiliar accent takes place, and how adaptation is affected by the variables relating to the speaker and the listener: strength of accent in the former case, and L2 listening proficiency and general familiarity with the accent in the latter. To cast light on these issues, this dissertation project, as mostly inspired by perceptual adaptation research working with both L1 and L2 listeners, is intended as an L2 perceptual adaptation study conducted in the context of a simulated TOEFL iBT listening test. The overarching goal is to explore whether or not the assistance of a 60-second exposure to an L2 speaker's accent improves comprehension of texts delivered by that speaker. Using texts recorded by two speakers whose rated intelligibility is slightly lower than that recommended by Ockey and French (2016) and whose accents are generally unfamiliar to the targeted L2 listeners, I investigated: 1) how passage-level listening comprehension scores and item-level response vary across four exposure conditions (i.e., audio-with-script, audio-only, no exposure, and baseline), among L2 listeners at three different proficiency levels (i.e., high, medium and low), and between two passage types (i.e., academic monologue lectures and conversations)/ three item types (i.e., main idea, explicit detail, implicit detail); 2) L2 listeners' perceived efficacy of the 60-second exposure; and 3) L2 listeners' attitudes towards speakers.

This dissertation seeks to address these questions. In the remaining chapters I will detail my efforts to investigate the feasibility of including a broader range of accents in large-scale listening tests without compromising test designers' fundamental concern over construct validity and fairness. Chapter 2, Literature Review, presents the theoretical background and empirical evidence to the present research. Chapter 3 presents the research questions, followed by a brief description of the research design, and then detailed explanation of rationales behind each

variable involved in the study. Chapter 4 details the two sequential phases of data collection: Phase 1 speaker audition, and Phase 2 main experiment, followed by data scoring, coding and analysis. Chapter 5 provides the results of the three main research questions. Chapter 6 is devoted to a discussion of the findings with regards to the research questions. Chapter 7 concludes the dissertation. It begins with a brief summary of the findings, then discusses implications of the findings. Finally, limitations of the study are briefly discussed in the context of recommendations for future work on this topic.

As a note, the terms native speaker (NS) and standard language (SL) are separate constructs, and therefore it is important to untangle them (Isaacs & Rose, 2022, p.407). The terms *prestige*, *non-prestige*, and *L2* were used throughout this paper while acknowledging that they are not the most accurate terms to describe such varieties. When used in this paper, the term *prestige* accents refer to General American (GA) and/or Standard Southern British English (SSBE), the two long upheld spoken standards or prestige varieties. The term *non-prestige* refers to all the accents except GA and SSBE. *L2* accents refer to varieties spoken by individuals who did not grow up speaking English their first language and belong to the category of *non-prestige* accents.

2 LITERATURE REVIEW

This chapter offers a review of the literature concerning two major themes: (1) the benefits and challenges of incorporating second language (L2) varieties of English in listening assessment and (2) perceptual adaptation studies. It opens with the ongoing debate on the inclusion of L2 varieties on high-stakes listening proficiency tests, then moves to prior endeavors to diversify English varieties on L2 listening comprehension tests without unfairly advantaging or disadvantaging some test takers. A detailed review of L1 perceptual adaptation research follows. A detailed critique of the methodologies of these studies is then presented, demonstrating the research gap for the present work. This section following reviews and critiques of L2 perceptual adaptation research, and further gaps are identified.

2.1 The Ongoing Debate on the Inclusion of Second Language (L2) Varieties

2.1.1 *Arguments from the Perspective of ELF*

With the spread of English as a lingua franca (ELF), English is now spoken by many more people as an L2 or foreign language (FL) than as a native language (Crystal, 2000; Lowenberg, 2002; Llorca, 2004). This is probably especially the case for the contexts of higher education in English-speaking countries (e.g., Australia, Canada, Ireland, New Zealand, the UK, and the US). In other words, in this target language use (TLU) domain, for which high-stakes English proficiency tests are designed, L2 listeners are likely to encounter not only historically prestige accents, but also a wide range of non-prestige accents (Abeywickrama, 2013; Brown, 2014; Canagarajah, 2006; Harding, 2011; 2012; 2014). This then calls for critical examinations with regard to the traditional centrality of prestige varieties of English underpinning international tests of English proficiency (Davies et al., 2003). And it seems logical to argue that traditionally prestige varieties are far from being sufficient for international students to deal with the demands

of academic studies as well as daily communication in a social setting characterized by negotiating successful communication with people from diverse lingua-cultural background (Galloway & Numajiri, 2019). What these international students need; instead, is “multidialectal competence,” or more precisely, multidialectal listening skills, when they “shuttle between multilingual communities.” (Canagarajah, 2006, p. 233). It is worth pointing out that domestic students in the same TLU domain would also need such multidialectal competence, though considering the context of the current study – high-stakes English proficiency tests, the scope of the investigation was restricted to international students.

High-stakes assessment, on the other hand, has been slow or reluctant to embrace the fact that the most common use of English today is English as an lingua franca (Elder & Harding, 2008; Harding, 2012, Kang, et al., 2019), and therefore the changing demographics and sociolinguistic realities of test takers’ ultimate language use in this TLU domain have not yet been reflected in current testing practices (Abeywickrama, 2013; Elder & Harding, 2008; Harding, 2012). It is true that some high-stakes English language proficiency tests (e.g., IELTS, TOEFL iBT, and TOEIC) are taking strides to be more inclusive and authentic by integrating non-GA and non-SSBE accents, such as Australian, Canadian, and New Zealand varieties, in recent years (Harding, 2012). However, these attempts, as Kang et al., (2019) pointed out, are still problematic in the sense that the newly incorporated accents are still traditionally native varieties. That is to say, “in most cases, the variety of English used has been, and still is, a native variety.” (Abeywickrama, 2013, p. 60). In the same vein, Harding and McNamara (2018) noted that despite a recent trend in assessment moving away from inner-circle language models by integrating a wider variety of accents into listening input, to date, this has mostly been limited to inner-circle varieties in standardized tests. Together, it appears that test developers of

standardized English proficiency tests still implicitly assume that L2 speakers will only interact with NSs in the TLU domain, which raises the question of whether such tests are “truly international and measure the criterion of global communicative behavior” (Isaacs & Rose, 2022, p. 408).

2.1.2 Arguments from the Perspectives of Test Authenticity, Construct Validity, and Communicative Competence

Authenticity refers to “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task” (Bachman & Palmer, 1996, p. 23), and provides “a means for investigating the extent to which score interpretations generalize beyond performance on the test to language use in the TLU, or to other similar non-test language use domain” (Bachman & Palmer, 1996, p. 24). Given that investigating the generalizability of score interpretations is a crucial part of construct validation, authenticity is thus linked to construct validity. Crudely speaking, the relationship between authenticity and construct validity is that the more authentic the test task, the more confident one can be about generalizing test results to test takers’ ability beyond the testing context, and the more valid are the inferences made about test takers’ ability based on their test performance (Wagner, 2016). This is likely the reason why it is authenticity that drives much of the current interest in integrating a variety of L2 accents into high-stakes listening proficiency tests (Harding, 2011).

To put this into perspective, accent diversity, as previously discussed, is one of the key features existing in the context of English-medium higher education or the TLU domain. As such, the practice of adhering to prestige varieties in creation of listening input used in high-stakes listening comprehension tests could be seen as lacking authenticity, which leads to an underrepresentation of English varieties found in academic domain which these high-stakes tests

have explicitly claimed to sample from (Elder & Harding, 2008). Such construct underrepresentation is described by Messick (1996) as one of the two major threats to construct validity (with other threat being construct irrelevant variance). Put simply, if an L2 listening test uses only traditionally prestigious varieties, its test score might be able to allow for valid inferences made about a test taker's ability to understand these varieties in a non-test context, but inferences made about that same test taker's ability to understand other varieties in the real world, are more suspect.

Hence, for high-stakes English proficiency tests to strengthen their construct validity, “acknowledging that L2 listeners will encounter and need to deal with a range of accents by including different varieties of English into their listening sections” would be the path to choose (Kang et al., 2018, p. 2). What should also be acknowledged, however, is the impossibility of replicating TLU tasks with “true verisimilitude” in a testing situation, or achieving situational authenticity, referring to “the extent to which the test tasks are perceived to share the characteristics of the target-language use tasks” (Buck, 2001, p. 106). This is simply because a test is a test and is different from the real-world situation (Buck, 2001). Nonetheless, this does not mean that the hands of test developers of high-stakes English proficiency are tied; rather, they can construct test tasks that have interactional authenticity, which is the second type of authenticity in Bachman's (1990) terms, defined as “a function of the interaction between the test taker and the test task” (Bachman, 1990, p. 317). This term was later renamed interactiveness in Bachman and Palmer (1996), defined as “ways in which the test taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task” (p. 25). Buck (2001) argued that it is interactiveness that “gets to the hearts of construct validity” (p. 126). That is to say, what really matters and can bolster the construct

validity of a test is that the interaction between the test-taker and the test task is similar to the interaction between the language user and the task in the TLU domain (Buck, 2001, p. 108).

Communicative competence has replaced the rigid adherence to ‘standard’ English norms in many testing contexts nowadays, high-stakes tests included (Elder & Harding, 2008; Wagner, 2016). For example, the purpose of TOEFL iBT test is “to measure the communicative language ability of people whose first language is not English . . . in situations and tasks reflective of university life.” (Jamieson et al., 2000, p. 10). In one of the most influential models of communicative competence – Bachman’s components of language competence--“sensitivity to dialect/variety” is directly related to the sub-competency “sociolinguistic competence” (Bachman, 1990, p. 87). That is to say, coping with, or adapting to, multiple accents is an important aspect of communicative competence in listening (Harding, 2011).

In sum, to make sure that test takers are prepared for the heterogeneity of accent varieties that characterizes the English-medium higher education and are equipped with essential multidialectal listening ability, an argument can be made for the inclusion of a diversity of L2 accents on listening assessments that are specifically designed to determine the extent to which test takers will be able to communicate in such a multidialectal TLU domain (Harding, 2011; Ockey & French, 2016).

2.1.3 Arguments from the Perspective of Washback

Washback, or consequential validity, or test consequences, refers to “the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning.” (Messick, 1996, p. 241). Ideally, introducing new tests to education systems or making changes to existing tests would be automatically followed by changes in teaching and learning. Evidence from previous studies

dedicated to this topic over the past decades, however, shows that such a simple linear relationship between teaching and testing may not be the reality (Alderson & Wall, 1993), and therefore it might be even “naive to hope that a substantial overhaul of English language testing would bring about the desired changes in teaching/learning attitudes and practice.” (Taylor, 2006, p. 54).

Nonetheless, many schools and teachers feel the pressure of tests (Fulcher, 2010). A case discussed in Harding’s (2011) concluding remarks may show such positive washback in action, presumably as a direct result of test pressure. That is, in response to the introduction of British and Australian accents on the listening section of TOEFL iBT, King George International College (a language school in Canada) modified its listening curriculum by adding British and Australian accents into their listening teaching.

Washback may well go beyond the realm of language learning and teaching, for it also pertains to linguistic and social justice. Still adhering to traditionally prestige varieties or not recognizing L2 varieties of English not only rejects the sociolinguistic reality of English but also creates a linguistic hierarchy in which L2 varieties are deemed as lacking legitimacy (Shohamy, 2006), which, in turn, may contribute to negative attitudes towards international professors, teaching assistants (ITAs), and students with L2 accents in higher education contexts in English-speaking countries (see Lindemann & Campbell, 2018 for a review). Nevertheless, language tests are intended to bring about beneficial consequences to our society (Messick, 1989).

2.1.4 Legitimate Concerns

The discussion so far has centered on the significance of incorporating a range of L2 accents into the assessment of L2 listening, which might lead to an overall impression of “conservatism within the testing community” (Elder & Harding, 2008, p. 34.3). This, however, is

not “driven by blindness on the part of language testers to the presence of such varieties” (Elder & Harding, 2008, p. 34.3), but rather by legitimate concerns with regard to fairness (test bias, in particular) and practicality (Harding, 2011; 2012; Kang et al., 2018; Taylor, 2006; Taylor & Geranpayeh, 2011).

Prominent among these concerns is test bias, “defined as existing when candidates of equal ability, but from different groups, have an unequal chance of getting an item correct, or of attaining the same test score”, due to a construct irrelevant factor (Harding, 2012, p.164). With relation to a listening test featuring L2 speakers, the concern for test bias relates to a shared-first language (shared-L1) effect (i.e., test takers may be advantaged when listening to L2 speakers of English who share their L1) and a familiarity effect (i.e., test takers may benefit from listening to L2 speakers of English to whom they have previously been exposed). A shared L1 effect, in essence, can be explained by familiarity with a specific accent or variety (Ockey & Wagner, 2018), and attributed to shared knowledge about the phonetics and phonology of the listener’s L1, alongside long-term exposure, and thus adaptation to that specific accent (Eger & Reinisch, 2019).

While findings from speech-processing and speech perception research have repeatedly shown that familiarity with a particular accent facilitates the processing and understanding of that type of accent as opposed to an unfamiliar one (Winke et al., 2013), the overall results from the line of research examining the effects of familiarity (with the focus on the investigation of the potential for shared-L1 advantage) on L2 listening comprehension was “not a clear yes or no, but sometimes”, as aptly summarized by Major et al., (2002, p. 185) two decades ago. That is, in certain circumstances, L2 listeners indeed perform better on a listening test featuring a speaker sharing their L1 background, but such parallel shared-L1 advantage does not hold for L2

listeners from other L1 backgrounds within the same study. Such inconclusive results have been repeatedly reported by prior studies looking into the nature of speaker-listener interaction and how this influences L2 listening comprehension (e.g., Abeywickrama, 2013; Bent & Bradlow, 2003; Kang et al., 2019; Major et al., 2002; Munro et al., 2006), independently of the listening tasks involved (e.g., transcription tasks, picture recognition, monologic lectures on retired TOEFL listening test), the accent strength or intelligibility level of the speakers (low, intermediate, high), and the sample size of the listeners involved (ranging from 40 to over 200).

It is important to note, however, that there is a critical methodological limitation existing in the above-mentioned studies, as pointed out by Harding (2012); that is, these studies only approach the issue of test bias indirectly by investigating the effect of different speakers within each L1 listener group (within-group differences). Recall the aforementioned definition of test bias, the issue of test bias should be examined by comparing between-group differences (i.e., shared-L1 listeners versus other L1 listeners on the same test featuring one L2 speaker). Another methodological limitation is attributed to the fact that all these studies relied solely on total or passage-level listening comprehension scores, leaving how L2 speech contributes to differential performance at the item level largely unknown (Shin et al., 2021).

In response to this, Harding (2012) drew on differential item functioning (DIF), a common approach to detect test bias in the language testing literature, to investigate the potential for a shared-L1 effect on an academic English test featuring speakers with Mandarin Chinese and Japanese accents. He found that while a shared-L1 advantage was relatively strongly supported for the L1 Chinese listeners ($n = 70$), but much less so for the L1 Japanese listeners ($n = 60$). Such a finding reflects many of the previous studies in which conflicting evidence is found for a shared-L1 advantage. Another interesting pattern observed is that test items which are more

likely to be flagged for DIF are the ones that require correct retrieval of specific words directly from the texts and hence tap into bottom-up listening skills. But this observation needs direct evidence, given that Mandarin and Japanese speakers in Harding's study delivered different texts so that the effects of "textual differences" and "accent differences" might have been intermingled (Shin et al., 2021, p. 583). This leads us to Shin et al., (2021), a very recent study that has made a significant contribution to the understanding of shared-L1 effects on L2 listening comprehension.

As with Harding (2012), Shin et al. (2021) also reported on an investigation of the potential for a shared-L1 benefit on L2 academic listening comprehension test at the item level using DIF analyses. What makes this study unique, however, lies in the fact that Shin et al. pursued this investigation by simultaneously taking into account speaker variables (i.e., L1 background and different levels of intelligibility, comprehensibility, and accentedness of L2 speech), listener variables (i.e., L1 background, familiarity with various English speech varieties), and listening text and item difficulty. Specifically, a total of 386 international undergraduate and graduate students (137 Indian, 149 Chinese, and 100 Korean) enrolled at a large Midwest university in the US participated as listeners in this study. Their average self-reported TOEFL iBT total score was 94.73 out of 120 ($SD = 24.49$) and their average TOEFL iBT listening score was 24.19 out of 30 ($SD = 5.10$). On the whole, this group of listeners could be considered advanced L2 listeners, according to the proficiency levels suggested by TOEFL iBT. They each took two listening tests. The first test served as an independent measure of listeners' English listening proficiency, the total score of which was used as an external matching criterion for DIF analyses. The second test was the experimental listening test, where listeners were assigned to one of the three versions comprising counterbalanced sets of American English, Indian-, and Chinese-accented lectures. Two types of DIF analyses were conducted: DIF within

speaker group (i.e., of the three L1 groups who listened to the same speaker, did any L1-listener group have an advantage?) and DIF within listener group (i.e., among all the listeners from the same L1, was any particular speaker easier or harder to understand?). The overall result showed that the shared-L1 effect was minimal. Within-speaker DIF analyses revealed that only seven items among a total of 36 (19%) were flagged as DIF items favoring Indian listeners, in comparison to two item (5.5%) exhibiting DIF in favor of Chinese listeners. Also, a shared-L1 effect was consistently associated with detail-oriented test items, with only one exception. Within-listener analyses detected an even smaller number of DIF cases, only two items, both of which were detailed-oriented test items. Another consistent trend was that all three listening groups attained the highest scores on lectures recorded by L1 American English speakers. These findings aligned with previous findings (e.g., Harding, 2012; Kang et al., 2019; Major et al., 2002), showing that not all L2 listeners benefited from or achieved better performance when listening to speakers who share their L1 background. By contrast, one finding contradicted prior studies, that is, some DIF items were associated with a more intelligible, more comprehensible and less accented L1 Hindi speaker. The explanation provided by the researchers is that the intelligibility, comprehensibility and accentedness among the L2 speakers involved in the study were not noticeably different from each other, L1 Hindi speakers, in particular.

To summarize, despite contradictory evidence for a shared-L1 benefit reported in the preceding studies, L2 listeners, in some circumstances at least, did perform better on a listening test featuring a speaker from their own L1 background. Such a shared-L1/ familiarity advantage would cancel itself out on a large-scale international exam, provided that a large number of L2 speakers could be used in this listening test, so that some test takers would be familiar with some accents while other test takers would be familiar with others (Ockey & Wagner 2018).

Unfortunately, due to logistical constraints, it is nearly impossible to take all test takers' linguistic backgrounds into consideration by including all English varieties that may be encountered in the TLU on high-stakes listening tests. Because of such impracticality, test developers would need to sample broadly from a large number of accents existing in the TLU domain, and this may lead to a selection of a handful of L2 varieties that are most frequently encountered in the TLU domain so as to approximate what listeners might experience in the real world (Elder & Harding, 2008; Ockey & Wagner, 2018). But this is precisely where test bias would creep in, and if existing in a systematic manner, it would further undermine the construct validity of listening scores elicited from such tests (Ockey & French, 2016).

Thus, it is fair to say that academic listening test designers, especially providers of large-scale high-stakes English proficiency tests, are facing a dilemma. That is, on the one hand, fairness is paramount (Taylor, 2006); on the other hand, authenticity, or integration of a diversity of accents (i.e., both 'standard' varieties and a whole range of L2 varieties spoken by international professors, teaching assistants and students), is desirable, for this is exactly what real-world L2 listeners would encounter in the TLU domain (Gu & So, 2015; Taylor & Geranpayeh, 2011). In response to this dilemma, Ockey and French (2016) and Kang et al., (2018) endeavored to identify some forms of threshold at which differences in speech varieties would negatively impact L2 listening comprehension, in the hope that such a threshold could provide some insights to test developers who desire to diversify English varieties on L2 listening comprehension tests without unfairly disadvantaging some test takers.

2.2 Prior Endeavors

Before delving into these two pioneer studies, it is necessary to clarify the three related and yet partially independent terms: accentedness, comprehensibility and intelligibility.

Following the constructs defined by Munro and Derwing (1995) in pronunciation literature, accentedness refers to “how strong the talker’s foreign accent is perceived to be,” comprehensibility refers to “listeners’ perceptions of difficulty in understanding particular utterances”, and intelligibility refers to “the extent to which an utterance is actually understood.” (Munro & Derwing, 1995, p. 291), which is often assessed by verbatim transcriptions of words, sentences, or longer units. Although it is reasonable to assume that as the strength of accent of a speaker increases, the comprehensibility and intelligibility of their speech decrease, this is, in fact, not always the case, for utterances rated as moderately or even heavily accented in Munro and Derwing (1995b) turned out to be perfectly intelligible and highly comprehensible.

2.2.1 Ockey and French (2016) – The Strength of Accent Threshold

The first pioneers are Ockey and French. In their 2016 study, they developed a Strength of Accent Scale (see Table 2.1). Accent was defined as “the degree to which an individual’s speech patterns are perceived to be different from the local variety, and how much this difference is perceived to impact comprehension of listeners who are familiar with the local variety” (Ockey & French, 2016, p. 695). Local variety in their definition refers to the ‘standard’ American English. As such, Ockey et al., (2016) argued that this definition takes into account both the speech variety of the speaker and that of the listener when making a judgement on the strength of accent. However, some may disagree, for this definition, in fact, conflates accentedness and comprehensibility.

Table 2.1 *Strength of Accent Scale (Ockey & French, 2016, p. 712)*

1	The speaker’s accent was NOT noticeably different than what I am used to and did NOT require me to concentrate on listening any more than usual. The accent did NOT decrease my understanding.
---	---

2	The speaker's accent was noticeably different than what I am used to but did NOT require me to concentrate on listening any more than usual. The accent did NOT decrease my understanding.
3	The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. However, the accent did NOT decrease my understanding.
4	The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. The accent slightly decreased my understanding.
5	The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. The accent substantially decreased my understanding.

To test the usefulness of this scale, Ockey and French (2016) conducted a large-scale empirical study. Specifically, twenty adult speakers were recruited, including two American, nine Australian, and nine British. These speakers were guided to record reading one of two academic scripts for TOEFL iBT listening comprehension section input (about five minutes in length). Using these recordings, two 20-second clips were created for each speaker. These clips were then subjected to judgement of 100 listeners using the Strength of Accent Scale. These 100 listeners, including both native and nonnative English speakers, were students and instructors from three US institutions, and they all resided in the US at the time of the study. All nonnative listeners were at advanced proficiency level. Accent ratings showed that all speakers were rated between '1' and '3' on the scale, indicating a narrow range of accent strength. Based on these results, nine out of 20 speakers were selected for the main experiment, with one speaking the local variety (i.e., 'standard' American English) and the other eight representing accent levels ranging from '1.7' to '2.7' on the scale.

These nine speakers further recorded one monologic lecture (686 words in length) as the listening stimulus to be tested on a real TOEFL iBT test. As such, participants involved were real-world TOEFL iBT test takers (N = 21,726) from 148 countries who took TOEFL iBT on

two consecutive weekends. They were randomly assigned to one of the nine conditions. The overall results revealed a negative relationship between strength of accent and L2 listening comprehension. Specifically, test takers who were assigned to speakers with accents of ‘2.1’ and/or stronger scored significantly lower than those assigned to the US speaker. On the other hand, test takers who listened to accents with strengths of ‘2’ and/or weaker did not obtain scores that were significantly different from those who listened to the US speaker. In light of the findings, the threshold at which the strength of accent negatively affects L2 listening comprehension was established as ‘2’ on the Strength of Accent Scale. In other words, accents with strengths below ‘2’ are not likely to have significantly negative impact on listening comprehension scores, whereas accents with strengths stronger than ‘2’ are likely to contribute to lower listening scores. Ockey et al., (2016) carried out a follow-up study with the same group of test takers but used an interactive lecture. Their results pointed to a slightly lower threshold – ‘2.2’ on the scale. Such a discrepancy, though small, was in line with some previous empirical studies in which interactive discourse was found generally easier to comprehend than monologic discourse (Ockey et al., 2016).

It is important to keep in mind that Ockey and French’s (2016) study remains narrow in focus, including only inner circle native-speaker varieties only (i.e., American, Australian, and British). Restricting inputs to traditional native-speaker varieties, as the researchers argued, would provide an approach that may lead to “a gradual broadening of the listening construct” – multidialectal listening ability (p. 711), which could not only allow the inclusion of a diversity of mild accents (i.e., below ‘2’ or ‘2.2’ on the Strength of Accent Scale, depending on the nature of discourse) without jeopardizing test fairness, but also bring about positive washback in the sense

that test takers would make efforts to develop multidialectal listening ability to achieve higher scores on the test.

2.2.2 Kang et al., (2018) – *The Intelligibility Threshold*

To go beyond English varieties exclusively from English-majority countries and thus to incorporate a wider range of varieties into high-stakes test of English listening proficiency, Kang et al., (2018) sought to establish an intelligibility threshold for English varieties that are equally intelligible to all test takers, regardless of their prior exposure. They did so by identifying phonological features on the basis of six English varieties represented by 18 speakers, including traditionally ‘standard’, ‘non-standard’, and ‘non-native’/L2 varieties. Before moving on to the details of this study, it is worth pointing out that in their study intelligibility was defined narrowly in phonological and perceptual terms, and an intelligibility threshold was operationalized in the context of a TOEFL-type monologic listening comprehension test.

The first step of this study concerned speaker audition. Initially, candidate speakers, who recorded TOEFL listening comprehension test passages, were evaluated by researchers themselves following Major et al.’s (2002) recommendations to ensure that these speakers sounded like professionals in the fields represented by TOEFL listening passages. The same recordings were also rated by eight expert raters, who were either graduates in Applied Linguistics or with a background in phonology and pronunciation, using a five-point comprehensibility scale (1 = ‘easy to understand’ to 5 = ‘difficult to understand’). These two procedures resulted in a list of 18 test speakers, with three from each of six distinctive groups, namely ‘standard’ English accents (i.e., General American and Received Pronunciation); ‘non-standard’ English accents spoken in contexts where English is an official language, but not necessarily the L1 of its speakers (i.e., Indian Hindi and South African Afrikaans); and EFL

speakers where English is not an official language, but used for international communication (i.e., Chinese Mandarin and Mexican Spanish). Speakers from conventionally non-standard English varieties were selected to represent a range of perceived comprehensibility, from high (i.e., 1- 2 out of 5), to mid (3 out of 5) and to low (4 - 5 out of 5).

Following this, these 18 speakers recorded two separate materials for the main study: one listening input passage (3 to 5 minutes in length) from TOEFL listening texts of monologic lectures (which was randomly assigned to each speaker) and 90 nonsense sentences. These recordings were then rated by 48 novice listeners on both strength of accent and comprehensibility, using two seven-point Likert scales (1 = 'no accent' to 7 = 'heavy accent'; 1 = 'easy to understand' to 7 = 'difficult to understand'). These novice raters consisted of undergraduate students (19), graduate students (21), and teachers (8). Their average evaluation on strength of accent for three speakers in each country was: 1.21, 1.27, and 1.33 for American speakers; 2.33, 2.56, and 2.62 for British speakers; 3.44, 4.29, and 5.11 for Indian speakers; 3.68, 5.27, and 5.53 for South African speakers; 2.04, 4.13, and 5.67 for Chinese speakers, and 3.60, 5.60, 5.84 for Mexican speakers. As for comprehensibility, while it was not explained why the expert and novice listeners were provided with different scales (five- versus seven-point, respectively), the ratings between these two groups were consistent as reported.

Moving on to the second step, the recorded speech of each speaker (including both TOEFL monologic lectures and nonsense sentences) was further subjected to phonetic and phonological analyses, including segmental, prosodic, and fluency features, for the reason that these features have been found to be strongly correlated with communicative success between native and non-native English speakers (e.g., Kormos & Dénes, 2004; Pickering, 2001). Specifically, segmental features covered vowels and consonants; prosodic features included

stress, intonation, and rhythm features; fluency features involved articulation rate, mean length of run, pauses, and hesitation markers. Using General American as a base for comparison, a trained phonetician identified all divergences in the pronunciation of other varieties of English. At the same time, sixty undergraduate and graduate students who were congruent with the same six countries (10 from each country) represented in the listening materials participated in the main study. These students completed the TOEFL listening comprehension test (18 passages in total) and an intelligibility test, which consisted of transcription of four missing content words in semantically nonsensical sentences, 90 sentences in total.

Finally, a speaker's intelligibility threshold was established on the pronunciation features of the 10 most intelligible speakers (i.e., three General American speakers, three Received Pronunciation speakers and one speaker representing each of the South African, Indian, Spanish, and Chinese accents), who were chosen based on listeners' comprehension scores which showed no significant difference among them. Specifically, in terms of content words, the speech of highly intelligible speakers rarely exhibited consonant deletion, inappropriate syllable reduction, and divergence in the pronunciation of consonant clusters relative to 'standard' American accent, but could show certain segmental divergence involving high functional-load vowels, low functional-load vowels, and high functional-load consonants, provided that the instance of such divergence is minimal. Segmental divergence in function words, however, was found to be less deleterious to listeners and thus could be discounted. Interestingly, speakers selected on the basis of these guidelines, as the researchers themselves commented, are comparable to speakers who would show very weak strength of accent in Ockey and French (2016). That is to say, despite all the differences between these two studies, they have come to almost the same conclusion. That

is, only speakers with high intelligibility or showing very weak strength of accent would fit the bill for high-stakes listening proficiency tests, irrespective of their English nativeness.

As a note, while different terms – comprehensibility and intelligibility were used in these two studies, they are collapsed and referred to as I/C in the following sections in order to emphasize a shared attribute of these concepts, that is, “a concern with successful message reception” (Lindemann & Subtirelu, 2013, p. 575).

2.2.3 Two Critical Questions Concerning the Two Thresholds

There is no doubt that the two thresholds, or more precisely, the two ways of determining whether a speaker is suitable for being a speaker on international English language proficiency tests, reviewed above would provide useful guidance on speaker selection for listening tests in general, and high stakes listening tests in particular, provided that the ability to adjust to unfamiliar accents is decided to be construct relevant for that test. One critical issue, though, especially in the case of L2 speaker selection, is that only the ones with high I/C would pass either of the two thresholds and to be selected. The challenges with limiting L2 speakers to those who are highly intelligible or comprehensible reside in achieving a fuller representation of accent strengths in the TLU domain (i.e., higher education) and a broader construct of listening comprehension (i.e., multidialectal listening skills). There is thus a need to find a responsible and feasible way to also include L2 speakers of relatively lower I/C so as to better reflect the TLU and to further broaden the listening construct without reducing the scores received.

More importantly, the notion of establishing high I/C thresholds is likely grounded on the assumptions that I/C is an “inherent feature” of speakers (particularly L2 speakers) or speech itself (Lindemann & Subtirelu, 2013, p. 575), and that I/C is “a purely fixed product” (Rajadurai, 2007, p. 95). In regard to the former assumption, it is indeed true that these pioneer researchers

have gone to great lengths to ensure the objectivity of process of establishing the thresholds or selecting speakers. To start with, both studies included “unsophisticated listeners” whose evaluation was deemed “especially important because they may provide insight into how understandable L2 speakers are when they interact with other members of their community.” (Munro, 2008, p. 200). Further, these novice listeners were heterogeneous, in terms of L1 backgrounds, areas of study, status (i.e., instructors, graduate students, and undergraduate students), and possibly even overall experience in communicating with L2 speakers (an assumption based on their area of study and status). Such employment of multiple listener groups would to a large extent avoid the native-speaker-centric approach commonly used in intelligibility studies (Rajadurai, 2007). Lastly, inter-rater reliability indicating fair internal consistency was reported in both studies.

However, as long as evaluation of I/C is dependent on human judgement, it is not immune to such listener-related factors as attitudinal variables (e.g., attitudes, expectations, stereotypes, and beliefs) and familiarity with L2 varieties (Kang, et al., 2018; Lindemann & Subtirelu, 2013; Munro, 2008; Rajadurai, 2007; Rajagopalan, 2010). Some empirical evidence supports the notion that positive attitudes towards and familiarity with a particular speaker or variety of English may play a facilitating role in I/C, while negative attitudes and unfamiliarity may act as a barrier to I/C (Rajadurai, 2007). In this sense, “intelligibility is not speaker- or listener-centered but is interactional between speaker and listener” (Smith & Nelson, 1985, p. 333). Furthermore, the mere presence of acceptable or high inter-rater reliability may not necessarily guarantee the validity of I/C measurements, be it relatively subjective instruments (e.g., a Likert scale) or objective instruments (e.g., a transcription task), in the sense that the high consistency may be attributed to “shared biases that exist within a population” rather than “a

shared objective assessment of speech,” as pointed out by Lindemann and Subtirelu (2013, p. 576). Rajadurai (2007) echoed this, arguing that “speech ratings are probably more indicative of subjective attitudes and prejudices than they are objective measures of intelligibility and comprehensibility.” (p. 92). Together, this draws into question whether there exists such a thing as ‘universal intelligibility’, or perhaps the real question is: to whom an accent or speech is intelligible or comprehensible?’ (Lindemann & Subtirelu, 2013; Rajagopalan, 2010).

As for the latter assumption, far from being a static notion, I/C is, in fact, dynamic. This has been empirically supported by speech-processing and speech perception research, which has repeatedly shown that despite initial difficulties in understanding unfamiliar L2 accents, L1 listeners eventually are able to adapt to these accents (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Gass & Varonis, 1984; Reinisch & Weber, 2012; Sidaras et al., 2009). But due to the dearth of such studies working with L2 listeners, “how unfamiliar the speech variety can be and how long it takes to adapt” are not yet known (Ockey & Wagner, 2018, p. 79). Thus, one more question remains: at which point during the course of listening may an accent or speech become intelligible or comprehensible? To answer these two questions, we could use some insights from L1 adaptation research. Before discussing this line of research, it is critical to point out that while the processes of L1 listening and L2 listening are similar, or at least not fundamentally different from each other, as suggested by existing work (Buck, 2001, p. 48), it is reasonable to suppose that it usually takes L2 listeners longer to adjust to a new or an unfamiliar accent than L1 listeners (Buck, 2001, p. 35).

2.3 Insights from and Limitations in L1 Perceptual Adaptation Research

The overall finding from L1 adaption studies demonstrates that L1 listeners, though they experience initial difficulties understanding foreign accents, are eventually able to adapt to

foreign accents of varying degrees of intelligibility. There are two ways allowing adaptation to occur: through ‘natural’ experience with previously unfamiliar accents (e.g., Sumner & Samuel, 2009; Witteman et al., 2013) and through short-term exposure under experimental lab conditions (e.g., Bradlow & Bent, 2008; Witteman et al., 2013), although it is the latter (i.e., short-term exposure) that is of particular relevance and interest to the current study (and possibly test developers) in that listening tests in general could only afford a rather limited exposure to assist test takers to ‘tune in’ to new accent(s) or voice(s), due to the concern over practicality (impractical to prolong a listening test). The key findings relevant to the current study in the existing literature are as follows.

To begin with, exposure to less than one minute was shown to be sufficient for listeners to overcome the initial decrease in processing speed with a foreign accent versus speech without one in Clarke and Garrett (2004). This rapid adaptation was measured through a cross-model word verification task, in which participants were asked to make a judgement whether a visually presented word matches the final word of a previous, auditorily presented sentence. In addition to adaptation to segmental variations (see Samuel and Kraljic, 2009 for an overview), such rapid adaptation also takes place at suprasegmental level. For instance, Reinisch and Weber (2012) demonstrated that a short exposure (about 2.3 minutes) allowed Dutch listeners to tune into lexical stress errors in Hungarian-accented speech in a word recognition task. Then, three types of adaptation were found in previous studies, that is, talker-dependent adaptation (i.e., adapting to the same talker in the training stage) (e.g., Clarke & Garrett, 2004; Gass & Varonis 1984), talker-independent adaptation (i.e., adapting to a novel talker who shares the L1 of the talker in the training stage) (e.g., Bradlow & Bent, 2008; Sidaras et al., 2009; Witteman et al., 2013; Xie et al., 2018), and accent-independent adaptation (i.e., adapting to a novel talker with a novel

accent) (e.g., Baese-Berk et al., 2013). Encouraging as these results are, it is crucial to consider their relevance to L2 testing contexts. In particular, speaker-independent adaptation may run directly against the prominent concern over test bias – a shared-L1/ familiarity advantage, in the sense that on an ideal high-stakes listening test it is a variety of different L2 accents rather than different L2 speakers from the same L1 background that should be opted for in order to avoid shared-L1/ familiarity advantage.

As for accent-independent adaptation found in Baese-Berk et al., (2013), the exposure stage was operationalized as an inclusion of speakers from five language backgrounds; more importantly, none of the 30 monolingual English listeners reported prior learning of the native languages of the speakers used in either the training or the test stage. These participants were subjected to further screening to ensure that they did not have significant exposure to foreign accents in general. Because of such homogeneity in terms of accent familiarity, or more precisely, unfamiliarity in this case, the researchers were able to draw a confident conclusion that the accent-independent adaptation found in their experiment was attributed to the exposure that those listeners received during the perceptual training stage. In the context of high-stakes testing, on the other hand, homogeneity of this sort rarely exists. Every year millions of test takers from all over the world take high-stakes English language proficiency tests, and these test takers represent more than 40 different native languages, as shown by demographic data released by major test agencies (e.g., ETS, IELTS). Further complicating the issue of accent familiarity is that test takers may have varying levels of familiarity with diverse accents other than their native tongue(s), via prior foreign language learning or general exposure (Winke et al., 2013). Such heterogeneity would prevent valid claims made about accent-independent adaptation in testing contexts, as it is impossible to separate out the effects of previous exposure from the effects of

training. For these reasons, the focus of the current dissertation project is speaker-dependent adaptation.

Another important and relevant finding is the factors contributing to this type of adaptation. As revealed by Witteman et al., (2013), the speed of perceptual adaptation to German-accented Dutch relied on two factors – the strength of foreign accent and listener experience with the accent. Results obtained from a cross-modal priming test (i.e., making a lexical decision for written target words) demonstrated that when no exposure was provided, participants with high prior experience with German-accented Dutch were able to interpret words at all levels of accent, but participants with low prior experience were only able to correctly interpret Dutch words with a medium and weak accent. In subsequent experiments, however, when provided with a short exposure phase (four minutes in duration) immediately before completing the task, low-experience-listeners' performance on the strongly accented words improved most when they were exposed to the same speaker producing comparable strongly-accented words, although those tokens were not included in the exposure phase. As such, the short exposure period resulted in equivalent performance to high prior experience. In addition, a short exposure to the same speaker without strongly accented items, as well as to a different German-accented speaker with strongly accented tokens also improved, though improvement in these cases was not observed immediately but occurred only in the second half of the experiment. Such discrepancy resulting from different operationalization of exposure materials for L1 listeners might also be relevant to L2 listeners and therefore should be considered when optimizing exposure input for L2 listeners.

One major limitation of this study, as pointed out by Porretta et al., (2016), is the “relatively coarse-grained” classification employed for both accent strength and listener

familiarity (p.2). As for the former, three levels of accent strength were operationalized only through three Dutch vowels, although these vowels were commonly found to be challenging for native German speakers. As for the latter, two levels of listener familiarity were based on the location of the universities that participants attended: near the German border (i.e., high prior experience) or in central Dutch that is far from the German border (i.e., low prior experience). Nevertheless, the takeaway point is that to be able to understand unfamiliar accents of varying degrees of I/C, extensive prior experience is not always essential; but rather, short-term perceptual learning mechanisms could assist the accommodation of variations in speech (Witteman et al., 2013).

Despite valuable insights into research method and promising findings, there are three main limitations to the aforementioned L1 perceptual adaptation studies when conducting an L2 perceptual adaptation study in a testing situation. Firstly, the task(s) involved in their experiments is limited to word- or sentence-level processing speed (e.g., Clark & Garrett, 2004), or accuracy in sentence recognition (e.g., Bradlow & Bent, 2008). By contrast, listening input often incorporated in high-stakes English language proficiency (e.g., TOEFL, IELTS) is extended discourse, some of which could last for five minutes or even longer (e.g., TOEFL monologic lecture). Secondly, L1 adaptation studies usually recruit well-educated native listeners in the effort to have a participant pool with high proficiency in a prestige dialect. L2 listeners' proficiency, on the other hand, varies wildly, and therefore might have a significant role to play in L2 perceptual adaptation. Finally, the fact that L1 speech perception capitalizes on lexical knowledge to resolve acoustically ambiguous speech could explain why exposure provided solely in audio form works for L1 listeners (Bradlow & Bent, 2008; Norris et al., 2003; Mitterer & McQueen, 2009). L1 listeners can learn to interpret ambiguous speech-sounds based on how

words should sound or disambiguating lexical contexts, and this lexically guided perceptual learning mechanism could lead to shifts in the perceptual category boundary, which could be further translated into being able to interpret previously unheard words (Mitterer & McQueen, 2009; Reinisch & Holt, 2014). An example given in Mitterer and McQueen (2009) is an ambiguous segment midway between /s/ and /f/ (“?”) that appear in sequences such as gira?. Since ‘giraffe’ is a word and ‘giras’ is not, L1 listeners learn to perceive the ambiguous sound as /f/. As such, lexically guided exposure leads to shifts in the perceptual /s/-/f/ category boundary, which can be used to interpret previously unheard words, such as recognizing cli? as cliff and che? as chef, etc. L2 listeners (especially non-advanced listeners), on the other hand, may not be able to access lexical analysis while interpreting ambiguous or unusual sounds (Hamada & Suzuki, 2021). That is to say, mere audio exposure or listening to an unfamiliar accent might not help L2 listeners adapt to that accent as it does for L1 listeners, at least not to the same extent. This might explain why in some L2 perceptual learning studies, lexical information was introduced to support L2 perceptual learning, such as the shadowing-with-script method employed in Hamada and Suzuki (2021), and the use of written subtitles in Mitterer and McQueen (2009). Critically, such an addition of lexical information has been found more effective in facilitating L2 perceptual adaptation than audio exposure alone in these studies, providing further evidence that lexical access is a key component driving perceptual adaptation.

2.4 Insights from and Limitations in L2 Perceptual Adaptation Research

While there is still a paucity of L2 adaptation research, three studies – Harding (2018), Hamada and Suzuki (2021), and Mitterer and McQueen (2009) provide a useful starting point for researching L2 perceptual adaptation in a testing context.

Harding (2018) is the first and so far, the only investigation into L2 adaptative behavior in a testing context. In addition to this, another major difference between this study and aforementioned L1 adaptation studies lies in the fact that instead of being exposed to an unfamiliar accent before being tested on their perceptual adaptation to that accent, L2 listeners involved in this study had to hear a listening comprehension passage delivered by a speaker with an unfamiliar accent from the beginning of the experiment. As such, it allowed the researcher to track listeners' adaptive behavior throughout the course of listening, which in turn helped to identify the point at which L2 listeners started to exhibit adaptation. With this information, future research within this line of enquiry would be better informed in terms of how long or short the exposure period could be.

Specifically, this study explored the experience of six L1 French listeners working on two academic lecture listening tasks featuring one familiar accent (i.e., Standard Southern British English) and one unfamiliar accent (i.e., Thai English). Findings revealed four types of listening difficulty, which were virtually all associated with the unfamiliar Thai accent, including general intelligibility difficulty, input decoding, word recognition/segmentation and distraction, as well as four types of strategies to cope with this unfamiliar accent: directed attention, selective attention, comprehension monitoring, and inferencing. The evidence of adaptation detected included two participants' explicit reporting of adjusting to the unfamiliar accent after the first part of the listening test, alongside the overall decreasing of accent-related issues reported by all participants combined after the exposure of the first two to three minutes. Note that the adaptation identified was considered "crude" by the researcher, given that it was "reported adjustment" elicited by retrospective verbal reports and stimulated recall (Harding, 2018, p. 111), meaning that the perceptual adaptation that may happen beneath listeners' conscious

understanding was left undetected. Note also that the I/C of the Thai English-accented speaker was not specified in the study, but as the findings suggested, it posed “a noticeable degree of challenge” (p. 100) to those L1 French listeners who were not familiar with this specific accent, although it did not result in “detrimental breakdowns in comprehension” in most cases at least (Harding, 2018, p. 99).

While being a qualitative study does limit the generalization of its findings, Harding (2018) has brought good news to test developers, as it demonstrates that L2 listeners may be able to adjust to an unfamiliar accent after a relatively short-term exposure stage (two to three minutes in this case), and that such adaptation occurs when listeners are completing academic lecture listening tasks. Furthermore, this study may provide indirect evidence that L2 listening proficiency may have played an important role in L2 adaptation. Although listener proficiency level was neither controlled for nor a variable in this study, the listener who reported the highest instances of accent-related listening difficulties was also the one whose proficiency was slightly lower than the others (i.e., B2 versus C1/C2 on the Common European Framework of Reference, 2020). In light of these findings, the researcher suggested future research working with more robust samples and setting stricter control over listeners’ L2 proficiency in order to provide more insights into L2 adaptation in testing contexts.

Compared with Harding (2018), the operationalization of exposure conditions in Hamada and Suzuki (2021) and Mitterer and McQueen (2009) are more ‘conventional’ or aligned with L1 adaptation studies. The most significant difference, however, is lexical information used alongside audio exposure that was intended to enhance lexical knowledge for L2 listeners. Hamada and Suzuki (2021) compared two exposure conditions, that is, shadowing alone and shadowing accompanied by a script, in the hope of developing a teaching technique that is more

effective in helping Japanese EFL listeners (with proficiency levels ranging from intermediate to upper intermediate) adapt to unfamiliar accents within a limited amount of time. Shadowing refers to “a paced, auditory tracking task which involves the immediate vocalization of auditorily presented stimuli.” (Lambert, 1992, p. 266). This technique has the merits of directing EFL listeners’ attention to phonological features to improve their bottom-up perception skills, but with relatively less attention on meaning processing or top-down process, especially in the case of low proficiency EFL listeners (Hamada & Suzuki, 2021). Such a drawback of shadow practice underpins the rationale behind the comparison between shadowing-only and shadowing-with-script, and the assumption was that shadowing (for promoting learners’ speech perception skills) used in collaboration with the assistance of a script (for lexically guided retuning) may be able to facilitate learners’ perception skills and to activate their lexical knowledge simultaneously. The findings backed up such an assumption, demonstrating that shadowing alone may not improve L2 perceptual adaptation, whereas shadowing with the assistance of a script promoted EFL listeners’ perceptual adaptation to the unfamiliar Chinese accent (a new speaker of the same accent involved in the exposure/practice stage), but not the Italian accent (which was not included in the exposure/practice stage).

Mitterer and McQueen (2009), a perceptual learning study, sought to find out whether subtitles would aid L2 listeners confronted with unfamiliar foreign regional accents. Similar to shadowing-with-script, the idea that subtitles may facilitate adaptation to unfamiliar foreign accents is based on the premise that subtitles would indicate which words are being spoken, and hence can boost lexically-guided learning about unfamiliar foreign speech. L1 Dutch listeners, fluent in both spoken and written English and as yet unfamiliar with Australian and Scottish varieties of English, watched a 25-minute video containing either ‘strongly accented’ Australian

English or Scottish English under three subtitle conditions, that is, English subtitles, Dutch subtitles, no subtitles. Note that subtitles in this case were not word-for-word transcriptions, though they gave most of the words in the speech stream. Note also that no information was provided as to how and why these two accents were defined as ‘strongly accented’, although to those who are only familiar with GA and/or SSBE, Australian and Scottish varieties might sound ‘strongly-accented’. After the exposure, participants were asked to repeat back audio excerpts spoken by the main characters in the videos, with half of the excerpts taken from the exposure material whereas the other half being completely new (but from the same speakers).

Results show that audiovisual exposure enabled rapid adaptation to unfamiliar accented foreign speech in the listener’s L2, and more importantly, English subtitles (in the language of videos) led to the best performance on both old and new items, whereas Dutch subtitles (in listener’s L1) only promoted performance on old items but resulted in worse performance on new items. Such enhancement effects brought by English subtitles on new items in particular seem to suggest that L2 listeners were able to retune their perceptual categories to pronunciation features of the exposure speakers, leading to long-term changes in speech perception, which in turn indicated that the retuning benefited from listeners knowing what words they were hearing and that listeners were using lexical knowledge to retune phonetic perception. The negative effect of Dutch subtitles attributed to its inconsistency with the spoken English words may in fact serve as further evidence to support the facilitative effect of lexical-phonological knowledge on perceptual learning.

Collectively, L2 perceptual studies support the facilitative role of lexically guided phonetic retuning by providing orthographic information for L2 listeners in classroom-based research and laboratory study (yet using real speech in a naturalistic setting). However, the

question remains whether such findings could extend to a testing situation where test takers are normally under great time pressure and with high anxiety to adjust to an unfamiliar accent in order to achieve higher scores on a test.

3 THE CURRENT STUDY

Considering the evidence presented so far, the critical question should no longer be whether or not L2 accents should be included, but, rather, how to incorporate various L2 accents in a responsible and feasible way so as to “create a more ecologically valid test (using different English accents) while maintaining fair conditions for test takers.” (Kang et al., 2019, p. 62). As such, the current study set out to explore the possibility of incorporating L2 accents with low familiarity and (relatively) low intelligibility to the targeted L2 listeners in high-stakes English proficiency tests. It was guided by the following research questions:

- 1.1 How do passage-level listening comprehension scores vary across four exposure conditions (i.e., audio-with-script exposure, audio-only exposure, no exposure, and baseline), among L2 listeners at different proficiency levels (i.e., high, medium, and low), and between two passage types (i.e., academic monologue lectures and conversations)?
- 1.2 How does item-level response vary across four exposure conditions (i.e., audio-with-script exposure, audio-only exposure, no exposure, and baseline), among L2 listeners at different proficiency levels (i.e., high, medium, and low) and three item types (i.e., main idea, explicit detail, and implicit detail)?
2. How do exposure condition (i.e., audio-with-script exposure and audio-only exposure), passage type (i.e., academic monologue lectures and conversations), and speaker (i.e., L2 speaker 1 and L2 speaker 2) influence L2 listeners’ perceived efficacy of exposure?
- 3.1 How do L2 listeners’ attitudes towards speakers differ between passage type (i.e., academic monologue lectures and conversations) and accent (i.e., ‘standard’ and L2)?

3.2 How do L2 listeners' attitudes towards speakers differ among exposure condition (audio-with-script exposure, audio-only exposure, and no exposure), between passage type (i.e., academic monologue lectures and conversations) and the two L2 speakers?

To answer these research questions, an L2 perceptual adaptation study in the context of a simulated TOEFL iBT listening comprehension test was designed in order to find out whether or not the assistance of a short-term exposure to an L2 speaker's accent would improve comprehension of texts delivered by that speaker. As discussed in Sections 2.3 and 2.4, variables relating to speaker(s), listener(s), and test task(s) would come into play and affect perceptual adaptation, which, in turn make a difference to L2 listening comprehension. Therefore, it is vitally important to take all these variables into consideration while carrying out this research. The following sections describe each variable included in the main experiment alongside its rationale.

3.1 L2 Speakers/Accents

First and foremost, which L2 speakers/accents to test? To get around shared-L1 effects and to broaden the range of accents in large-scale listening tests, the current study took an innovative approach by using L2 accents that are rated around '3' on Ockey and French' (2016) accent scale, defined as mid-I/C in the current study, and that are unfamiliar to the targeted L2 listeners.

Reasons for excluding high I/C were laid out earlier (see Section 2.2.3). Rationales for excluding Low I/C speakers were as follows. Low I/C speakers are probably not able to meet admissions requirements in terms of English language proficiency in the first place, and thus presumably are not very common in the TLU domain. Then, even though speakers with low I/C do exist, in real life L2 listeners usually have at their disposal much more time and additional

resources (e.g., slides shared by instructors, opportunities to ask classmates to clarify or repeat) to adapt to these speakers. Finally, the chance of L2 listeners being able to adapt to low I/C speaker would be reasonably low based on findings from previous studies, or exposure period required to allow such adaptation to happen is likely to be much longer than the one that could be incorporated into the listening sections of high-stakes standardized tests without causing practicality related issues. For these reasons, using L2 speakers with mid-I/C would be the option that provides a more realistic picture of the TLU domain of higher education where L2 speakers may not have optimal intelligibility, which in turn ensures that test takers would be reasonably challenged in order to assess a more multidialectal construct of listening comprehension.

With respect to using unfamiliar accents, the rationales behind this are threefold. Theoretically, it aligns with what Buck (2001) proposed, that is, to select an accent that is equally unfamiliar to all test takers so that shared-L1/familiarity effect would be largely minimized. From the perspective of interactiveness (see Section 2.1.2), if the ability to process or cope with unfamiliar varieties of English is decided as part of the construct of a listening test, then a listening passage featuring accents of low familiarity is more likely to engage a test taker's competence, or knowledge and skills in this regard, as opposed to a test featuring accents of high familiarity, and thus would allow more valid inferences made about a test taker's ability to tolerate, rapidly attune to unfamiliar English accents, ultimately to successfully negotiate meanings in the TLU domain. Last but not least, from a practical standpoint, reducing this listener variable to just one level rendered the data collection manageable, which in turn consolidated the study design. As for why using two different speakers/accents instead of one, it is related to the purpose of the experiment: to test L2 listeners' ability to adapt to a new variety, rather than their ability to understand a particular one.

As an L2 listener myself, I am fully aware of the challenges and discomfort facing test takers who participated in the main experiment; and as a researcher, I believe that it is best to design the characteristics of the test task that promote feelings of comfort and/or safety in test takers, which in turn elicits their best performance. However, as Bachman and Palmer (1996) pointed out, “there needs to be a balance between what test taker feels comfortable with and what we want to measure” (p. 66). To mitigate the amount of anxiety and discomfort, an exposure to each of the L2 speakers, conceived of as a ‘buffer zone’, was provided in part for this purpose (see Section 3.2 for more details), although such as exposure was not provided for test takers assigned to the conditions that were intended for no exposure, and the primary and sole reason of this is to enhance the robustness of study design.

A final note with respect to accent familiarity is that it is the level of familiarity with each individual speaker in the specific context of a L2 listening comprehension test (or local familiarity) rather than with the entire L1 group that each testing speaker represents (or global familiarity) that is the main focus of interest of the current study. Collecting global accent familiarity data is, in fact, the common practice in the literature, with the Language Experience Questionnaire developed in Harding (2011) being a case in point. However, data elicited would be global accent familiarity with a given variety, and yet it probably lacks legitimacy to generalize such overall accent familiarity to the familiarity with a specific speaker from that language variety.

3.2 Length of Exposure

Given the dearth of investigation into L2 perceptual adaptation and the exploratory nature of this dissertation project, a small-scale pilot study was conducted to trial the instruments and procedures to inform a full-fledged experiment. In particular, the pilot study sought to find out

whether or not a short-term exposure (i.e., four 20-second clips) to an L2 speaker would largely prepare L2 listeners for an ensuing monologic lecture listening comprehension task featuring the same speaker. Results showed that all participants ($N = 6$), except one, deemed having four 20-second clips sufficient. Amongst the five participants, four reported that having two or three clips was enough, and one stated that their performance would not show much variation even without an exposure. In light of such results, the final exposure input was set at one minute (60 seconds). Logistically, a 60-second exposure seems feasible in a testing situation. It was also hoped that a relatively short exposure would better distinguish test takers at different proficiency levels.

3.3 Exposure Conditions

Although lexically guided phonetic retuning by providing orthographic information has been observed for L2 listeners in classroom-based research (Hamada & Suzuki, 2021) and laboratory study (Mitterer & McQueen, 2009), it is not yet known how this practice would play out in a testing situation. As such, three different exposure conditions were operationalized, namely, audio-with-script exposure, audio-only exposure, and no exposure. The difference between audio-with-script exposure and audio-only exposure, as their names suggest, lies in with or without a script accompanying its audio. Participants were randomly assigned to one of three conditions for the passages featuring L2 speakers. At the same time, they were also tested on parallel listening passages recorded by ETS voice actors, which is the baseline condition. With counterbalanced design (see Section 4.6.2.4 for explanation), test takers worked on the same listening passages, though speaker(s) who recorded or participated in the listening passages varied, depending on exposure conditions and counterbalanced design. This way, comparisons can be made for the same listening passage under different conditions so as to find out whether and to what extent a 60-second exposure to those L2 speakers would facilitate L2 perceptual

adaptation, and ultimately, improve comprehension of texts delivered by them. In addition, it is also important to delve into test takers' perception of exposure opportunity.

3.4 Listener Proficiency

Very few studies set strict controls over listeners' proficiency in the existing literature on comprehension of nonnative English speech, leaving this important listener factor "underresearched and thus undertheorized" (Kang et al., 2019, p. 2). To this date, only one study, Kang et al., (2019), has specifically used L2 listeners' proficiency level as a variable to uncover its role in L2 listening comprehension, although perceptual adaptation was not the focus of that study. In their study, L2 proficiency level was defined using TOEIC scores: 805 or higher = *advanced*, between 605 and 800 = *intermediate*; and scores of 600 or lower = *beginner*. Their findings demonstrated a strong association between proficiency and listening comprehension, and more importantly, a complex interplay of speaker intelligibility and listener proficiency. Specifically, in the case of high-comprehensibility speakers, neither advanced nor beginner L2 listeners' comprehension scores showed a difference across speakers (both native and nonnative speakers included); in other words, advanced listeners scored equally well across speakers, whereas beginner listeners scored poorly, irrespective of speakers. By contrast, intermediate listeners' scores were significantly different between British speakers (scored highest) and Mexican speakers (scored lowest), though no other comparisons between speakers were statistically significant. In the case of low-comprehensibility speakers, on the other hand, both advanced- and intermediate-level listeners exhibited significant differences. Beginning-level listeners, again, had low scores across the board. Note that in this study speakers' comprehensibility was determined using a seven-point scale (1 = easy to understand; 7 = difficult

to understand). High-comprehensibility speakers were rated between 1 and 2, and low-comprehensibility speakers were rated between 4.5 to 5.

Furthermore, recall that when L2 listeners' adaptive behavior was documented during the course of listening in Harding (2018), the listener whose proficiency level was the lowest amongst the participants involved contributed the highest frequency of pronunciation-, accent-, intelligibility-related episodes, defined as, "instances of talk where listeners oriented towards, or problematized, the speaker's pronunciation, accent or intelligibility" (Harding, 2018, p. 103). These pieces of evidence together seem to suggest that L2 proficiency level has an important role to play in L2 perceptual adaptation, and yet it remains unclear the degree to which this variable would exert its influence on L2 perceptual adaptation. Lastly, proficiency level varies in real-world testing contexts. Hence, probing into L2 listeners' proficiency levels would not only shed light on how this listener variable would influence L2 adaptation, but also broaden the generalizability of research findings.

3.5 Passage Type

Prior research on the effects of nonnative accent on L2 listening comprehension has predominately dealt with monologic discourse, academic lecture in this case. In real-life academic settings, however, listening contexts concerns also interaction between a speaker and one or more listener(s), and such interactive discourse has already been reflected in most high-stakes listening tests. Dialogues differ from monologues in the degree of orality (Bloomfield et al., 2010). Orality, according to Tannen (1982), is the extent to which a passage contains features of spoken language as opposed to features typical of written language, such as more disfluencies and redundancy, and simpler syntax. Generally speaking, a conversation, even between a professor and a student, would be more oral than a formal lecture delivered by a professor.

Findings regarding the effects of orality on L2 listening comprehension suggest that L2 listeners have less difficulty understanding passages that are more oral (see Bloomfield et al., 2010 for a review). Results from studies carried out in the field of L2 language assessment are inconclusive, with some echoing whereas others challenging such a finding (Ockey et al., 2016, p.86). The results from Ockey et al., (2016) introduces an added complication, speaker accent, into the mix. That is, test takers were affected by speakers' accent in interactive discourse almost to the same extent as they were listening to monologic lectures (Ockey et al., 2016). Therefore, it is worth investigating the degree to which the effect of short-term exposure on performance of monologic discourse would differ from or echo that on interactive discourse.

In addition to addressing this research gap, incorporating conversations into the main experiment was also a response to one pilot study participant's suggestion that listening to a recording that is about half the length of a monologic lecture would alleviate their 'suffering'. In this sense, using conversations in which an L2 accent replaces one of the two speakers while the other speaker's accent remains 'standard' seems to be a reasonable option. At the same time, the length of half of a conversation, about three minutes in length, should give sufficient time for normalization in the case of L2 listeners who tend to be slower to adjust to the characteristics of voice and delivery than L1 listeners (Field, 2013). Hence, in the main study both academic monologic lectures and conversations were included to explore how a short-term exposure would affect these two passage types differently. To make these two passage types as comparable as possible, instead of using conversations between students and administrative staff, conversations between a student and a professor were chosen.

3.6 Item Type

Finally, very few studies directly investigate how item type may further contribute to item difficulty when listening to an unfamiliar accent. Results from DIF analysis conducted in both Harding (2011) and Shin et al., (2021) suggest that items targeting specific details were more vulnerable to unfamiliar L2 speech. Further, in Ockey et al., (2016), the one item that showed the biggest differential difficulty across the nine different accents was also the most difficult items in a six-item set across all nine speakers. This might suggest, as the researchers argued, that item difficulty and accent strength may interact; in other words, accent strength might have more impact on difficult items than on easier ones. However, it is not yet known how different exposure conditions would interact with different item types.

4 METHODS

This chapter first describes two sequential phases of data collection for the main experiment: Phase 1 speaker audition, laying out each step involved in identifying the final two L2 speakers, and Phase 2 main experiment, explaining the participants and instruments. This chapter then presents data scoring, coding, and analysis.

Phase 1 Speaker Audition

4.1 Recruiting and Initial Screening of L2 Speakers

Flyers were posted on social media platforms (e.g., Facebook and Twitter) to recruit native speakers of Turkish and Ukrainian. The decision to choose L2 English speakers from these two particular L1s was guided by the idea of using English varieties or accents that are less likely to be familiar to the targeted listeners involved in Phase 2, who were undergraduate and graduate students residing and studying in mainland China. This assumption was later confirmed by accent familiarity data collected at Phase 2 (see Section 5.1). All speakers who agreed to participate were asked to make a sample recording using the same script, which was an excerpt from a retired TOEFL iBT monologue lecture. They were asked to practice beforehand so that they sound like a real instructor or professor delivering a lecture; in other words, they should speak fluently and confidently, without worrying about the so called ‘non-standard accent’ but focusing on avoiding mispronunciation, especially in the case of technical terms. To control for rate of speech, speakers were asked to follow the guidance of 2.2 to 2.8 words per second (all the sample recordings received were about 30 seconds in length). They were allowed to record the sample speech file using whatever software they found most comfortable with as long as they did it in a quiet room. Most of them opted for their smart phone devices, which turned out to produce overall high-quality sound files. Speech files were then communicated through work email.

Following this was the initial screening. I worked with a group of seven raters representing both native and non-native speakers of English as well as Applied Linguistics and non-Applied Linguistics majors. We used the Strength of Accent Scale (Ockey & French, 2016; see Section 2.2.1) with threefold considerations: first, this scale was found useful in Ockey and French (2016), a large-scale empirical study; second, using this scale would allow direct comparison with Ockey and French (2016); third, it is relatively easy to use for raters, regardless of their background. The seven raters were provided with directions on how to judge each speaker's accent using the scale. In particular, they were asked to base their judgements on 'standard' American accent. Note that it is the original scale that was used at this stage (see Section 4.3.2 for a revised rating scale), due to the need of screening out speakers whose accent 'was NOT noticeably from than what I am used to', indicating 'standard' American accent. That is to say, only speakers whose accent demonstrated clear L2 accent were able to pass the initial screening. It is worth noting that the goal of this initial screening was to select speakers whose accent approximates mid-I/C or '3' on the Strength of Accent Scale (Ockey & French, 2016) so as to maximize chances of identifying the 'best' mid-I/C speakers for the main experiment. This initial screening resulted in 12 speakers with six from each L1 background, from a pool of 30 speakers with 20 Turks and 10 Ukrainians.

4.2 Candidate Speakers and Audition Recordings

The twelve L2 speakers (six speakers of Turkish and six speakers of Ukrainian) who passed the initial screening participated as candidate speakers. They each recorded two experiment recordings, the length of which ranged from 30 to 50 seconds. Recording texts, again, came from retired TOEFL iBT monologue lectures. To choose these texts, the principles developed in Ockey and French (2016) were applied: (1) No two clips were identical; (2) Each

clip began at the start of a sentence; (3) Clips with low frequency words, especially those unique to specific fields of study, were avoided; (4) Segments that required additional context to be comprehended were excluded. In addition, all candidate texts were subject to VocabProfiler (Cobb, T. *Web Vocabprofile*; Heatley et al., 2002) for a brief vocabulary analysis. The selected texts fell within the range of 82 – 92 per cent for the most common 1,000 and 2,000 words.

Instructions on how to make and share speech files were the same as those of the initial screening. Once they were received, the speech files were scrutinized for length, mispronunciation, and sound quality. Speakers were asked to rerecord if the length of their recordings fell out of the range of 2.2 to 2.8 words per second, and/or if a word was left out or added, and/or if their sound files were in poor quality. After all speech files were collected and screened, they were edited visually and aurally using both iMovie and Audacity to ensure that there were no unnecessary pauses nor background noise and that the volume across recordings was at the same level.

4.3 Raters, Rating Scale, and Speaker Audition Survey

4.3.1 Raters

Following the procedure in Kang et al., (2018), two groups of raters were recruited for candidate speaker screening: ‘expert raters’ who have a background in Applied Linguistics and/or ESL and ‘novice raters’ who are from other areas of study. Nineteen ‘expert raters’, comprising instructors (2), graduate students (9), and undergraduates (8), were recruited within the department of Applied Linguistics and ESL at a large southern US university. Among these 19 ‘expert raters’, 12 were L1 English speakers and 7 were L2 English speakers.

Novice listeners were recruited through instructors of a graduate level academic writing course at a large southern US university and an undergraduate level elementary Chinese course

at a small college in the south of US. Initially, there were 21 of them, but one rater's data had to be discarded, as some of the ratings contradicted their corresponding comments, meaning that this rater might have read the rating scale in reverse order. The remaining 20 were diverse in terms of first or second language English speakers (first [14], second [6]), major area of study (e.g., Biology, Business Management, Communication, Computer Science, Economics, Political Science, Neuroscience), and academic status (i.e., instructor [1], graduate student [2], undergraduate student [17]). All raters involved were residing in the US at the time of data collection.

4.3.2 Revised Rating Scale

To better serve the purposes of the current study whose focus is L2 accents, the original Strength of Accent Scale went through necessary changes. Table 4.1 shows the revised version. The first major change was related to Level 1 on the original scale. It was removed given that no candidate speakers were rated at this level during the initial screening. This leads to the second change to the scale. That is, the part concerning accentedness at each level was also left out. The resulting four-level scale concentrates on comprehensibility and intelligibility. With these changes, level 1 on this new scale corresponds to Level 2 on the original scale, level 2 as level 3, and so on. Note that the two rater groups used the same revised Strength of Accent Scale for the sake of consistency, and hence easier comparison of results between them.

Table 4.1 *Revised Strength of Accent Scale*

Level	Description
1	The speaker's accent did NOT require me to concentrate on listening any more than usual. The accent did NOT decrease my understanding.
2	The speaker's accent did require me to concentrate on listening more than usual. However, the accent did NOT decrease my understanding.
3	The speaker's accent did require me to concentrate on listening more than usual. The accent slightly decreased my understanding.

4	The speaker's accent did require me to concentrate on listening more than usual. The accent substantially decreased my understanding.
---	---

4.3.3 *Speaker Audition Survey*

Two 20-second clips were prepared for each speaker and uploaded to Qualtrics to create the speaker audition survey. All 24 clips were randomized to prevent an order effect. Raters were instructed to use the modified Strength of Accent Scale to judge the accent of each candidate speaker. They were allowed to enter a score that is between levels (e.g., 1.7; 2.2; 3.5) and to leave comments on any speaker(s) as they wish. One practice recording was provided so that raters could familiarize themselves with the scale and judging process, as well as adjust the volume of their computer/ headphones/earbuds if necessary (see Appendix B.1 for the link to this survey). Once it was finalized, the speaker audition survey was sent out to course instructors who then shared the survey link with their students.

4.4 **Selecting the Final Speakers**

Accent rating (average rating of the 39 raters) for each candidate speaker is shown in Table 4.2. Speakers were ordered, based first on country of origin and then on strength of accent. It is worth noting that rating assigned by 'expert raters' was quite comparable with those given by 'novice raters', though, 'expert raters' overall were a bit harsher than 'novice raters'. Rating assigned by L1 speakers also matched those given by L2 speakers, although L2 speakers tended to be a little stricter than their L1 counterparts. From these results, five (two Turkish and three Ukrainian) out of the 12 candidate speakers whose accents were rated around '2' on the revised scale (equivalent to level 3 on the original one) proceeded to the final screening.

Table 4.2 *Accent Rating of Candidate Speakers*

Country of origin	Gender	Average rating			
		First	Second	Mean	<i>SD</i>

Turkey	Female	1.38	1.60	1.49	.68
Turkey	Male	1.80	1.91	1.86	.71
Turkey	Female	1.87	2.31	2.09	.96
Turkey	Female	2.19	2.32	2.26	.88
Turkey	Female	2.49	2.07	2.28	.85
Turkey	Male	2.54	2.12	2.33	.75
Ukraine	Female	1.90	1.70	1.80	.78
Ukraine	Male	1.75	1.85	1.80	.87
Ukraine	Female	1.90	1.71	1.81	.82
Ukraine	Female	1.80	1.83	1.82	.85
Ukraine	Female	1.56	2.15	1.86	.79
Ukraine	Male	2.65	2.10	2.38	.83

Note. Bold indicates selected to participate in the study

During the final screening stage, I had lengthy discussions with a native speaker of Turkish who has a PhD degree in Applied Linguistics, and a native speaker of Ukrainian who is a current PhD student in Applied Linguistics before consensuses were reached. Essentially, the two final speakers were selected with fourfold considerations. Firstly, their accents exhibit typical pronunciation features of that L1 speech community. Secondly, they have the mature voice quality and pitch of an instructor in the field represented by the TOEFL iBT listening passages. This choice may run the risk of reinforcing stereotypes about particular voice characteristics, though the purpose was to use voices similar that of TOEFL iBT voice actors. Thirdly, they are both female and of similar age groups (late 20s and early 30s) so that the chance of gender and age acting as confounding variables would be to a large extent minimized. Finally, they were able to make time for recording exposure and testing materials for Phase 2 of the study. See Section 4.7.2.2 for recordings made by these two L2 speakers.

Phase 2 Listening Tests Administration

4.5 Listeners

A total of 317 L2 listeners participated in Phase 2 of the study. This group of listeners was comprised of undergraduate (227) and graduate students (90) enrolled at universities across

rankings (within Tier 1 category or the Top 100 universities, according to the Chinese university ranking) and regions in mainland China, though they mainly came from three cities, Beijing, Shanghai, and Wuhan. Participants were recruited through their respective university level and/or departmental level WeChat groups. They represented two age groups, 18 to 22 and 23 to 27, which corresponded to the undergraduate/graduate student distinction. The former age group accounted for about 72% of participants, and the latter one made up about 28%. Two hundred and forty-three of them were female, 64 were male, and the remaining 10 chose “prefer not to answer” while responding to the question in relation to gender. The participants were diverse in terms of major area of study (business [28], engineering [32], humanities [197], law [4], medicine [8], natural sciences [16], social sciences [32]). In terms of prior experience with TOEFL iBT test, only 53, accounting for about 17% of the total number, reported having experience preparing for or taking TOEFL iBT test. My initial plan was to recruit university students enrolled in TOEFL iBT test preparatory courses so as to include potential high-stakes standardized test takers (an important stakeholder group). However, due to the situation of COVID-19 pandemic in the US, many Chinese students opted to further their studies in countries other than the US, such as Australia and the UK. The number of students on TOEFL preparation courses therefore witnessed a steady decrease beginning in early 2020. In light of this, a decision was made to remove the constraint of enrollment in TOEFL iBT preparation courses and to work with registered undergraduate and graduate students instead.

4.6 Instruments

Two major instruments were administered in Phase 2 of the study: a pre-test and an experimental test.

4.6.1 Pre-test

A pre-test, the listening section of a sample Michigan English Test (MET), was employed to gauge L2 listening proficiency, one of the two key listener factors under investigation (familiarity with the chosen L2 accents being another). The MET was used as free sample tests were available online; and more importantly, the chance of the targeted listeners being familiar with this test would be slim, and thus minimize test familiarity or practice effect. A follow-up interview conducted with 160 participants confirmed this assumption, who collectively reported having no prior experience in taking this test.

The structure of this test can be found in Table 4.3. There are 50 multiple-choice items on the test. Test takers are given 35 minutes to finish the test. Note that MET, originally being a paper-and-pencil test, underwent necessary modifications to be delivered on Qualtrics. The instructions, for example, were revised and re-recorded by an L1 American English speaker, though the original recordings for the test items were used as they were. However, all efforts were taken to simulate the original delivery format as much as possible. See Appendix B.2 for the link to this test.

Table 4.3 *The Structure of MET Listening Section*

Parts	Format	N of questions
Part 1	Short conversations are each followed by a question.	19
Part 2	Longer conversations between two people are each followed by three or four questions.	14
Part 3	Short talks are delivered by a single speaker and followed by several questions.	17 (4 sets)

Using their test scores obtained from this pre-test (with the highest score being 50), L2 listeners were categorized into three groups: 100 *high* (36 or higher; Mean = 39.98, *SD* = 2.72), 112 *medium* (26 – 35; Mean = 30.22, *SD* = 3.12), and 105 *low* (12 – 25; Mean = 20.38, *SD* =

3.38). It is worth pointing out that these cut-off scores were arbitrary and chosen to achieve an (roughly) even distribution L2 listeners across the three groups.

As discussed in Section 4.5, the L2 listeners involved in the present research were either undergraduate or graduate students in China. Their English education covers a total span of 11-13 years. They typically started taking regular (compulsory) English courses from primary school (Grade 3, mostly) and throughout secondary school (Borg & Liu, 2013; Wu, 2001). At the tertiary level, college English (CE) is a required course in the first two years of university study for all students except those whose major is English language and literature and who study a foreign language (as a major) other than English. At the end of the second year, students are required to take a national English test called College English Test Band 4. The result of this test is considered as a significant indicator of non-English-major students' English proficiency, as it is recognized in Chinese tertiary institutions as well as in the job market (Borg & Liu, 2013).

At the time of data collection, 77 (about 24%) listeners were non-English-major students who were either freshmen or sophomores, meaning that they might not have taken the College English Test Band 4 yet. Further, Chinese university students' general English proficiency levels vary greatly depending on where they come from. Usually, learners from economically more developed areas tend to have higher proficiency, whereas those from economically less developed areas are likely to have lower proficiency. Such an uneven development in English proficiency levels might be particularly the case for listening skills, given the in some provinces listening comprehension tasks were not introduced to college entrance English exams until recent years. Therefore, it is difficult to gauge the overall English proficiency of this group of listeners, though the lower band are likely to sit at B1 level on CEFR.

4.6.2 *Experimental Test*

4.6.2.1 **Listening Passages**

Three non-operational TOEFL iBT listening tests were obtained from Educational Testing Service (ETS) upon request. To select passages for the experimental test, I worked with an experienced TOEFL iBT preparatory course instructor. We first took all three tests separately. We then compared our test results and discussed such aspects of each listening passage as the perceived overall difficulty level (including both listening inputs and test items), structure, length, potential topic familiarity (to the targeted L2 listeners), change of topics and turns (in conversation), density of terminology, and depth and breadth of background knowledge required of the listening passages. Ultimately, we settled on four listening passages: two conversations (between a student and a professor) and two monologic academic lectures. This selection was largely driven by the desire to include listening passages representing different fields of study (minimizing potential effects of topic familiarity) and yet comparable in terms of length, number of turns and change of topics (conversation only). It is important to point out that instead of conversations between a student and a staff member, I chose those between a student and a professor in order to allow easier comparisons with monologic lectures, given that almost half of the two chosen conversations (timewise) were discussions about academic topics. A detailed description of each listening passage is presented in Table 4.4.

Table 4.4 *Details of Each Listening Comprehension Passage on the Experimental Test*

Listening passage	Major topic(s)	Field of study	Word count	Length of original recording	N of turns	N of questions
-------------------	----------------	----------------	------------	------------------------------	------------	----------------

Conversation 1	Late essay submission, party preparation, compilation of articles in anthropology, speciation	Anthropology	498	2'56"	14	5
Conversation 2	Breathing and respiration at high altitudes, test content, study session	Biology	505	3'07"	12	5
Lecture 1	Glaciers	Geology	701	4'41"	N/A	6
Lecture 2	Zipf's law	Sociology	748	5'27"	N/A	6

Turning now to questions or test items on TOEFL iBT listening test, they are mostly multiple-choice items with one correct answer, although four of them require two answers (one in Conversation 1, one in Lecture 1, two in Conversation 2). Also, for some test items an excerpt of the listening passage is replayed in the question. In other words, for this type of item, the key information is replayed up to three times, which is explained in detail in Section 4.8.2.

It is worth pointing out that all NSs were female, except the one who recorded Lecture 1.

4.6.2.2 Testing Recordings

Three main sets of recordings were used on the experimental test: instruction recordings made by the same L1 American English speaker who recorded instructions for the pre-test, original ETS recordings for the baseline stage and one side of two conversations (student in each of the original conversation recordings) for the testing stage, and recordings made by the two selected L2 speakers for the testing stage, including two monologic lectures, one side of two conversations (professor in each of the original conversation recordings), and one exposure input.

For L2 speakers to make recordings of testing passages, scripts of all four listening passages were shared. As for their corresponding sound files, different decisions were made for lectures and conversations. Lecture recordings were not shared in order to avoid possible priming effect. When it comes to recordings of conversations, speakers were given two options, that is, listening to student's parts in the original recordings or practicing with me or their friend(s) before making their recordings (with me or their friends performing the student part). One speaker chose the former, the other one practiced with her friend. Turning to exposure recordings, each L2 speaker recorded themselves reading one script lasting for about two minutes. One script came from a retired TOEFL iBT listening test obtained from ETS, the other one from TOEFL iBT test preparation materials.

The instructions on how to make and share the above-mentioned speech files were the same as those of the audition clips, except that they were also instructed to use the same recording device for all recordings. Once they were received, speech files were subjected to the same editing process as with the audition recordings. During this round of editing, however, special attention was given to the length of each testing passage recording to ensure that they approximated that of the original recordings. The two testing speakers were paid upon completion of their recordings.

4.6.2.3 Trialing and Revisions

All instructions, rating questions (targeted at comprehensibility, perceived efficacy of short exposure, accent familiarity and attitudes towards speakers), demographic questions were first drafted. Then, three sample tests were constructed, with each corresponding to one exposure condition. They were piloted with three Chinese college students and two instructors (one university instructor of general English and one TOEFL iBT preparatory course instructor) on

Zoom. Feedback on the comprehensibility of instructions and questions and, more importantly, the efficacy of questions in eliciting crucial data as intended, from each trial were discussed with an experienced ESL instructor. Changes were made, accordingly, to the phrasing of instructions and some questions for the sake of clarity. These trials resulted in the final version of the experimental test, detailed explanation of which is in the following sections. Note that all efforts were made to simulate the delivery of the listening section on the TOEFL iBT test as much as possible. For example, a critical feature of the TOEFL iBT listening test in terms of test delivery is that it is an after-listening test, meaning that instead of having test items in front of them while listening to its associated listening passage, test takers would first listen to a listening passage. Only after that, test items would be presented in both aural and written form (in that order). This feature was maintained in the experimental test.

One major difference, however, is that a time limit was set for answering each question (35 seconds) instead of for the entire test, which, according to the ETS website, lasts for 41 to 57 minutes depending on the number of listening passages included on a given test. This was mainly due to technical challenges, as it is not possible to control for the whole testing time when test items are not in the same block on Qualtrics, and data analysis, as it would be the best if participants had time to answer all the questions rather than leaving questions towards the end (testing stage) blank.

4.6.2.4 Structure of the Experimental Test

Table 4.5 outlines the final version of the experimental test repeated under each of the three exposure conditions, namely audio-with-script exposure, audio-only exposure, and no exposure (see Appendix B.3 for the link to this test). To control for effects of presentation order (which might be of particular significance for a perceptual adaptation study), topic familiarity (of

listening passages) and fatigue, alongside any extraneous factors related to differential task/item difficulty, a counterbalanced design was adopted, resulting in four groups within each exposure condition, and 12 conditions in total (three exposure conditions * four groups under each exposure condition). L2 listeners were randomly assigned to one of these 12 conditions by Qualtrics. Note that with the order of L2 speakers remaining the same across groups (L2 speaker1 preceding L2 speaker2), a limitation exists in such a design. To have a ‘full’ version of a counterbalanced design, however, would require at least another 300 participants, which is beyond the scope of this dissertation project. Table 4.6 displays the numbers of L2 listeners in each group across the three exposure conditions. Table 4.7 presents the numbers of L2 listeners at each proficiency level across the three exposure conditions. Within each proficiency level, L2 listeners were almost evenly distributed to different exposure conditions. Within each exposure condition, the numbers of L2 listeners at the three proficiency levels were also relatively even.

Table 4.5 *Structure of Experimental Test Under One Exposure Condition*

Group	Section 1/Baseline stage		Section 2/Testing stage		Section 3
Group 1	Con1/NS1	Lec1/NS2	Con2/L2 speaker1	Lec2/L2 speaker2	Demographic survey
Group 2	Con2/NS3	Lec2/NS4	Con1/L2 speaker1	Lec1/L2 speaker2	
Group 3	Lec1/NS2	Con1/NS1	Lec2/L2 speaker1	Con2/L2 speaker2	
Group 4	Lec2/NS4	Con2/NS3	Lec1/L2 speaker1	Con1/L2 speaker2	

Note. Con = Conversation; Lec = Lecture.

Table 4.6 *L2 Listeners by Groups Across Exposure Conditions*

Exposure Condition	Group 1	Group 2	Group 3	Group 4	Total
Audio script	26	27	24	25	102
Audio only	33	29	23	26	111
No exposure	31	27	20	26	104
Total	90	83	67	77	317

Table 4.7 *L2 Listeners by Proficiency Levels and Exposure Conditions*

	Audio-with-script	Audio-only	No exposure	Total
High	33	32	35	100
Medium	38	40	34	112
Low	31	39	35	105
Total	102	111	104	317

As shown in Table 4.5, each group consisted of three main sections. Section 1 was the baseline stage, associated with two listening passages delivered by two L1 American English speakers (i.e., TOEFL iBT speakers): a conversation between two speakers (a professor and a student) and a monologic lecture. Section 2 was the testing stage, dedicated to another two parallel listening passages featuring the two L2 testing speakers. With or without a 60-second exposure and with or without a script accompanying that exposure before coming into the listening passages delivered by L2 speakers distinguished the three exposure conditions (see Section 4.7.2.5 and Figure 4.1 for more detail). This resulted in each listening passage having seven separate versions (one from NS + three from L2 speaker1 + three from L2 speaker2). The last section was the demographic survey, collecting such listener information as age, gender, academic status (undergraduate student, graduate student), major, prior TOEFL iBT experience.

4.6.2.5 Structure of One Listening Passage

Figure 4.1 illustrates the structure of one listening comprehension passage on the experimental test. There are four of such on each version of the test. Each listening passage was made up of three or five modules, depending on the exposure condition. Note that modules enclosed by dotted lines appeared only under audio-with-script and audio-only exposure conditions in the case of L2 speakers delivering a listening passage.

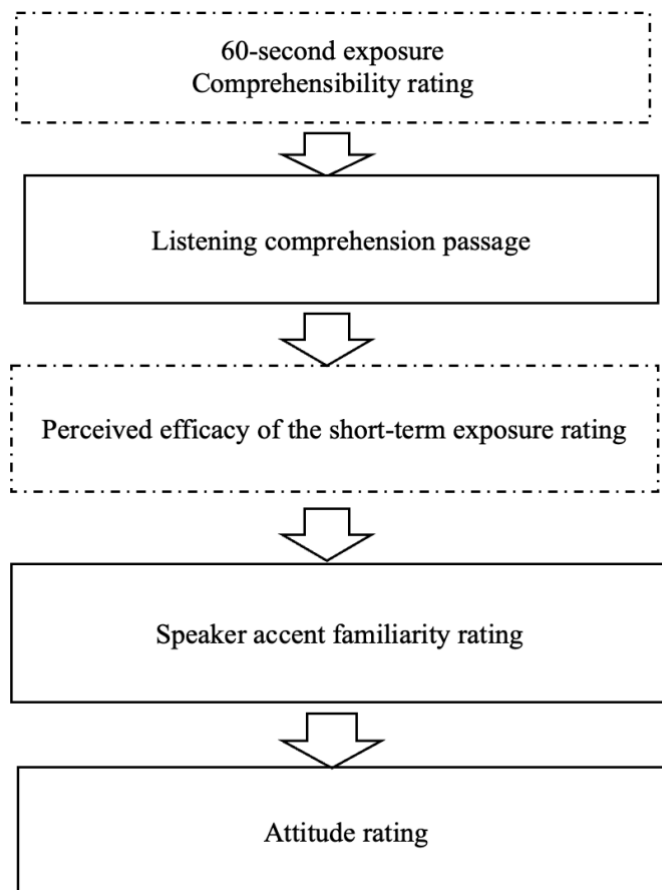


Figure 4.1 Structure of One Listening Passage on the Experimental Test

If assigned by Qualtrics to any of the eight conditions with an exposure, right after listening to the 60-second exposure clip, listeners were asked to rate the professor's comprehensibility on a seven-point scale (1 = *difficult to understand*, 7 = *easy to understand*), with the question being 'How easy is it for you to understand this professor?'. This was followed by one listening comprehension passage. Upon completion listeners were prompted to answer five questions in regard to the efficacy of 60-second exposure. The first four asked participants to rate four statements on seven-point scales (1 = *strongly disagree*, 7 = *strongly agree*). The statements were: 'The 60-second recording helped me get used to the professor's speaking style'; 'The 60-second recording helped me get used to the professor's accent'; 'The 60-second

recording made me less anxious'; 'The 60-second recording was long enough for me to get used to the professor's accent'. Note that Chinese translation was given to "speaking style" and "accent" respectively to avoid misinterpretation. The final question was open-ended and optional, allowing participants to share their opinions on 'In what other ways, if any, did the 60-second recording help?' The final two modules, speaker accent familiarity rating and attitude rating, appeared on all 12 conditions. The question for accent familiarity was 'How familiar were you with the professor's accent?'. Participants responded to a seven-point scale (1 = *not at all familiar*, 7 = *very familiar*).

L2 listeners' attitudes towards each speaker were elicited by four questions. The first three were seven-point-scale questions: 'The professor has bad/good pronunciation'; 'If this professor were my instructor, I would be unhappy/happy'; 'If this professor were included on an English listening test such as TOEFL iBT or IELTS, I would be unhappy/happy' (1 = *bad pronunciation*, 7 = *good pronunciation*; 1 = *unhappy*, 7 = *happy*). These three questions were adapted from Harding's (2011) Speaker Evaluation Task to fit the context of the current investigation: evaluating attitudes towards speakers in a formal academic speech situation and in high-stakes testing contexts. The last one was an open-ended and optional question: 'What comments, if any, do you have about this professor?'. It should be acknowledged that placing these attitude questions right after the listening comprehension passages may obscure the situation where a test taker's negative attitude towards an L2 speaker is triggered by their perceived poor performance. To largely reduce such a risk would be to put these questions at the end of the exposure clip, though this is impossible under the no-exposure conditions, and it does not make much sense to place these questions at different positions across 12 conditions.

4.7 Testing Procedure

The entire testing procedure was administered exclusively online, using Qualtrics. Upon agreement on participation, L2 listeners were provided with a testing manual guiding them through the process. Specifically, there were told that they could finish the pre-test (i.e., the listening section of a sample MET to gauge L2 listening proficiency) and the experimental test on the same day or two separate days, depending on their preference and/or availability. If taking two tests on the same day, they were suggested to take a 20- to -30-minute break. The pre-test took about 35 minutes. The experimental test took about 30 to 60 minutes, depending on the exposure condition. To be able to access the audio recordings embedded on both tests, they were directed to use Google Chrome, Microsoft Edge, or Firefox, instead of mainstream domestic (Chinese) web browsers (e.g., Baidu, 360, QQ Browser). Further, a separate sample test consisting of three mock TOEFL iBT listening comprehension passages (without associated questions) was created to allow participants to check the status of their Internet connection before taking the two main tests, and they were asked to do so before taking each test.

The manual also contained a brief introduction of the two tests, including how the test items and audios were presented (while-listening or after-listening), when to take notes, and so on. In addition, given that most participants had neither prepared for nor taken the TOEFL iBT test before, they were highly recommended to take some sample tests before taking the experimental test. Follow-up interviews revealed that some participants did heed the advice, some did not. Instruction was also given if running into issues related to poor Internet connection, that is, to pause the test by closing the tab instead of pressing on. Fortunately, cases of such were not common and resolved one-on-one. L2 listeners were paid upon completion of both tests.

4.8 Scoring, Coding and Analysis

4.8.1 Test Item Scoring

All items on the pre-test and experimental test were scored dichotomously (i.e., right/1 or wrong/0), and no exception was made to the four multi-select test items on the experimental test (with one in Conversation 1 and Lecture 1, and two in Conversation 2), which is how ETS scores this type of test items. When participants selected the correct answer, they received 1 point; when they chose incorrect answer(s) or did not choose any answer, they received 0 point.

4.8.2 Item Type Coding

According to TOEFL iBT Test Framework and Test Development (2010), listening test items are categorized into five item types depending on the specific ability assessed by each item: ‘understanding main ideas,’ ‘understanding important details,’ ‘recognizing a speaker’s attitude or function,’ ‘understanding the organization of the information presented and relationships between the ideas presented,’ and ‘making inferences or connections among pieces of information’. Since I did not have access to ETS’ confidential item specifications, I worked with two PhD students and one PhD graduate, all of whom specialize in language assessment and testing. We coded the 22 items on the experimental test independently. Disagreements were discussed, and notes were taken on the rationale. The final coding scheme classified items into three main categories: ‘understanding main ideas’ (main idea), ‘understanding important details’ (explicit detail), and ‘making inferences’ (implicit detail) which entailed both ‘recognizing a speaker’s attitude or function’ and ‘making inferences or connections among pieces of information’. The primary reason for this categorization was two-fold: 1) key previous finding regarding item type effects was between main idea versus explicit detail; 2) speaker’s attitude or function and inference both require higher-level processing, and therefore they make sense as a

category together. Note that there is no item falling under ‘understanding the organization of the information presented and relationships between the ideas presented’ on the experimental test.

Test items were also coded for how many times listeners heard the key information. Note that as with TOEFL iBT listening test, the experimental test did not give test takers the option to choose to listen more than once. In general, items targeting main idea and important details were once only, whereas items requiring implicit details varied, once, twice, or three times, mostly twice or three times, though.

Table 4.8 shows the breakdown of test items by item type along with their play times. Each item was represented by four numbers, with the first one being the passage number (1 = Conversation 1, 2 = Lecture 1, 3 = Conversation 2, and 4 = Lecture 2), the second being the question number within each listening passage (1 = question 1, 2 = question 2, and so on), the third describing item type (1 = main idea, 2 = explicit detail, 3 = implicit detail), and the last one showing how many times the key information were listened to (1 = once, 2 = twice, 3 = three times). Item 1111, for example, reads as the first question in Conversation 1, targeted at main idea and heard once; Item 2421 as the fourth question in Lecture 1, targeted at explicit detail and heard once; and Item 4633, as the six (or last) question in Lecture 2, targeted at implicit detail and heard three times.

Table 4.8 *Test Items by Item Types and Listen Times*

Main idea (N = 4)	Play Times	Item type			
		Explicit detail (N = 10)	Play Times	Implicit detail (N = 8)	Play Times
1111	1	1221	1	1431	1
2111	1	1321	1	1533	3
3111	1	2221	1	2532	2
4111	1	2321	1	2632	2
		2421	1	3331	1
		3221	1	3533	3
		3421	1	4231	1

4321	1	4633	3
4421	1		
4521	1		

4.8.3 *Ratings on Accent Familiarity, Efficacy of Exposure Clip, and Attitudes Towards*

Speakers

As mentioned in Section 4.7.2.5, items asking for familiarity with the accent of each speaker, for usefulness and sufficiency of exposure stage, and for attitudes towards speakers were all seven-point-scale questions (i.e., 1 to 7). The higher the number, the higher the familiarity, the higher the perceived usefulness and sufficiency, and the more positive the attitude.

4.8.4 *Statistical Analysis*

Prior to answering the main research questions, it was necessary to check whether or not the participants assigned to the three exposure conditions (i.e., audio-with-script, audio-only and no exposure) differ from each other in terms of L2 listening proficiency. This was done by running a one-way independent ANOVA using *R* (R Core Team, 2022), with pre-test score as the dependent variable and exposure condition as the independent variable. The result showed that there was no significant difference among the participants assigned to the three exposure conditions, $F(2, 314) = .589, p = .556, \omega^2 = .004$. This result could also be seen in the mean scores across the three exposure conditions: audio-with-script: $M = 30.50, SD = 7.95$; audio-only: $M = 29.34, SD = 8.47$; no exposure: $M = 30.34, SD = 8.99$. This further lends weight to the assumption that these three exposure conditions were of comparable L2 English listening proficiency.

The assumption with respect to L2 listening proficiency, however, should be that test takers at high proficiency should outperform those at intermediate level, who should achieve

better performance than their low- proficiency counterparts, on the experimental test. To check this assumption, a one-way independent ANOVA was conducted again using R , with experimental test score as the dependent variables and proficiency level as the independent variable. Results indicated that there was a significant effect of proficiency on experimental test scores, $F(2, 314) = 276.6, p < .0001, \omega^2 = .63$. A Bonferroni *post hoc* comparison revealed that the mean scores for all three proficiency levels were significantly different from each other (high: $M = 15.80, SD = 2.93$; medium: $M = 9.90, SD = 3.52$, low: $M = 6.09, SD = 2.30$).

4.8.4.1 RQ1

To address the first research question (How do passage-level listening comprehension scores and item-level response vary across four conditions [i.e., audio-with-script exposure, audio-only exposure, no exposure, and baseline]), among L2 listeners at different proficiency levels [i.e., high, medium and low], and between two passage types [i.e., academic monologue lectures and conversations]/three item types [i.e., main idea, explicit detail, implicit detail]?), Many-Facet Rasch Measurement (MFRM; Linacre, 1989) was employed. A brief introduction of MFRM is provided in the following sections.

4.8.4.1.1 MFRM

Rasch models are a widely used family of probabilistic models of item response. The basic Rasch model was developed by Georg Rasch (1960/1980), showing how the probability of a person responding correctly to an item could be modelled as a function of the item's difficulty and the person's ability. MFRM extends the basic Rasch model by incorporating additional variables or *facets*, such as raters, tasks, scoring criteria, and time of testing than the two (i.e., person/examine and item) that are typically involved in an assessment situation, and therefore provides "a fine-grained analysis of multiple variables potentially having an impact on test or

assessment outcomes” (Eckes, 2009, p. 3). These *facets* can be anything that test developers/researchers assume might have an influence on test scores in a systematic way (Eckes, 2009, p. 2). The essential aspect is that the measures of the *elements* of the *facets* add (or subtract) to produce the *observations*, such as two facets: persons and items; three facets: persons, items and raters; four facets: persons, items, raters and tasks. Specific examinees, items, raters, etc., in each facet, such as individual test takers in a person/examinee facet, or individual items in an item facet, are referred to as elements of the facet in question.

MFRM models are most frequently associated with rater-mediated performance assessments, where a rating scale is commonly used. For this reason, the MFRM typically specifies the probability, P_{njimk} , that “person n of ability B_n is observed by judge j of severity S_j in category k of item i of difficulty D_i while performing task m of difficulty T_m ” as opposed to the probability $P_{njim(k-1)}$ of the person being observed in category $(k-1)$, as expressed in Equation:

$$\text{Loge} (P_{njimk}/P_{njim(k-1)}) = B_n - S_j - D_i - T_m - F_k$$

where

P_{nmik} = Probability of person n being rated k on item i in task m by rater j

P_{nmik-1} = Probability of person n being rated $k-1$ (i.e., one level down on the rating scale) on item i in task m by rater j

B_n = Proficiency of person n

S_j = Severity of rater j

D_i = Difficulty of item i

T_m = Difficulty of task m

F_k = Difficulty of receiving a rating of k relative to a rating $k-1$ (adjacent categories, equally probable to be observed)

As can be seen from this equation, a MFRM model, in essence, is an additive linear model that is based on a logistic transformation of observed ratings to a logit or log-odds scale. The logistic transformation of ratios of adjacent category probabilities (log odds) can be seen as the dependent variable with multiple facets, such as examinees, raters, tasks conceptualized as independent variables working together to influence these log odds (Eckes, 2009).

4.8.4.1.2 MFRM for Bias/Interaction between Facets

Facets in a model can interact with each other in various ways. For example, it is possible that individual raters (element of the rater facet) display a bias when rating argumentative essays (element of the writing task facet), or put simply, they consistently rate writing prompt A essays higher than writing prompt B essays, or vice versa. Also, in addition to two-way interactions, three or even more facets might come together and have a combined impact on test scores “in subtle, yet systematic ways” (Eckes, 2009, p. 2). Such interactions can be detected by conducting a bias analysis with MFRM (McNamara et al., 2019). Basically, a bias analysis utilizes the overall severity of each rater (across all writing tasks) and the overall difficulty of each writing task that have already been estimated as the basis for determining whether or not this overall pattern holds for particular rater with an essay elicited by a certain prompt. It then predicts the likely score a rater would give when rating a certain essay if this were applied in the same way the rater rated other essays. If the predicted and the observed or actual scores are sufficiently different, and such differences happen consistently when rating essays elicited by a certain prompt, bias is then identified.

Depending on the purpose of an interaction analysis, two types of interaction analyses can be called for: exploratory and confirmatory. As their names suggest, an exploratory

interaction analysis seeks to identify “systematic deviations from model expectations without any specific hypothesis in mind” (Eckes, 2009, p. 32), whereas a confirmatory interaction analysis is to test a specific interaction hypothesis on the basis of a theoretical rationale, of existing literature, or of repeated observations. The interaction analyses conducted to address RQ 1 can be classified as an exploratory interaction analysis, given that there was no theoretical rationale nor specific interaction hypothesis that explicitly states which facets or which subgroups of elements of particular facets are likely to be involved in generating patterns of systematic violations of model expectations; but rather, each and every combination of elements from two or three different facets is examined for significant differences between observed and expected scores.

4.8.4.1.3 MFRM models constructed for RQ1

Translating to the focus of RQ1, two models were constructed, one for passage level analysis, and the other for item level. In the passage-level MFRM model, ID (i.e., test-taker), exposure condition, and listening passage were entered as the primary facets, whereas proficiency level and passage type were treated as dummy facets (i.e., a facet that does not contribute to main measurement, but used to investigate interactions). Listening passage was non-centered. In most MFRM analyses one facet must be non-centered, otherwise the estimates are over-constrained. In the item-level MFRM model, ID (i.e., test-taker), exposure condition and test item were entered as the primary facets, while proficiency level and item type were treated as dummy facets. Test item were non-centered. See Table 4.9 for a summary of the facets’ specification.

Table 4.9 Facets for MFRM Analysis on the Experimental Test Data

Facets							
	Test taker	Exposure condition	Listening proficiency	Passage type	Item type	Listening passage	Test items
Facet values	1-317	1	1 high	1	1	1 C1	1-22
		audio script		conversation	main idea		
		2	2 medium	2	2	2 L1	
		audio only		lecture	explicit detail		
		3	3 low		3	3 C2	
		no exposure			implicit detail		
		4				4 L2	
		baseline					

It is important to note, however, that both models are ambiguous, due to disjoint subsets which could be attributed to the counterbalanced design adopted in the current study (see Table 4.5), where test takers were either in groups 1 and 3 or groups 2 and 4. This issue is called subset connectedness or disjoint subsets, explained in Section 4.9.4.1.4). there are two disjoint subsets. Note also that the variable speaker was not included as a facet in either of the two models. This is because this facet is nested in the facets exposure condition and listening passage (e.g., Conversation 1 + baseline = NS1), and because RQ1 was mainly concerned with exposure conditions, rather than specific speakers. Also, to investigate difficulty level of the three experimental conditions (i.e., audio-with-script, audio-only and no exposure), data from the two L2 speakers were collapsed, and this is justifiable given that the two speakers were at comparable difficulty level. That is, results from MFRM analyses where speaker was incorporated as a facet showed that the distances between L2 speaker1 and L2 speaker2 were a scant .03 logits in both passage-level and item-level models, which is so small as to be of no practical significance, indicating that the two L2 speakers were at similar difficulty level.

As noted in Section 4.9.4.1.2, exploratory interaction analyses were conducted for RQ1. In other words, each and every combination of facets in the passage- and item-level models was scanned to uncover interactions/bias. Hence, seven Rasch bias/interaction analyses were run for each model, including interactions between exposure condition and listening passages/test items (1), exposure condition and listening proficiency (2), exposure condition, listening passages/test items, and listening proficiency (3), exposure condition and passage type/item type (4), exposure condition, passage type/item type and listening proficiency (5), exposure condition, listening passages/test items, passage type/item type (6), exposure condition, listening passages/test items, passage type/item type, and listening proficiency (7). To calculate effect sizes for each contrast, formula $d = 2t/\sqrt{df}$ was followed (Rosenthal & Rosnowm, 2008), with .20, .50 and .80 as thresholds for a small, medium, and large effect size, respectively (Cohen, 1988, p.40).

4.8.4.1.4 Subset Connectedness

In MFRM, connectedness is needed, although it is not necessary, for example, to have all raters rate all examinees on all items. In a connected data set, there exist a network of links through which every element involved in producing an observation is either directly or indirectly connected to every other element of the same assessment context (Eckes, 2009). Lack of connectedness among elements of a particular facet (e.g., among raters) would make it impossible to calibrate all elements of that facet on the same scale, and therefore the measures constructed for these elements (e.g., rater severity measures) could not be directly compared. Only measures within the same subset are comparable, but not across subsets.

If FACETS warns that the data form subsets, there are several actions that can be taken. The one used in the current study was Group-Anchoring to identify equivalent distributions of elements (FACETS manual). The rationale behind this is that if examinees are randomly

assigned to raters, it is then reasonable to assert that the different subsets of examinees are equally able, on average. This is the case for the test taker facet in the current design, since all test takers were assigned at random to one of the 12 conditions by Qualtrics. Hence, in both models the test taker facet was group anchored to resolve disjoint subsets. Treating the facets proficiency level, passage type and item type as dummy facets in the two models was also intended to resolve disjoint subsets, since each test taker only belonged to one proficiency level, each listening passage only belonged to one passage type, and each item only belonged to one item type.

4.8.4.1.5 Measurement Outcomes

MFRM analyses in the present study were conducted using the computer program FACETS version 3.84.0; Linacre (2022). When a MFRM analysis is run, the specified facets are analyzed simultaneously and calibrated onto a single linear scale (i.e., the logit scale), thus making it possible to directly compare all parameter estimates on a common scale (Eckes, 2009), such as test takers, exposure conditions, listening passages/test items, and rating scales. The output from the MFRM analysis to demonstrate such a graphical description of facet statistics is called Wright map (see Figure 5.1 as an example) and was one of the three aspects of output from a MFRM analysis employed in the following chapter (i.e., Chapter 5 Results) to explain answers to RQ1. In addition to a Wright map, a MFRM analysis also provides facet elements measure reports (see Table 5.3 as an example), listing each element of each facet and their estimates.

As for interaction/bias analysis, there are two sets of output from the MFRM analysis: the bias report and the pairwise bias report. The former gives information about, for example, the interaction between exposure condition and listening passage, or the bias size of each listening

passage under each exposure condition, which is estimated based on the overall difficulty of each listening passage and that of each exposure condition. The latter, for instance, provides a contrast value that demonstrates the difference in bias measures of the difficulty of Lecture 1 when under the audio-only exposure condition versus when under the baseline condition, which is essentially a paired t-test. The present study used the pairwise bias report, since it shows the logic difference of the same listening passage or test item under two different exposure conditions, although a bias report of the interaction between exposure condition and listening passage was demonstrated for illustrative purpose.

4.8.4.2 RQ2

The second research question was concerned with L2 listeners' perceived efficacy of the 60-second exposure. RQ2 How do exposure condition (i.e., audio-with-script exposure and audio-only exposure), passage type (i.e., academic monologue lectures and conversations), and speaker (i.e., L2 speaker 1 and L2 speaker 2) influence L2 listeners' perceived exposure efficacy? It is worth remembering that that only test takers who were assigned to either audio-with-script or audio-only conditions were prompted to answer questions regarding the efficacy of the 60-second exposure clip, hence only data from 213 participants, rather than the whole 317 participants, were available to answer the three sub-RQs. To answer this RQ, four three-way factorial ANOVAs were conducted in *R*. In each of these, the independent variables included exposure condition, passage type, and speaker, whereas the dependent variable was the ratings of the four aspects of perceived efficacy of exposure clip (i.e., 'the 60-second recording helped me get used to the professor's speaking style'; 'the 60-second recording helped me get used to the professor's accent'; 'the 60-second recording made me less anxious'; 'the 60-second recording was long enough for me to get used to the professor's accent'). As for the interpretation of effect

sizes for ANOVA analysis, omega squared (ω^2), an unbiased estimator of the population's partial eta-squared (η^2), was used, and it has been suggested that values of .01, .06 and .14 represent small, medium, and large effects respectively (Kirk, 1996, as cited in Field et al., 2012).

Before running each ANOVA, all assumptions were tested. Results from Shapiro-Wilk Tests indicated that the data for all four models were not normally distributed, but the assumption of normality was met by a general equivalence of the group sizes shown in Table 4.6 (Field et al., 2012). The assumption of homogeneity of variance were met, as shown by Levene's Test ($p > .05$). To analyze the data, in each ANOVA model three-way interaction was first entered. In the case of no significant three-way interactions being found, two-way interactions were entered in the second model. If all two-way interactions were insignificant, then all interactions were removed in the third model, where only the main effect of the three independent variables was tested.

4.8.4.3 RQ3

RQ3 addressed L2 listeners' attitudes towards speakers on the experimental test. RQ 3.1 How do L2 listeners' attitudes differ between accents (i.e., 'standard' and L2) and passage types (i.e., academic monologue lectures and conversations)? 3.2 How do L2 listeners' attitudes differ among exposure conditions (audio-with-script exposure, audio-only exposure, and no exposure), and between passage types (i.e., academic monologue lectures and conversations) and the two L2 speakers (i.e., L2 speaker1 and L2 speaker2)? To answer this question, two sets of factorial ANOVAs were conducted, with the first focusing on the two accents (i.e., native versus L2) and the second L2 speakers. The dependent variables in these two sets of ANOVA models were the same, that is, the ratings of the three attitude traits, namely 'good pronunciation', 'happy to have

as my instructor’, and ‘happy to have on a listening test’. The independent variables, however, varied, with accent and passage type included in the first set of models, whereas exposure condition, passage type, and speaker in the second. The process of conducting ANOVA analysis was the same as that of RQ 2.

As a note, a post-task interview was conducted with 160 of the listeners (as mentioned in Section 4.6.1). However, this set of data did not go through full analyses for the current study, but some of the general findings were drawn on in the discussion of the results.

5 RESULTS

This chapter reports the results of the data analyses on the main experiment data. It begins with ratings of familiarity with the six speakers involved in the experimental test, then moves on to reliability coefficients of the 12 conditions of the experimental test. Finally, and most importantly, it reports in detail the results of analyses of the three main research questions.

5.1 Familiarity Ratings

As discussed in Section 4.1, the two L2 speakers were selected based on the assumption that the accents of Turkish and Ukrainian speakers of English would be generally unfamiliar to the targeted Chinese listeners. Table 5.1 shows the familiarity ratings reported by this group of listeners. As we can see, L2S1, the Turkish speaker was rated 2.68, and L2S2, the Ukrainian speaker, were rated and 3, with an average of 2.84 on a 7-point scale (1 = not at all familiar, 7 = very familiar). NSs, on the other hand, were rated around 5.13 on average. The assumption was thus confirmed.

Table 5.1 *L2 Listeners' Familiarity with Speakers' Accent (out of 7)*

Speaker	<i>N</i> of listeners	Mean	<i>SD</i>
NS1	157	5.23	1.73
NS2	157	4.48	1.92
NS3	160	5.33	1.62
NS4	160	5.47	1.63
L2S1	317	2.68	1.53
L2S2	317	3.00	1.71

5.2 Coefficient alpha

Coefficient alpha was .88 ($N = 317$) for the pre-test. Coefficient alpha for each of the 12 conditions of the experimental test (see Table 4.6 for the study design) is shown in Table 5.2,

with an average of .83 (N = 317), ranging from .78 to .89. Thus, 10 tests among 12 had high reliability coefficients and the remaining two had acceptable reliability coefficients.

Table 5.2 *Coefficient Alpha by Test Version*

Exposure Condition	Group 1	Group 2	Group 3	Group 4
Audio script	.78	.83	.84	.84
Audio only	.84	.78	.82	.85
No exposure	.84	.80	.86	.89

5.3 RQ1

For the purposes of this section, I present the results in an order that aims to facilitate the understanding of important steps in the analysis. First, I specify the MFRM model on which the analysis was based. Then, I discuss a graphical display (Wright map) of the joint calibration of ID (i.e., test taker), exposure condition, and listening passage, and the rating scales.

Subsequently, I present measurement results (measure reports) for each facet separately, beginning with the exposure condition facet, followed by result for the listening passage facet/the test item facet. Finally, I discuss the seven interaction/bias analyses conducted for each model.

5.3.1 RQ 1.1

RQ 1.1 asked how passage-level listening comprehension scores would vary across four exposure conditions (i.e., audio-with-script exposure, audio-only exposure, no exposure, and baseline), among L2 listeners at three different proficiency levels (i.e., high, medium, and low), and between two passage types (i.e., academic monologue lectures and conversations). To answer this RQ, a main MFRM model was constructed with ID (i.e., test taker), exposure condition, and listening passage entered as primary facets, while proficiency level and passage type treated as dummy facets. Seven Rasch bias/interaction analyses were then run for this

model, including interactions between exposure condition and listening passage (1); exposure condition and listening proficiency (2); exposure condition, listening passages, and listening proficiency (3); exposure condition and passage type (4); exposure condition, passage type and listening proficiency (5); exposure condition, listening passage, and passage type (6); and exposure condition, listening passage, passage type, and listening proficiency (7).

5.3.1.1 The Wright Map of ID, Exposure Condition, and Listening Passage

MFRM analysis at passage level used scores that test takers obtained from each listening passage to estimate exposure condition difficulties and listening passage difficulties. The program calibrated test taker (labeled “ID” in the analysis), exposure condition, and listening passage onto the same equal-interval scale (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis. Figure 5.1 displays the Wright map representing the calibrations of ID (i.e., test taker), exposure condition and listening passage.

Measr	+ID	+Exposure condition	+Listening passage	S.1	S.2	S.3	S.4
4	.			(5)	(6)	(5)	(6)
3	.*			---	---	---	---
2	***			4	5	4	5
1	*****	baseline		---	---	---	---
0	*****	* audio only	* Con2 <u>Lec2</u>	---	3	---	3
	*****	audio script	<u>Lec1</u>		---	---	---
	*****	no exposure	Con1	2	2	2	2
-1	*****			---	---	---	---
-2	***			1	1	1	1
-3	.			---	---	---	---
-4	.			(0)	(0)	(0)	(0)
Measr	* = 3	+Exposure condition	+Listening passage	S.1	S.2	S.3	S.4

Figure 5.1 Wright Map of ID, Exposure Condition, and Listening Passage

Note. Con = Conversation; Lec = Lecture.

The logit scale (“Measr” – measure) appears as the first column in the map. All measures of test takers, exposure conditions and listening passages are positioned on this scale, -4 to +4 in this case. The second column (labeled “ID”) displays the estimates of test taker proficiency on the four listening passages. Each star represents three examinees, and a dot represents one or two examinees. We can see that test takers are spread over about seven logits. The ID facet is positively oriented, as indicated by the plus sign before its facet name (i.e., ID). In other words, higher-scoring test takers appear at the top of the column, and lower-scoring test takers appear at the bottom. The third column (labeled “Exposure condition”) compares the four exposure conditions in terms of their relative difficulties. In this analysis, the exposure condition facet has a positive orientation, again, as indicated by the plus sign before the facet name (i.e., Exposure condition); that is, the higher the exposure condition logit value, the higher the raw score, and the easier the exposure condition. As shown in Figure 5.1, baseline, appearing at the top, was the easiest exposure condition, whereas audio-with-script and no exposure, appearing at the bottom, were the more difficult ones. This finding suggested that the three experimental conditions, listening passages involving the two L2 speakers, were on the whole more difficult for test takers than the equivalent baseline tasks.

The fourth column (labeled “Listening passage”) compares the four listening passages in terms of their relative difficulties. As with the ID and exposure condition facets, listening passage facet is also positively oriented, therefore listening passage appearing higher in the column were easier than those appearing lower. That is to say, Conversation 2 and Lecture 2 were relatively easier than Lecture 1 and Conversation 1. The last four columns show the five-category rating scale for conversation and six-category rating scale for lecture, mapping the two scales to the equal-interval logit scale. The highest and lowest observed scale levels are shown in

brackets. When comparing the locations of the test takers on the logit scale with the rating scale in the right-hand columns, we can see what score test takers could be likely to receive at specific logit levels. For instance, test takers at a logit value of 2 would be likely to score a 4 on conversation and a 5 on lecture. The horizontal dashed lines are category thresholds.

5.3.1.2 Measurement Report of Exposure Condition (Based on Listening Passages)

As noted in Section 4.9.4.1.5, the second aspect of the output from the MFRM analysis is the measurement report for each facet included in the main measurement. Relevant to RQ 1.1 are exposure condition measure report at passage level (see Figure 5.2) and listening passage measurement report (Figure 5.3). The fifth or “Measure” column in Figure 5.2 provides us with each exposure condition’s position on the logit scale, same as we see in the Wright map (Figure 5.1). That is, baseline was the easiest condition of all. No exposure condition, on the other hand, was the most difficult. The distance between these two exposure conditions was .64 (.42 - -.22) logits. Audio-only condition was the easiest among the three experimental conditions. The distance between audio-only and audio-with-script was very close to that between audio-with-script and no exposure: .09 logits versus .07 logits. The sixth or “Model S.E.” column or Standard error in Figure 5.2 gives us an indication of the precision of the measure. The low standard errors here indicate that the exposure condition measure is relatively precise. The fixed chi-square test tests the hypothesis that all exposure conditions are at the same difficulty level. We can see a chi-square value (i.e., 86.8), the degrees of freedom (3 in this case, as there are four exposure conditions), and a p-value (i.e., .00) in Figure 5.2. A significant result indicates that at least two exposure conditions are significantly different, which is the case of the current data. More revealing, the exposure condition separation index (i.e., “Strata” in Figure 5.2) indicates the number of statistically different level (or strata) of exposure condition performance in the

sample. In the current data, there are 5.8 statistically distinct levels of exposure condition performance. The reliability of the exposure condition separation index (“Reliability” in Figure 5.2) is the Rasch equivalent of Cronbach’s alpha or KR-20 and provides us with an indication of the reliability with which the exposure conditions in the sample are separated. If these four conditions were at the same difficulty level, the reliability index would be zero. A reliability of .94 in the current data, however, shows high reliability, confirming that the difficulty level of the four conditions was highly distinct.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N Exposure condition
1814	634	2.86	2.84	.42	.04	.97	-.5	.94	-1.1	1.05	.82	.80	4 baseline
521	222	2.35	2.33	-.06	.07	1.09	.9	1.07	.7	.89	.76	.79	2 audio only
499	204	2.45	2.23	-.15	.08	.96	-.4	.97	-.2	.99	.77	.78	1 audio script
494	208	2.38	2.17	-.22	.08	.95	-.4	.94	-.6	1.01	.80	.80	3 no exposure
832.0	317.0	2.51	2.39	.00	.07	.99	-.1	.98	-.3		.78		Mean (Count: 4)
567.0	183.1	.21	.26	.25	.01	.06	.6	.05	.7		.02		S.D. (Population)
654.8	211.5	.24	.30	.29	.02	.07	.7	.06	.8		.03		S.D. (Sample)

Model, Populn: RMSE .07 Adj (True) S.D. .24 Separation 3.52 Strata 5.02 Reliability .93
 Model, Sample: RMSE .07 Adj (True) S.D. .28 Separation 4.10 Strata 5.80 Reliability .94
 Model, Fixed (all same) chi-squared: 86.8 d.f.: 3 significance (probability): .00
 Model, Random (normal) chi-squared: 2.9 d.f.: 2 significance (probability): .23

Figure 5.2 Exposure Condition Measurement Report (Based on Listening Passages)

5.3.1.3 Measurement Report of Listening Passages

The listening passage measurement report shown in Figure 5.3 is structured in the same way as the exposure condition measurement report (Figure 5.2). The fifth or “Measure” column in Figure 5.3 shows that the easiest listening passage was Lecture 2, while the most difficult one was Conversation 1. The distance between these two listening passages was .69 logits. The sixth or “Model S.E.” column or Standard error in Figure 5.3, again, shows that the estimation of the listening passage measures is relatively precise. The fixed chi-square test in this case tests the hypothesis that all listening passages are of equal difficulty level. A significant chi-square value, in the case of the current data set, indicates that at least two listening passages are significantly different from each other in terms of difficulty level. Again, the more useful statistic is the

listening passage separation index (“Strata” in the table). The index of 6.65 in the current data set indicates that there are 6.65 distinct levels of listening passage difficulty, with a reliability of .96. Finally, if examining listening passages according to passage type, conversation was more difficult than lecture (see Figure 5.4, column “Measure”), and the distance between them was .25 logits. Again, low standard errors (see Figure 5.4 column “Model S.E.”) show that the estimation of passage type measures is relatively precise.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	PtExp	Group	N Listening passage
978	317	3.09	2.91	.03	.06	1.06	.8	.99	.0	.95	.82	.83	2	4 Lec2
819	317	2.58	2.53	-.08	.07	1.01	.1	.99	.0	.97	.77	.77	1	3 Con2
879	317	2.77	2.44	-.26	.06	.87	-1.6	.87	-1.6	1.11	.84	.82	2	2 Lec1
652	317	2.06	1.88	-.66	.07	.99	.0	1.01	.1	.98	.76	.76	1	1 Con1
832.0	317.0	2.62	2.44	-.24	.06	.99	-.2	.97	-.4		.79			Mean (Count: 4)
118.4	.0	.37	.37	.26	.01	.07	.9	.06	.7		.03			S.D. (Population)
136.7	.0	.43	.43	.30	.01	.08	1.0	.06	.8		.04			S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .25 Separation 4.08 Strata 5.77 Reliability .94
 Model, Sample: RMSE .06 Adj (True) S.D. .29 Separation 4.74 Strata 6.65 Reliability .96
 Model, Fixed (all same) chi-squared: 65.5 d.f.: 3 significance (probability): .00
 Model, Random (normal) chi-squared: 2.9 d.f.: 2 significance (probability): .24

Figure 5.3 Listening Passage Measurement Report (Individual Listening Passage)

Note. Con = Conversation; Lec = Lecture.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	PtExp	Group	N Listening passage
832.0	317.0	2.62	2.44	-.24	.06	.99	-.2	.97	-.4		.79			Mean (Count: 4)
118.4	.0	.37	.37	.26	.01	.07	.9	.06	.7		.03			S.D. (Population)
136.7	.0	.43	.43	.30	.01	.08	1.0	.06	.8		.04			S.D. (Sample)
735.5	317.0	2.32	2.20	-.37	.07	1.00	.1	1.00	.0		.76		1=con	Mean (Count: 2)
83.5	.0	.26	.33	.29	.00	.01	.1	.01	.1		.00		1=con	S.D. (Population)
118.1	.0	.37	.46	.41	.00	.01	.2	.01	.1		.00		1=con	S.D. (Sample)
928.5	317.0	2.93	2.67	-.12	.06	.97	-.4	.93	-.8		.83		2=lec	Mean (Count: 2)
49.5	.0	.16	.23	.14	.00	.10	1.2	.06	.8		.01		2=lec	S.D. (Population)
70.0	.0	.22	.33	.20	.00	.13	1.7	.09	1.1		.01		2=lec	S.D. (Sample)

Figure 5.4 Listening Passage Measurement Report (Passage Type)

Note. Con = Conversation; Lec = Lecture.

Moving on now to the bias/interaction analyses, as noted in Section 4.9.4.1.5, it is pairwise bias reports that were employed to explain each bias/interaction conducted for both passage and item level MFRM analysis in the subsections to follow, although a bias report of

exposure condition and listening passage was used to demonstrate where bias measures come from and that these two reports provide the same information, just in two formats.

5.3.1.4 Rasch Interaction: Exposure Condition and Listening Passage

Figure 5.5 gives information about the bias/interaction of each listening passage under each exposure condition (Column ‘Bias Size+’), estimated based on the overall difficulty (Columns ‘measr +’ on the right-hand side of the figure) of each exposure condition and that of each listening passage. These measures are the same as those presented in Figure 5.2 and Figure 5.3, respectively. Figure 5.5 shows that there are three instances of bias (as enclosed in green box), all connected to Lecture 2. Specifically, for Lecture 2, the observed score for audio-with-script condition was 118 and the expected score is 138.40; the difference between observed and expected score is shown in the ‘Obs – Exp Average’ column (where ‘-’ is a minus) is -.41 score points. The bias size is indicated in logit unit, -.40 logits. This difference is significant as indicated by a significant p -value ($p < .01$). This means that Lecture 2 was significantly harder under the audio-with-script condition than expected. Similarly, under the audio-only condition, the difference between observed and expected score is -.30, corresponding to a bias size of -.30 logits, which was also statistically significant ($p < .05$). By contrast, Lecture 2 was .27 logits easier under the baseline condition than predicted, and the difference was significant, as indicated by a significant p -value ($p < .01$).

Figure 5.6 presents the interaction between exposure condition and listening passage, using the bias size shown in Figure 5.5. It, again, demonstrates that Lecture 2 was significantly easier under the baseline condition than all the experimental conditions. By contrast, for the other three listening passages, the bias was not significant, as shown by the bias size being closer to zero.

Obsvrd Score	Expcd Score	Obsvrd Count	Obs-Exp Average	Bias+ Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq N	Exposure condition	Exposure con meas+r	Listening p N	Li meas+r		
109	103.25	52	.11	.16	.17	.96	51	.3408	1.0	1.0	1	1	audio script	-.15	1	Con1	-.66
99	97.08	55	.03	.05	.17	.32	54	.7485	1.1	1.1	2	2	audio only	-.06	1	Con1	-.66
96	98.73	53	-.05	-.08	.17	-.46	52	.6448	.9	.9	3	3	no exposure	-.22	1	Con1	-.66
348	353.35	157	-.03	-.05	.10	-.51	156	.6122	1.0	1.0	4	4	baseline	.42	1	Con1	-.66
147	139.60	52	.14	.15	.14	1.04	51	.3027	.9	.8	5	1	audio script	-.15	2	Lec1	-.26
135	128.62	55	.12	.12	.14	.89	54	.3758	.9	.9	6	2	audio only	-.06	2	Lec1	-.26
142	133.34	53	.16	.17	.14	1.22	52	.2272	.8	.8	7	3	no exposure	-.22	2	Lec1	-.26
455	477.72	157	-.14	-.14	.08	-1.81	156	.0722	.8	.8	8	4	baseline	.42	2	Lec1	-.26
125	118.04	50	.14	.18	.16	1.13	49	.2631	1.0	1.0	9	1	audio script	-.15	3	Con2	-.08
144	135.60	56	.15	.20	.16	1.30	55	.1983	1.1	1.1	10	2	audio only	-.06	3	Con2	-.08
119	120.27	51	-.02	-.03	.16	-.21	50	.8356	.9	.9	11	3	no exposure	-.22	3	Con2	-.08
431	445.17	160	-.09	-.12	.09	-1.32	159	.1876	1.0	.9	12	4	baseline	.42	3	Con2	-.08
118	138.40	50	-.41	-.40	.14	-2.82	49	.0068	.8	.8	13	1	audio script	-.15	4	Lec2	.03
143	159.85	56	-.30	-.30	.13	-2.23	55	.0295	1.1	1.1	14	2	audio only	-.06	4	Lec2	.03
137	142.00	51	-.10	-.10	.14	-.70	50	.4848	1.1	1.0	15	3	no exposure	-.22	4	Lec2	.03
580	537.83	160	.26	.27	.08	3.36	159	.0010	1.0	.9	16	4	baseline	.42	4	Lec2	.03
208.0	208.05	79.3	.00	.00	.14	.01			1.0	1.0			Mean (Count: 16)				
148.5	146.46	45.8	.17	.19	.03	1.53			.1	.1			S.D. (Population)				
153.3	151.27	47.3	.18	.19	.03	1.58			.1	.1			S.D. (Sample)				

Fixed (all = 0) chi-squared: 37.7 d.f.: 16 significance (probability): .00

Figure 5.5 Bias Report for Exposure Condition and Listening Passage

Note. Con = Conversation; Lec = Lecture.

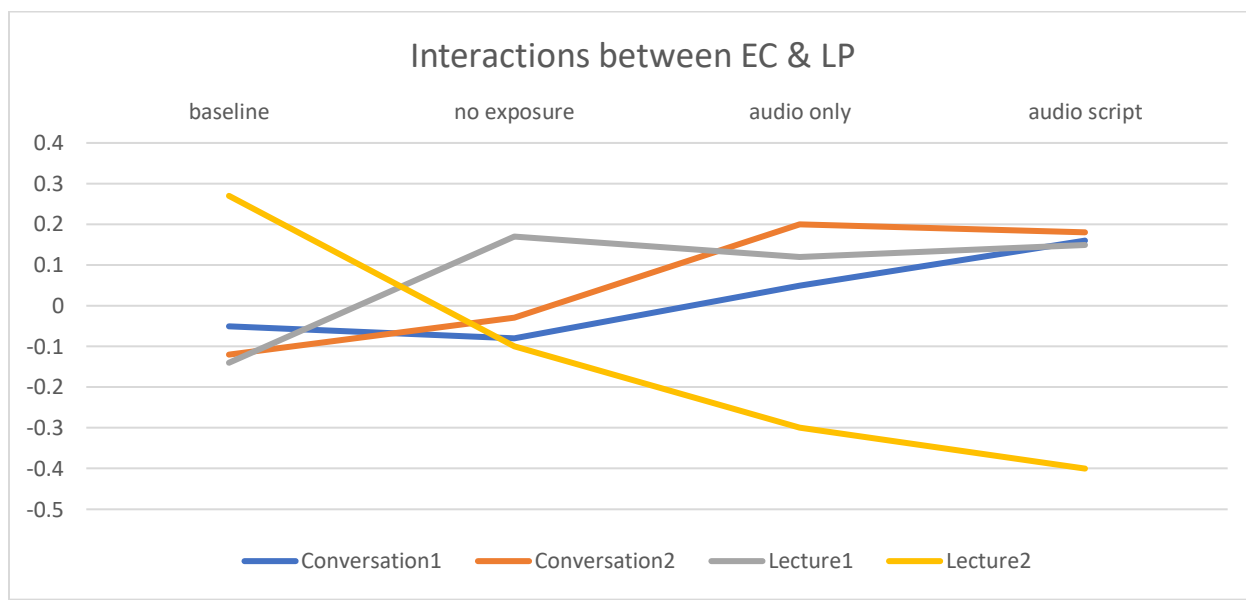


Figure 5.6 Interaction between Exposure Condition and Listening Passage

Note. EC = exposure condition; LP = listening passage.

Table 5.3 is the extracted pairwise bias reports with significant results only. See Appendix B.1 for the complete pairwise bias report. It has the same information as Figure 5.5, though it contrasts the same listening passage under two of the conditions. Before explaining the pairwise bias results, it is necessary to explain where the numbers in Columns ‘Target Measure +’ in Table 5.3 come from (‘+’ means positively oriented). Target measure for Lecture 2 on each condition is bias size + listening passage measure in Figure 5.5, as enclosed in blue box. On the audio-with-script condition, Lecture 2 was $-.40 + .03 = -.37$ logits more difficult; on the audio-only condition, Lecture 2 was $-.30 + .03 = -.27$ logits more difficult; on the no exposure condition, Lecture 2 was $-.10 + .03 = -.07$ logits more difficult; on the other hand, on the baseline condition, Lecture 2 was $.27 + .03 = .30$ easier. Turning now to the pairwise bias/interaction, there are three statistically significant results between exposure condition and listening passage, all related to Lecture 2. Specifically, Lecture 2, the target listening passage, was .37 logits more difficult under the audio-with-script condition, whereas .30 logits easier under the baseline condition. Overall, for Lecture 2, the audio-with-script condition was .67 logits ($-.37 - .30$) more challenging than the baseline condition, or baseline condition was .67 logits easier than the audio-with-script condition. This contrast was statistically significant, as indicated by $p < .001$. Put simply, Lecture 2 was significantly more difficult under the audio-with-script condition than under the baseline condition. Likewise, Lecture 2 was .57 logits harder under the audio-only condition than under the baseline condition. This contrast was also statistically significant, $p < .001$. Similarly, Lecture 2 was .37 logits more difficult under the no exposure condition than under the baseline condition, and the contrast was, again, statistically significant, $p < .05$.

Table 5.3 *Pairwise Bias Report for Exposure Condition and Listening Passage*

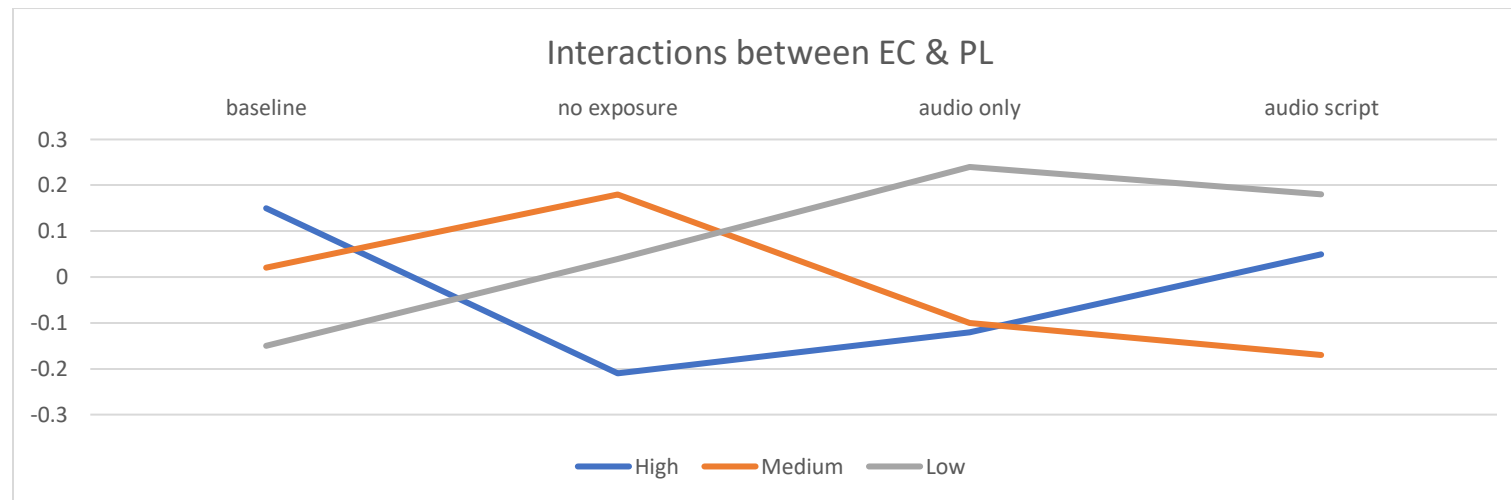
Target listening passage	Target Measure	S.E.	Exposure condition	Target Measure	S.E.	Exposure condition	Target contrast	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
	+			+			+					
Lecture 2	-.37	.14	audio script	.30	.08	baseline	-.67	.16	-4.11	82	.000	-.91
Lecture 2	-.27	.13	audio only	.30	.08	baseline	-.57	.16	-3.63	96	.000	-.74
Lecture 2	-.07	.14	No exposure	.30	.08	baseline	-.37	.16	-2.27	84	.026	-.50

5.3.1.5 Rasch Interaction: Exposure Condition and Proficiency Level at Passage Level

The pairwise bias/interaction analysis between exposure condition and listening proficiency revealed three statistically significant results. This can be found in Table 5.4 (see Appendix B.2 for the complete pairwise bias report). No significant difference was detected for medium-level test takers. As for high-proficiency test takers, they performed significantly worse under the no exposure condition than the baseline condition. The contrast was .36 logits, and reached statistical significance, $p < .05$. However, no such bias was identified between the audio-with-script and baseline conditions nor between the audio-only and baseline conditions, despite the fact that test takers' performance on the audio-with-script and audio-only conditions was lower than the baseline condition. That is to say, for high-proficiency test takers, the 60-second exposure helped them to be tested on listening passages featuring the two L2 speakers, although to a lesser extent when compared with low-proficiency test takers. Regarding low-proficiency test takers, their performance under the audio-with-script exposure condition was .33 logits better than that under the baseline. Similarly, they performed .40 logits better under the audio-only exposure condition than the baseline. Both contrasts were significant, $p < .05$. These findings can also be seen in Figure 5.7, which shows the interaction between exposure condition and listening passage, using the bias size. Together, the results seem to suggest that a 60-second exposure, especially in the form of audio only, was most helpful for the low-proficiency group, and it did make a difference for the high-proficiency group as well, audio-with-script exposure in particular, since they performed better than without it.

Table 5.4 *Pairwise Bias Report for Exposure Condition and Proficiency Level*

Target proficiency	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	-.21	.12	no exposure	.15	.08	baseline	-.36	.15	-2.42	136	.017	-.42
Low	.18	.14	audio script	-.15	.08	baseline	.33	.16	2.03	98	.045	.41
Low	.24	.13	audio only	-.15	.08	baseline	.40	.15	2.63	134	.010	.45

**Figure 5.7** *Interaction between Exposure Condition and Proficiency Level at Passage Level*

Note. EC = exposure condition; PL = proficiency level.

Recall that in Section 4.8.4 a one-way independent ANOVA was conducted using experimental test score as the dependent variables and proficiency level as the independent variable to check if the three proficiency levels were significantly different from each other on the experiment test. Results showed that there was a significant effect of L2 listening proficiency on experimental test scores, with high-proficiency test takers outperforming medium-proficiency level test takers, who did better than their low-proficiency counterpart. One limitation of using mean scores to run ANOVA, however, was that scores obtained under the baseline condition were lumped together with those attained under the three experimental conditions. In other words, the ANOVA results did not really tell us how test takers at different proficiency levels performed under the three experimental conditions separately. Fortunately, MFRM analyses provided this piece of evidence, which is shown in Figure 5.8, using the observed average score for each proficiency level under each exposure condition. As we can see, just like their performance under the baseline condition, under each of the three experimental conditions high-proficiency test takers' performance remained the highest, followed by medium-proficiency test takers, and low-proficiency test takers' performance were the lowest across the board.

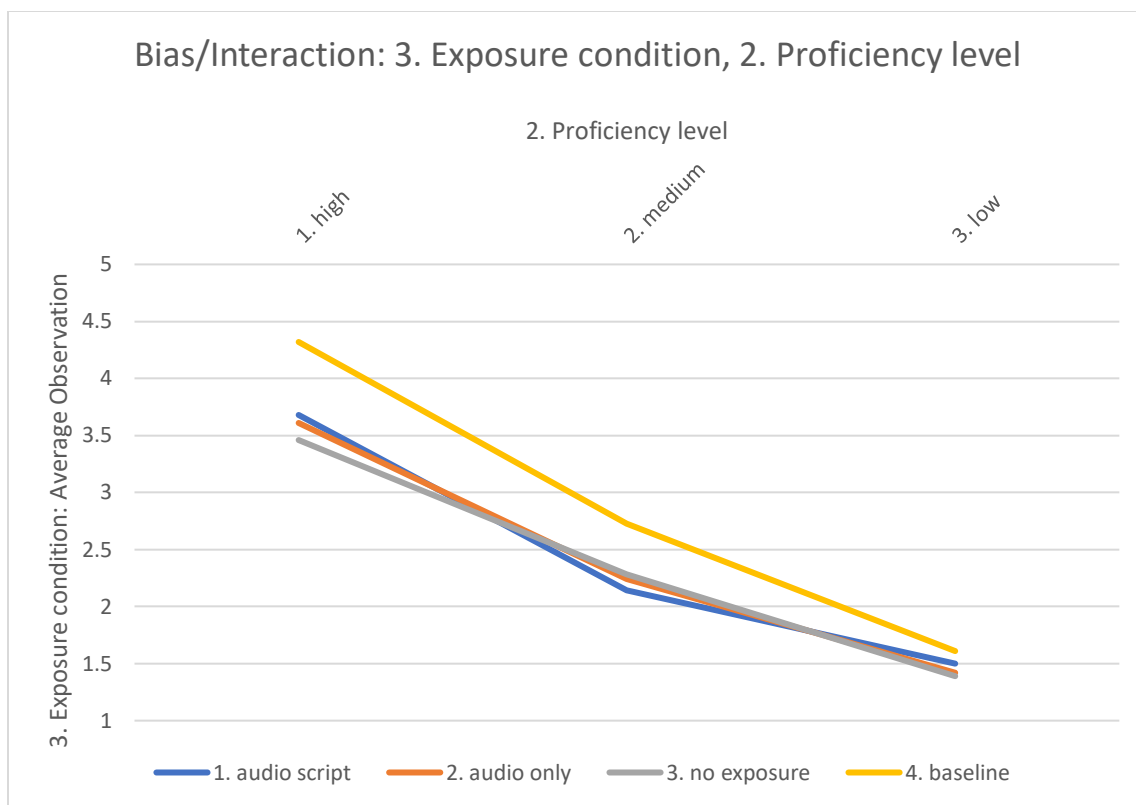


Figure 5.8 Interaction between exposure condition and proficiency level (based on observation average score)

5.3.1.6 Rasch Interaction: Exposure Condition, Listening Passage, and Proficiency Level

The pairwise bias/interaction report for the interaction between exposure condition, listening passage, and proficiency level is presented in Table 5.5 (see Appendix B.3 for the complete pairwise bias report). For high-proficiency L2 listeners, all statistically significant bias was linked to Conversation 2 and Lecture 2. For Conversation 2, both instances of bias were within experimental exposure conditions. That is, Conversation 2 was easiest under the audio-with-script exposure condition, which was 1.45 logits easier than under the audio-only exposure, and 1.35 logits easier than the no exposure condition. By contrast, statistically significant bias related to Lecture 2 was between the baseline and experimental conditions. That is, Lecture 2

was 1.03 logits more difficult under the audio-with-script condition than under the baseline condition, and .81 logits more difficult under the no exposure condition than under the baseline condition.

With respect to medium-proficiency L2 listeners, the significant results were also related to Conversation 2 and Lecture 2. In the former case, the audio-only condition turned out to be .72 logits easier than the baseline condition; in the latter case, the audio-with-script condition was .84 logits more difficult than the baseline condition. Likewise, the audio-only condition was .91 logits more difficult than the baseline condition. The no exposure condition, however, was not significantly more difficult than the baseline condition.

Different from their high- and medium-proficiency counterparts, for low-proficiency L2 listeners all significant bias was related to Lecture 1, and quite interestingly, all three experimental conditions were easier than the baseline condition: the audio-only condition was 1.03 logits, the no exposure condition was .90 logits, and the audio-with-script condition was .79 logits easier, than the baseline condition.

Table 5.5 *Pairwise Bias Report for Exposure Condition, Listening Passage, and Proficiency Level*

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	Conversation 2	.97	.41	audio script	-.48	.30	audio only	1.45	.51	2.84	21	.010	1.24
High	Conversation 2	.97	.41	audio script	-.37	.31	no exposure	1.35	.52	2.61	21	.016	1.14
High	Lecture 2	-.64	.27	audio script	.39	.16	baseline	-1.03	.31	-3.28	20	.004	-1.47
High	Lecture 2	-.42	.24	no exposure	.39	.16	baseline	-.81	.29	-2.75	27	.011	-1.06
Medium	Conversation 2	.43	.24	audio only	-.29	.16	baseline	.72	.29	2.49	37	.017	.82
Medium	Lecture 2	-.50	.21	audio script	.34	.13	baseline	-.84	.24	-3.47	36	.001	-1.16
Medium	Lecture 2	-.57	.21	audio only	.34	.13	baseline	-.91	.24	-3.74	36	.001	-1.25
Low	Lecture 1	.01	.29	audio script	-.78	.16	baseline	.79	.33	2.39	21	.026	1.04
Low	Lecture 1	.24	.24	audio only	-.78	.16	baseline	1.03	.29	3.53	36	.001	1.18
Low	Lecture 1	.12	.28	no exposure	-.78	.16	baseline	.90	.32	2.79	25	.010	1.12

In short, while the overall pattern was that all four listening passages were significantly easier under the baseline condition than under the experimental conditions, the contrast between the baseline condition and the three experimental conditions was particularly pronounced on Lecture 2, which was also the easiest of the passages, and this resulted in the significant bias detected between the baseline condition and the three experimental conditions. A 60-second audio-only exposure was particularly useful for the low-proficiency group, and the high-proficiency groups also benefited from having an exposure. The interaction between exposure condition, listening passage and listening proficiency uncovered two contrasting patterns. That is, Lecture 2 was significantly easier under the baseline condition than the three experimental conditions for high- and medium-proficiency listeners, whereas Lecture 1 was significantly easier under the experimental conditions than the baseline condition for low-proficiency listeners. In addition, the only significant bias between the experimental conditions and the baseline condition was connected with conversations (Conversation 2 in this case), which was measurably easier under the audio-only exposure than under the baseline condition for medium-level test takers only.

5.3.1.7 Rasch Interaction: Exposure Condition and Passage Type

The Rasch bias/interaction analysis yielded no statistically significant results between exposure condition and passage type (see Appendix B.4 for the complete pairwise bias report).

5.3.1.8 Rasch Interaction: Exposure Condition, Passage Type and Listening Proficiency

Table 5.6 shows the pairwise bias report for interaction between exposure condition, passage type, and proficiency level (see Appendix B.5 for the complete pairwise bias report). All significant results were linked to lecture. In the case of high-proficiency test takers, lecture was

.45 logits more challenging under the no exposure condition than under the baseline condition. Regarding medium-level test takers, lecture was .45 logits more difficult under the audio-with-script condition than the baseline condition. Likewise, the audio-only condition was .47 logits more difficult than the baseline condition. Among the three experimental conditions, the no exposure condition was the easiest, which was .50 logits easier than the audio-with-script condition, and .52 logits easier than the audio-only condition. When it comes to low-proficiency test takers; however, lecture was .42 logits easier under the no exposure condition than under the baseline condition.

Table 5.6 *Pairwise Bias Report for Exposure Condition, Passage Type, and Listening Proficiency*

Target proficiency level	Passage type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	lecture	-.31	.16	no exposure	.14	.11	baseline	-.45	-.19	-2.34	70	.022	-.56
Medium	lecture	-.28	.16	audio script	.22	.17	no exposure	-.50	-.23	-2.19	69	.032	-.53
Medium	lecture	-.28	.16	audio script	.17	.09	baseline	-.45	-.18	-2.5	61	.015	-.64
Medium	lecture	-.30	.15	audio only	.22	.17	no exposure	-.52	-.23	-2.3	70	.025	-.55
Medium	lecture	-.30	.15	audio only	.17	.09	baseline	-.47	-.18	-2.67	66	.010	-.66
Low	lecture	.28	.18	no exposure	-.14	.10	baseline	.42	.21	2.03	57	.047	.54

In summary, when listener proficiency facet was not taken into consideration, no statistically significant bias was detected between exposure condition and passage type. However, with L2 listening proficiency level added to the interaction, all bias pointed to lecture, although the direction of bias was opposite for high- and medium-level test takers versus low-level test takers. That is, the baseline condition was easier than the experimental conditions for high- and medium-level test takers, whereas the no exposure condition was easier than the baseline condition for low-level test takers. As such, it appears that the assistance of a 60-second exposure was not of much help for L2 listeners across all three proficiency levels.

As noted in Section 4.8.4.1.2, the type of Rasch interaction analysis carried out in the current study was exploratory interaction analysis. Hence, interaction analyses were also conducted among exposure condition, listening passage, and passage type, as well as among exposure condition, listening passage, passage type, and listening proficiency. However, the results for these two interactions were the same as interaction between exposure condition and listening passage, and interaction among exposure condition, listening passage, and listening proficiency, respectively. This is probably because the facet passage type was nested in the facet listening passage.

5.3.2 RQ 1.2

RQ 1.2 addressed how item-level response would vary across four exposure conditions (i.e., audio-with-script exposure, audio-only exposure, no exposure, and baseline), among L2 listeners at three different proficiency levels (i.e., high, medium, and low) and three item types (i.e., main idea, explicit detail, implicit detail). To answer RQ 1.2, a main MFRM model was constructed with ID (i.e., test taker), exposure condition, and test item as primary facets, while proficiency level and item type as dummy facets. As with passage level analysis, seven Rasch

bias/interaction analyses were conducted for the model, including interactions between exposure condition and test item (1); exposure condition and listening proficiency (2); exposure condition, test item, and listening proficiency (3); exposure condition and item type (4); exposure condition, item type and listening proficiency (5); exposure condition, test item and item type (6); exposure condition, test item, item type and listening proficiency (7).

5.3.2.1 The Wright Map of ID, Exposure Condition and Test Item

MFRM analysis at item level used scores that test takers obtained from each test item to estimate exposure condition difficulties and test item difficulties. The program calibrated test taker (labeled “ID” in current analysis), exposure condition, and test item onto the same the logit scale, thus facilitating comparisons within and between the various facets. Figure 5.6 displays the Wright map representing the calibrations of ID (i.e., test taker), exposure condition and test item. As with the passage level Wright map, all facets are positively oriented, meaning that the higher the logit values (or measures), the higher-scoring test takers, the easier exposure conditions and test items. The first three columns in this item-level Wright map are the same as those in the passage-level Wright map (Figure 5.1), that is, the logit scale (“Mear” – measure), the ID facet, and the exposure condition facet. What is different from the passage-level Wright map (Figure 5.1) are the fourth and fifth columns. The fourth column shows the item type corresponding to each item in the fifth column (for explanation of item number, see Section 4.9.2). The purpose of doing this was to allow easier examination of the distribution of test items falling under different item types in the Wright map. As it shows, the easiest item is an implicit test item, and it comes from Conversation 2. By contrast, the hardest item is a main idea test item, and from Conversation 1. Further, while most (five out of eight) implicit detail items are above 0, the

opposite is true for explicit test items, with six out of 10 below 0. Also, there is no rating scale in this map since all test items were scored dichotomously (see Section 4.9.1).

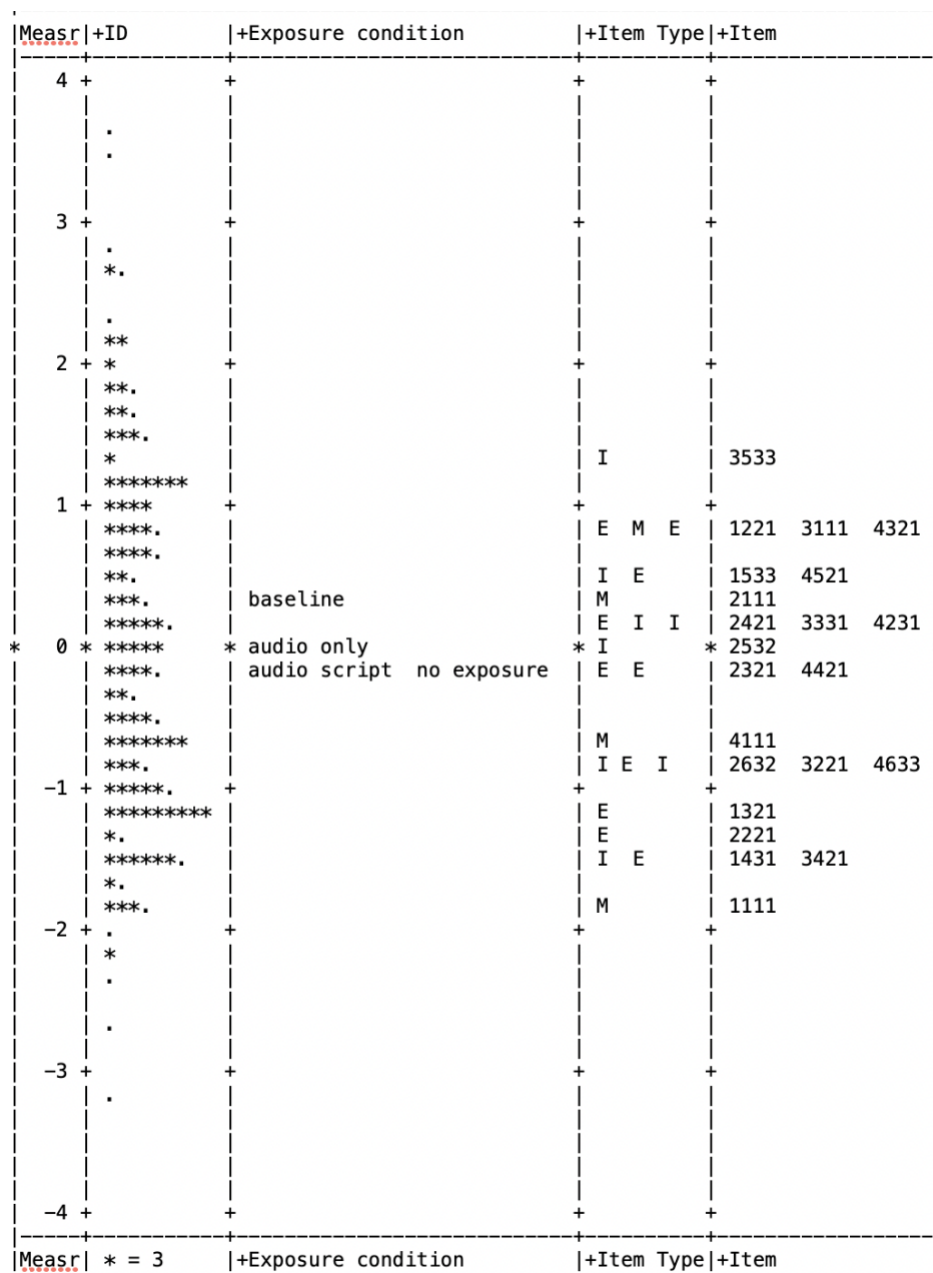


Figure 5.9 Wright Map of ID, Exposure Condition and Test Item

Note. M = main idea; E = explicit detail; I = implicit detail.

5.3.2.2 Measurement Report of Exposure Condition (Based on Test Items)

Exposure condition report at item level can be found in Figure 5.7. As shown in Figures 5.6 and 5.7 (“Measure” column), the difficulty level of the four exposure conditions estimated by test items was in the same order as that estimated by listening passages, with baseline being the easiest condition among four conditions, audio-only condition being the easiest among the three experimental conditions, and no exposure condition being the most challenging of all. The distance between the easiest, baseline, and the most difficult, no exposure, was .56 logits, which is a little smaller than that for passage level (i.e., .64 logits). The distance between the audio-only condition and audio-with-script was again very close to that between audio script and no exposure, .08 logits versus .06 logits. The low standard errors in column “Model S.E.” suggest that the exposure condition measure is relatively precise. Just as the passage-level fixed chi-square test, the hypothesis is that all exposure conditions are of equal difficulty level. A significant chi-square value, however, indicates that at least two exposure conditions are significantly different. The exposure condition separation index (“Strata” in Figure 5.7), 5.43 shows that there are 5.43 measurably distinct levels of exposure condition difficulty, with a reliability of .94.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	N Exposure condition
1814	3487	.52	.53	.37	.04	.94	-3.1	.94	-1.6	1.10	.57	.54	4 baseline
521	1221	.43	.43	-.05	.07	1.05	1.6	1.11	1.7	.89	.51	.54	2 audio only
499	1122	.44	.41	-.13	.07	1.04	1.2	1.14	2.3	.90	.50	.53	1 audio script
494	1144	.43	.39	-.19	.07	1.04	1.2	1.12	1.8	.91	.52	.55	3 no exposure
832.0	1743.5	.46	.44	.00	.06	1.02	.3	1.08	1.1		.53		Mean (Count: 4)
567.0	1007.3	.04	.05	.22	.01	.04	2.0	.08	1.6		.03		S.D. (Population)
654.8	1163.1	.04	.06	.26	.02	.05	2.3	.09	1.9		.03		S.D. (Sample)

Model, Populn: RMSE .06 Adj (True) S.D. .21 Separation 3.27 Strata 4.69 Reliability .91
 Model, Sample: RMSE .06 Adj (True) S.D. .25 Separation 3.82 Strata 5.43 Reliability .94
 Model, Fixed (all same) chi-squared: 75.9 d.f.: 3 significance (probability): .00
 Model, Random (normal) chi-squared: 2.9 d.f.: 2 significance (probability): .23

Figure 5.10 Exposure Condition Measurement Report Based on Test Items

5.3.2.3 Measurement Report of Test Items

Figure 5.8 presents test item measurement report. As can be seen in Figures 5.6 and 5.8 (“Measure” column), the easiest item was Item 3533, which is from Conversation 2, targeting at implicit detail. By contrast, the most difficult item is Item 1111, which is from Conversation 1, targeting at main idea. These two items were more than 3 (3.07) logits apart. Further, there are 11 items with negative logits, meaning that they are relatively more difficult compared with the remaining half with positive logits. These 11 items are almost evenly distributed across the four listening passages (as indicated by the first figure in the item number), with three items from Conversation 1, Lecture 1, and Lecture 2, respectively, and the remaining two from Conversation 2. They cover all three item types, with six asking for explicit detail (10 items in total), three targeting at implicit detail (eight items in total), and two looking for main idea (four items in total). Interestingly, amongst the six explicit detail items four of them (Item 1321, Item 2321, Item 3221, and Item 3421) are the ones requiring two answers, and item 2221 asks for the correct order of a process with four steps. This seems to suggest that items requiring multi-select were more challenging in comparison with items requiring mono select.

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Group	Nu	Item
237	317	.75	.79	1.31	.14	.98	-.2	1.06	.3	1.01	.41	.41	3	16	3533 I
212	317	.67	.70	.83	.13	.93	-1.1	1.03	.3	1.10	.48	.45	1	12	3111 M
212	317	.67	.70	.83	.13	.94	-.9	1.02	.1	1.09	.48	.45	2	19	4321 E
209	317	.66	.68	.77	.13	1.14	2.2	1.22	1.7	.71	.36	.46	2	2	1221 E
194	317	.61	.62	.51	.13	.89	-1.9	.88	-1.1	1.22	.54	.47	3	5	1533 I
191	317	.60	.61	.46	.13	.84	-3.0	.75	-2.6	1.37	.59	.47	2	21	4521 E
184	317	.58	.58	.34	.13	.76	-4.5	.69	-3.6	1.52	.64	.48	1	6	2111 M
177	317	.56	.55	.22	.13	1.06	1.1	1.09	1.0	.85	.44	.49	3	18	4231 I
173	317	.55	.54	.15	.13	.98	-.2	1.03	.3	1.03	.50	.49	3	14	3331 I
171	317	.54	.53	.12	.13	1.28	4.6	1.29	3.0	.40	.30	.49	2	9	2421 E
168	317	.53	.52	.07	.13	.88	-2.2	.81	-2.2	1.29	.58	.49	3	10	2532 I
153	317	.48	.45	-.19	.13	1.03	.5	.99	.0	.96	.48	.50	2	20	4421 E
150	317	.47	.44	-.24	.13	.99	-.1	.98	-.1	1.02	.50	.50	2	8	2321 E
129	317	.41	.35	-.60	.13	.88	-1.9	.87	-1.4	1.22	.58	.50	1	17	4111 M
117	317	.37	.31	-.81	.13	1.00	.0	1.01	.1	.99	.49	.49	3	11	2632 I
116	317	.37	.30	-.83	.13	1.17	2.5	1.20	1.9	.70	.38	.50	2	13	3221 E
116	317	.37	.30	-.83	.13	1.07	1.1	1.01	.1	.90	.46	.50	3	22	4633 I
97	317	.31	.23	-1.18	.14	.96	-.5	.96	-.2	1.06	.50	.48	2	3	1321 E
89	317	.28	.21	-1.34	.14	.94	-.8	.97	-.2	1.08	.51	.47	2	7	2221 E
82	317	.26	.18	-1.49	.15	1.14	1.7	1.67	4.0	.71	.32	.46	3	4	1431 I
81	317	.26	.18	-1.52	.15	1.01	.1	1.13	.9	.96	.45	.47	2	15	3421 E
70	317	.22	.15	-1.76	.15	1.02	.2	1.06	.3	.97	.43	.45	1	1	1111 M
151.3	317.0	.48	.45	-.24	.14	1.00	-.2	1.03	.1		.47				Mean (Count: 22)
48.5	.0	.15	.19	.86	.01	.11	2.0	.20	1.7		.08				S.D. (Population)
49.6	.0	.16	.20	.88	.01	.12	2.0	.20	1.7		.09				S.D. (Sample)
Model, Populn: RMSE .14 Adj (True) S.D. .85 Separation 6.28 Strata 8.70 Reliability .98															
Model, Sample: RMSE .14 Adj (True) S.D. .87 Separation 6.43 Strata 8.91 Reliability .98															
Model, Fixed (all same) chi-squared: 826.6 d.f.: 21 significance (probability): .00															
Model, Random (normal) chi-squared: 20.5 d.f.: 20 significance (probability): .43															

Figure 5.11 *Test Item Measurement Report (Individual Item)*

Note. M = main idea; E = explicit detail; I = implicit detail.

Moving on now to test items with positive logits, again, they almost equally spread in the four listening passages, with two from Conversation 1, and three from Conversation 2, Lecture 1, and Lecture 2, respectively. With respect to item type, five were implicit detail items (eight items in total), four were explicit detail items (10 items in total), and the remaining two were main idea items (four items in total). As for the standard errors in column “Model S.E.”, they indicate that we can be relatively sure of the precision of the measure, although some error does exist.

The fixed chi-square test in this case tests the hypothesis that all test items are at the same difficulty level, though a significant result shows otherwise; that is, at least two test items were significantly different. Furthermore, the test item separation index (“Strata” in Figure 5.8) indicates that there are 8.91 measurably distinct groups of test items in the current sample. And the reliability with which the test items in the sample are separated is .98.

When examining test items according to item type as displayed in Figure 5.9 (column “Measure”), the difficulty levels of explicit detail and main idea were virtually identical (-.31 and -.30 logits, respectively), and were more difficult than implicit detail (-.11 logits).

Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Group	Nu Item
151.3	317.0	.48	.45	-.24	.14	1.00	-.2	1.03	.1		.47			Mean (Count: 22)
48.5	.0	.15	.19	.86	.01	.11	2.0	.20	1.7		.08			S.D. (Population)
49.6	.0	.16	.20	.88	.01	.12	2.0	.20	1.7		.09			S.D. (Sample)
148.8	317.0	.47	.45	-.30	.14	.90	-1.9	.91	-1.1		.53		1=M	Mean (Count: 4)
54.4	.0	.17	.21	.99	.01	.09	1.8	.15	1.7		.08		1=M	S.D. (Population)
62.8	.0	.20	.24	1.14	.01	.11	2.0	.17	1.9		.10		1=M	S.D. (Sample)
146.9	317.0	.46	.43	-.31	.14	1.03	.4	1.05	.4		.46		2=E	Mean (Count: 10)
46.7	.0	.15	.19	.83	.01	.12	2.1	.15	1.5		.08		2=E	S.D. (Population)
49.2	.0	.16	.20	.87	.01	.13	2.2	.16	1.6		.09		2=E	S.D. (Sample)
158.0	317.0	.50	.48	-.11	.13	1.00	-.1	1.07	.3		.47		3=I	Mean (Count: 8)
46.7	.0	.15	.18	.83	.01	.08	1.4	.24	1.7		.07		3=I	S.D. (Population)
49.9	.0	.16	.20	.89	.01	.09	1.5	.26	1.8		.08		3=I	S.D. (Sample)

Figure 5.12 Test Item Measurement Report (Item Type)

Note. M = main idea, E = explicit detail, I = implicit detail.

5.3.2.4 Rasch Interaction: Exposure Condition and Test Item

Table 5.7 shows the pairwise bias/interaction report for the interaction between exposure condition and test item (see Appendix B.6 for the complete pairwise bias report). As we can see, there are nine statistically significant results that were associated with six test items, only one of which (Item 3111, the second easiest test item) was not among the more difficult items. Item 1321 (explicit detail) was .94 logits easier under the audio-with-script condition than under the baseline condition. Similarly, Item 2221 (explicit detail) was .84 easier under the audio-with-script condition than under the baseline condition. Item 3421 (explicit detail) was .86 logits easier under the audio-with-script and 1.01 logits easier under the audio-only than under the baseline. Item 3111 (main idea) was .92 logits easier under the audio-only condition than under the baseline condition. On the other hand, Item 4111 (main idea) was .98 logits more difficult under the audio-with-script condition than under the baseline condition, and was 1.08 logits more

challenging under the no exposure than under the baseline. Likewise, Item 4633 (implicit detail) was .88 logits more challenging under the audio-with-script than under the baseline, and was .92 logits more difficult under the audio-only than under the baseline.

Taken together, the number of items that were more difficult under the baseline condition was twice as many as the ones that were easier under the baseline condition (four versus two), and three of them were test items tapping into explicit detail. Further, three out of four items that were easier under the experimental conditions were from the two conversations. By contrast, both items that were easier under the baseline condition came from Lecture 2.

Table 5.7 *Pairwise Bias Report for Exposure Condition and Test Item*

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
1321 E	-.64	.33	audio script	-1.58	0.2	baseline	.94	.39	2.43	92	.017	.51
2221 E	-.86	.33	audio script	-1.7	0.21	baseline	.84	.39	2.15	92	.034	.45
3111 M	1.5	.34	audio only	.58	0.19	baseline	.92	.39	2.36	91	.021	.49
3421 E	-1.03	.35	audio script	-1.89	0.21	baseline	.86	.41	2.1	86	.039	.45
3421 E	-.88	.33	audio only	-1.89	0.21	baseline	1.01	.39	2.6	103	.011	.51
4111 M	-1.15	.36	audio script	-.17	0.18	baseline	-.98	-.41	-2.42	75	.018	-.56
4111 M	-1.26	.37	no exposure	-.17	0.18	baseline	-1.08	-.41	-2.63	76	.010	-.60
4633 I	-1.43	.38	audio script	-.54	0.18	baseline	-.88	-.42	-2.1	73	.040	-.49
4633 I	-1.46	.36	audio only	-.54	0.18	baseline	-.92	-.40	-2.28	85	.025	-.49

Note. M = main idea; E = explicit detail; I = implicit detail.

5.3.2.5 Rasch Interaction: Exposure Condition and Proficiency Level at Item Level

Table 5.8 shows the pairwise bias/interaction report for the interaction between exposure condition and proficiency level at item level (see Appendix 5.7 for the complete pairwise bias report). The results almost echoed those at passage level, in that significant bias was only found for high- and low-proficiency groups. In addition, as with the direction of bias at passage level, high-proficiency test takers performed significantly worse under the no exposure condition than the baseline condition. The contrast was $-.38$ logits, and significant as indicated by a significant p -value ($p < .01$). On the other hand, low-proficiency test takers performed $.36$ logit better under the audio-only exposure condition than the baseline. This contrast was also significant, $p < .05$. Hence, a 60-second audio-only exposure seems to be most useful for the low-proficiency group.

Table 5.8 *Pairwise Bias Report for Exposure Condition and Listening Proficiency at Item Level*

Target proficiency	Target measure	S.E.	Exposure condition	Target measure	S.E.	Exposure condition	Target contrast	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
	+			+			+					
High	-0.21	0.12	no exposure	0.16	0.08	baseline	-0.38	0.14	-2.62	764	0.009	-0.19
Low	0.22	0.12	audio only	-0.14	0.07	baseline	0.36	0.14	2.59	749	0.010	0.19

5.3.2.6 Rasch Interaction: Exposure Condition, Test Item and Proficiency Level

Table 5.9 illustrates the pairwise bias report for the interaction between exposure condition, test item, and listening proficiency (see Appendix B.8 for the complete pairwise bias report). For high-proficiency test takers, there are nine instances of significant bias associated with six test items, three out of which were from Lecture 2. All six test items were easier under the baseline condition when compared with the experimental conditions. Specifically, Item 4111 (main idea) was 1.87 logits more difficult under the audio-with-script than under the baseline. Item 4421 (explicit detail) was 2.21 logits more difficult under the no exposure than under the baseline. Item 4521 (explicit detail) was 2.86 logits more difficult under the audio-with-script than the baseline. It is worth noting that there was no significant bias between the audio-only version of these three items and their baseline version. The remaining three items were each from the remaining three listening passages. Item 1431 (implicit detail) was 1.37 logits more difficult under the audio-with-script condition than under the baseline. Item 2421 (explicit detail) was 1.48 logits more difficult under the audio-only condition than under the baseline. Item 3221 (explicit detail) was 1.74 logits more difficult under the audio-only condition than the baseline, and was 1.97 logits more difficult under the no exposure condition than under the baseline. Among the three experimental conditions for this test item, the audio-with-script condition was the easiest, 2.25 logits easier than the audio-only condition, and 2.48 logits easier than the no-exposure condition.

Table 5.9 *Pairwise Bias Report for Exposure Condition, Test Item, and Listening Proficiency*

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	1431 I	-2.77	.58	audio script	-1.41	.33	baseline	-1.37	.66	-2.05	33	.048	-.71
High	2421 E	-1.61	.57	audio only	-.13	.41	baseline	-1.48	.70	-2.12	31	.042	-.76
High	3221 E	-.39	.65	audio script	-2.63	.61	audio only	2.25	.89	2.52	24	.019	1.03
High	3221 E	-.39	.65	audio script	-2.86	.69	no exposure	2.48	.95	2.6	24	.016	1.06
High	3221 E	-2.63	.61	audio only	-.89	.30	baseline	-1.74	.68	-2.58	22	.017	-1.10
High	3221 E	-2.86	.69	no exposure	-.89	.30	baseline	-1.97	.76	-2.61	19	.017	-1.20
High	4111 M	-1.57	.63	audio script	.30	.38	baseline	-1.87	.74	-2.55	20	.019	-1.14
High	4421 E	-1.4	.56	no exposure	.81	.45	baseline	-2.21	.72	-3.09	34	.004	-1.06
High	4521 E	-.80	.63	audio script	2.06	.73	baseline	-2.86	.97	-2.95	45	.005	-.88
Medium	1221 E	1.97	.63	no exposure	.23	.29	baseline	1.73	.69	2.52	23	.019	1.05
Medium	2421 E	.66	.54	no exposure	-.62	.28	baseline	1.29	.61	2.11	25	.045	.84
Medium	3421 E	-.87	.50	audio only	-2.65	.49	baseline	1.78	.70	2.55	56	.013	.68
Medium	4111 M	-1.43	.57	audio only	-.12	.30	baseline	-1.31	.64	-2.04	31	.050	-.73
Medium	4521 E	-.41	.47	audio only	1.02	.33	baseline	-1.43	.57	-2.51	40	.016	-.79
Medium	4633 I	-3.16	1.04	audio script	.15	.53	no exposure	-3.31	1.16	-2.85	29	.008	-1.06
Medium	4633 I	-3.16	1.04	audio script	-.30	.30	baseline	-2.86	1.08	-2.66	23	.014	-1.11
Medium	4633 I	-2.27	.75	audio only	.15	.53	no exposure	-2.42	.92	-2.62	34	.013	-.90
Medium	4633 I	-2.27	.75	audio only	-.30	.30	baseline	-1.97	.81	-2.43	26	.023	-.95
Low	1321 E	-.20	.57	audio only	-2.01	.53	baseline	1.81	.78	2.33	52	.024	.65

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	2111 M	.71	.56	audio script	-.68	.34	baseline	1.39	.65	2.12	23	.045	.88
Low	2221 E	-.89	.78	audio script	-4.19>	1.42	baseline	3.31	1.62	2.04	66	.046	.50
Low	2321 E	1.01	.53	no exposure	-.68	.34	baseline	1.69	.63	2.7	28	.012	1.02
Low	2421 E	.39	.58	audio script	2.4	.53	audio only	-2.01	.78	-2.57	29	.016	-.95
Low	2421 E	2.40	.53	audio only	.51	.28	baseline	1.89	.60	3.17	30	.004	1.16
Low	3111 M	1.6	.48	audio only	.27	.30	baseline	1.33	.57	2.33	32	.026	.82
Low	3421 E	-.14	.57	audio only	-1.88	.53	baseline	1.73	.78	2.22	48	.031	.64

Note. M = main idea; E = explicit detail; I = implicit detail.

Turning now to medium-level test takers, as with high-level test takers, there were nine instances of significant bias connected to six test items, and half of them were from Lecture 2, though the direction of bias was a mix, with items from the two conversations and Lecture 1 were harder under the baseline, while items from Lecture 2 easier under the baseline. Specifically, Item 1221 (explicit detail) was 1.73 logits easier under the no exposure condition than under the baseline. The same pattern was observed for Item 2421 (explicit detail), 1.29 logits easier under the no exposure condition than under the baseline. For Item 3421 (explicit detail), it was 1.78 logits easier under the audio-only condition than under the baseline. Notice that all three items were explicit detail. By contrast, Item 4111 (main idea) was 1.31 logits more difficult under the audio-only condition than under the baseline. Likewise, item 4521 (explicit detail) was 1.43 logits more difficult under the audio-only condition than the baseline. For item 4633 (implicit detail), the audio-with-script condition was 2.86 logits, and the audio-only condition was 1.97 logits more difficult than the baseline. Amongst the three experimental conditions, the no exposure condition was the most difficult, which was 3.31 logits more difficult than the audio-with-script condition, and 2.42 logits more difficult than the audio-only condition.

With respect to low-proficiency test takers, the patterns observed were almost the opposite of those observed for their high- and medium-proficiency counterparts. That is, none of the eight significant instances of bias (linked to seven test items) were associated with Lecture 2, and all seven test items were more challenging under the baseline condition than the experimental conditions. Specifically, Item 1321 (explicit detail) was 1.81 logits easier under the audio-only condition than the baseline. Item 2111 (main idea) was 1.39 logits easier under the audio-with-script condition than the baseline. This pattern holds for Item 2221 (explicit detail), which was 3.31 logits easier under the audio-with-script condition than the baseline. For Item

2321(explicit detail), the no exposure condition was 1.69 logits easier than the baseline. Item 2421 (explicit detail) was 1.89 logits easier under the audio-only condition than the baseline. Significant bias was also identified among the experimental conditions, with the audio-only condition being 2.01 logits easier than the audio-with-script condition. For Item 3111(main idea), the audio-only condition was 1.33 logits easier than the baseline. The same pattern emerged for Item 3421, which was 1.73 logits easier under the audio-only condition than under the baseline.

To sum up, for high-proficiency test takers, while all six test items were more difficult under the experimental conditions than the baseline, none of them demonstrated bias against all three experimental conditions, rather, only one of the three experimental conditions was more challenging than the baseline condition in most cases (except test item 3221). It is also worth noting that half of these items came from the same listening passage, Lecture 2. Among the six items demonstrating significant bias for medium-level test takers, the numbers of test items showing bias for and against baseline condition were the same (three versus three). The three items that were easier under the baseline condition were all from Lecture 2. As for low-level test takers, all seven items were easier under the experimental exposure conditions, and four of them were from Lecture 1.

5.3.2.7 Rasch Interaction: Exposure Condition and Item Type

No statistically significant bias/interaction was identified between exposure condition and item type. The complete pairwise bias report can be seen in Appendix B.9.

5.3.2.8 Rasch Interaction: Exposure Condition, Item Type, and Proficiency Level

Table 5.10 shows the pairwise bias report for the interaction between exposure condition, item type, and proficiency level (see Appendix B.10 for the complete pairwise bias report). For high proficiency L2 listeners, two significant biases were identified, and both were explicit

detail. That is, the audio-only condition was .57 logits more difficult than the baseline. Among the three experimental conditions, the audio-with-script condition was .54 logits easier than the audio-only condition. For medium proficiency L2 listeners, the only significant bias was linked to implicit detail, which was .55 logits more difficult under the audio-only exposure than under the no exposure. In other words, no bias was detected between the experimental conditions and the baseline condition for listeners at this proficiency level. For low-level L2 listeners, significant interactions were found for items looking for main idea and for explicit detail. In the former case, the audio-only condition was .80 logits easier than the baseline, and was 1.04 logits easier than the no exposure condition; in the latter case, the audio-only condition was .52 logits easier than the baseline.

In short, significant biases/interactions between baseline condition and experimental conditions were identified for high- and low-proficiency L2 listeners only. For high-proficiency listeners, explicit detail item was more difficult under the audio-only condition than the baseline. By contrast, for low-proficiency listeners, explicit detail items were easier under the audio-only condition than the baseline, as were main idea test item.

Table 5.10 *Pairwise Bias Report for Exposure Condition, Item Type, and Listening Proficiency*

Target proficiency level	Item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	explicit detail	.11	.19	audio script	-.44	.18	audio only	.54	.26	2.08	322	.038	.23
High	explicit detail	-.44	.18	audio only	.14	.12	baseline	-.57	.22	-2.64	300	.009	-.30
Medium	implicit detail	-.27	.18	audio only	.27	.20	no exposure	-.55	.27	-2.03	284	.043	-.24
Low	main idea	.37	.28	audio only	-.67	.36	no exposure	1.04	.45	2.31	133	.022	.40
Low	main idea	.37	.28	audio only	-.43	.18	baseline	.80	.33	2.45	143	.016	.41
Low	explicit detail	.38	.17	audio only	-.13	.11	baseline	.52	.20	2.53	347	.012	.27

Exposure condition									
Audio script	204	3.91	1.75	3.81	1.76	3.95	1.83	3.32	1.68
Audio only	222	4.17	1.66	3.96	1.75	4.00	1.92	3.68	1.87
Passage type									
Conversation	213	4.30	1.61	4.20	1.67	4.30	1.81	3.84	1.78
Lecture	213	3.79	1.77	3.58	1.78	3.66	1.89	3.18	1.74
Speaker									
L2S1	213	4.16	1.68	3.96	1.77	4.00	1.98	3.46	1.78
L2S2	213	3.93	1.74	3.82	1.74	3.95	1.77	3.55	1.79

5.4.1 *Helped Adapt to Speaking Style*

Table 5.12 shows descriptives for ANOVA analyses conducted for the ratings of the four aspects of exposure clip. The results of the three-way ANOVA revealed a significant main effect of passage type on ratings of ‘helped adapt to speaking style’, $F(1, 422) = 8.98, p < .01, \omega^2 = .02$, meaning that the 60-second exposure clip was deemed to be significantly more useful for adjusting to the speaking style of a professor participating in a conversation than that of a professor delivering a lecture. By contrast, neither the main effect of exposure condition nor that of speaker was significant.

Table 5.12 ANOVA Descriptive Statistics for L2 Listeners' Perception of the Efficacy of 60-second Exposure

Exposure condition	Passage type	Speaker	Helped adapt to speaking style		Helped adapt to accent		Made me less anxious		Was long enough	
			Mean	<i>SD</i>	Mean	<i>SD</i>	Mean	<i>SD</i>	Mean	<i>SD</i>
Audio script	Conversation	L2S1	4.13	1.68	4.23	1.69	3.94	1.82	3.28	1.61
Audio script	Lecture	L2S1	3.80	1.72	3.55	1.76	4.06	2.17	3.27	1.68
Audio only	Conversation	L2S1	4.48	1.49	4.27	1.69	4.53	1.96	4.05	1.80
Audio only	Lecture	L2S1	4.14	1.83	3.67	1.86	3.33	1.83	3.12	1.92
Audio script	Conversation	L2S2	4.37	1.54	4.18	1.50	4.43	1.50	3.92	1.64
Audio script	Lecture	L2S2	3.38	1.93	3.30	1.91	3.42	1.67	2.87	1.68
Audio only	Conversation	L2S2	4.18	1.78	4.08	1.84	4.24	1.90	4.08	1.98
Audio only	Lecture	L2S2	3.87	1.58	3.77	1.62	3.81	1.83	3.44	1.69

5.4.2 *Helped Adapt to Accent*

The three-way ANOVA revealed similar results for ratings of ‘helped adapt to accent’. There was a significant main effect of passage type, $F(1, 419) = 7.41, p < .01, \omega^2 = .03$. That is to say, the 60-second exposure clip was perceived to be significantly more helpful for adapting to the accent of a professor participating in a conversation than that of a professor delivering a lecture. However, the main effects of exposure condition and speaker were not significant.

5.4.3 *Made Me Less Anxious*

The results of the three-way ANOVA for ratings of ‘made me less anxious’ revealed a significant passage type * exposure condition * speaker interaction, $F(1, 418) = 6.98, p < .01, \omega^2 = .01$, though further pairwise comparisons between each level of all independent variables (or main effects) only identified two significant interactions. That is, ratings of L2S1 (i.e., the Turkish speaker) under the audio-only exposure condition were significantly lower for lecture ($M = 3.33, SD = 1.83$) than for conversation ($M = 4.53, SD = 1.96$), $d = -.63$. This indicated that, for test takers the 60-second exposure clip was deemed more useful for reducing their anxiety while listening to the Turkish speaker participating in conversations than listening to her delivering lectures. Then, ratings of the Ukrainian speaker under the audio-with-script exposure condition while delivering lectures ($M = 3.42, SD = 1.67$) were significantly lower when compared with the Turkish speaker under the audio-only exposure condition while delivering one side of conversations ($M = 4.53, SD = 1.96$), $d = -.61$. This result, along with a non-significant interaction between exposure condition and speaker while a significant interaction between passage type and exposure condition, and between passage type and speaker appears to suggest that passage type played a bigger role in this interaction.

5.4.4 *Was Long Enough*

The three-way ANOVA for ratings of ‘was long enough’ detected a non-significant passage type * exposure condition * speaker interaction effect, but a significant passage type * speaker interaction, $F(1, 418) = 4.43, p < .05, \omega^2 < .01$. Specifically, ratings of the Ukrainian speaker while delivering one side of conversations ($M = 4.00, SD = 1.81$) were significantly higher than the Turkish speaker delivering lectures ($M = 3.19, SD = 1.80$), $d = .45$, and herself (Ukrainian speaker) delivering lectures ($M = 3.17, SD = 1.70$), $d = .47$. These results seem to suggest that the 60-second clip was regarded as more sufficient for listening to conversations than for listening to lectures, considering that neither of the other two-way interactions (passage type * exposure condition and exposure condition * speaker) were significant. In sum, the results of all four aspects of the exposure clip suggest that having a 60-second exposure was considered more useful for listening to L2 speakers participating in conversations than delivering lectures.

5.5 RQ3

RQ3 focused on L2 listeners’ attitudes towards speakers on the experimental test. Table 5.13 gives descriptive statistics for ratings of the three attitude traits, namely ‘The professor has bad/good pronunciation’, ‘If this professor were my instructor, I would be unhappy/happy’, and ‘If this professor were included on an English listening test such as TOEFL iBT or IELTS, I would be unhappy/happy’. What stands out in the table is that ratings of all three traits were higher for TOEFL iBT speakers than for L2 speakers. Amongst TOEFL iBT speakers, NS2, who delivered Lecture 1, was consistently rated lowest. This probably had less to do with his accent, since all TOEFL iBT speakers speak ‘standard’ American accent, but more to do with his voice quality and the way he paused in the recording, as commented by some test takers. Another interesting trend is that each speaker was rating highest on ‘good pronunciation’, followed by

‘happy to have as my instructor’, and the lowest ratings went to ‘happy to have on a listening test’. This seems to suggest that this group of participants held the highest standards for being a speaker on a listening test, or perhaps they are just never happy to have a listening test.

Table 5.13 *Descriptive Statistics for L2 Listeners’ Attitudes Towards Six Speakers on the Experimental Test*

Speaker	N	Good pronunciation		Happy to have as my instructor		Happy to have on a listening test	
		Mean	SD	Mean	SD	Mean	SD
NS1	157	6.08	1.38	5.90	1.38	5.55	1.55
NS2	157	5.26	1.65	4.89	1.68	4.50	1.82
NS3	160	6.04	1.19	5.89	1.26	5.46	1.53
NS4	160	6.09	1.20	5.88	1.31	5.66	1.54
L2S1	317	2.79	1.58	2.59	1.59	2.22	1.57
L2S2	317	3.02	1.73	2.89	1.74	2.47	1.67

5.5.1 Attitudes Towards Two Accents

RQ 3.1 addressed how L2 listeners’ attitudes speakers would differ between accents (i.e., ‘standard’ and L2) and passage types (i.e., academic monologue lectures and conversations). Table 5.14 gives descriptive statistics for L2 listeners’ attitudes to the two accents. As the table shows, ratings of NS accent and conversation were consistently higher than their counterparts (i.e., L2 accent, lecture) across all three attitude traits (i.e., ‘good pronunciation’, ‘happy to have as my instructor’, and ‘happy to have on a listening test’). Table 5.15 demonstrates ANOVA descriptives for the ANOVA analyses. The results from two-way independent ANOVA analyses conducted for ratings of ‘good pronunciation’ indicated a significant main effect for accent, with ratings of NS accent being significantly higher than L2 accent, $F(1, 1264) = 553.07, p < .0001, \omega^2 = .49$. Ratings of ‘good pronunciation’ was significantly higher for conversation compared to

lecture, $F(1, 1264) = 10.06, p < .01, \omega^2 = .03$. However, there was no significant interaction effect between accent and passage type on ratings of ‘good pronunciation’.

Table 5.14 *Descriptive Statistics for L2 listeners’ Attitudes Towards Two Accents on the Experimental Test*

	N	Good pronunciation		Happy to have as my instructor		Happy to have on a listening test	
		Mean	SD	Mean	SD	Mean	SD
Accent							
NS	634	5.87	1.41	5.64	1.48	5.29	1.68
L2	634	2.91	1.66	2.74	1.67	2.34	1.63
Passage type							
Conversation	634	4.65	2.03	4.44	2.09	4.03	2.18
Lecture	634	4.13	2.20	3.95	2.17	3.61	2.23

Table 5.15 *ANOVA Descriptive Statistics for L2 listeners’ Attitudes Towards Two Accents on the Experimental Test*

Accent	Passage type	Good pronunciation		Happy to have as my instructor		Happy to have on a listening test	
		Mean	SD	Mean	SD	Mean	SD
NS	Conversation	6.06	1.29	5.90	1.32	5.50	1.54
NS	Lecture	5.68	1.50	5.39	1.58	5.09	1.78
L2	Conversation	3.23	1.61	2.98	1.66	2.55	1.66
L2	Lecture	2.58	1.64	2.50	1.66	2.14	1.56

Similar results were revealed for ratings of ‘happy to have as my instructor’. That is, ratings of NS accent were significantly higher than those of their L2 counterpart, $F(1, 1264) = 552.98, p < .0001, \omega^2 = .46$. Ratings of conversation was, again, significantly higher in comparison to those of lecture, $F(1, 1264) = 16.58, p < .0001, \omega^2 = .02$, though, no significant interaction effect was found. The results for ratings of ‘happy to have on a listening test’ echoed those for ratings of ‘good pronunciation’ and ‘happy to have as my instructor’. Specifically, ratings of NS accent once again were significantly higher than L2 accent, $F(1, 1264) = 515.19, p$

$<.0001$, $\omega^2 = .45$, and ratings of conversation was significantly higher when compared with those of lecture, $F(1, 1264) = 10.38$, $p < .01$, $\omega^2 = .02$. Passage type * accent interaction, however, was not significant.

5.5.2 Attitudes Towards L2 Speakers

RQ 3.2 dealt with how L2 listeners' attitudes to speakers would differ among exposure conditions (audio-with-script exposure, audio-only exposure, and no exposure), between passage types (i.e., academic monologue lectures and conversations), and the two L2 speakers (i.e., L2 speaker1/ the Turkish speaker and L2 speaker2/the Ukrainian speaker). Descriptive statistics for L2 listeners' attitudes towards the two L2 speakers can be found in Table 5.16. In terms of exposure condition, the differences among the three conditions were very small for all three attitude traits (i.e., 'good pronunciation', 'happy to have as my instructor', and 'happy to have on a listening test'). As for passage type, the result was in line with that of attitudes towards the two accents. That is, conversation was associated with more positive attitudes than lecture was. With respect to speaker, the Ukrainian received relatively higher ratings on all three traits.

Table 5.16 Descriptive Statistics for L2 Listeners' Attitudes Towards L2 Speakers on the Experimental Test

	Good pronunciation			Happy to have as my instructor		Happy to have on a listening test	
	N	Mean	SD	Mean	SD	Mean	SD
Exposure condition							
Audio-script	204	3.00	1.53	2.76	1.51	2.39	1.50
Audio-only	222	2.87	1.56	2.77	1.65	2.31	1.55
No exposure	208	2.84	1.87	2.69	1.85	2.33	1.82
Passage type							
Conversation	317	3.23	1.61	2.98	1.66	2.55	1.66
Lecture	317	2.58	1.64	2.50	1.66	2.14	1.56
Speaker							

L2S1	317	2.79	1.58	2.59	1.59	2.22	1.57
L2S2	317	3.02	1.73	2.89	1.74	2.47	1.67

Table 5.17 shows ANOVA descriptives for the ANOVA analyses. The results of the three-way ANOVA for ratings of ‘good pronunciation’ showed that there were significant main effects of passage type, $F(1, 629) = 27.37, p < .0001, \omega^2 = .04$, and speaker, $F(1, 629) = 5.21, p < .05, \omega^2 = .01$ on listeners’ ratings of L2 speakers’ pronunciation. However, no significant main effect of exposure condition was found. The three-way ANOVA analyses for ratings of ‘happy to have as my instructor’ revealed a significant passage type * exposure condition * speaker interaction, $F(2, 622) = 3.60, p < .05, \omega^2 = .01$, although further pairwise differences between each level of all independent variables (or main effects) only identified one significant interaction. That is, listeners’ attitude ratings of the Ukrainian speaker under the audio-with-script exposure condition delivering one side of conversations ($M = 3.43, SD = 1.51$) were significantly higher than ratings of the Turkish speaker under the audio-only exposure condition delivering lectures ($M = 2.10, SD = 1.28$), $d = .95$. The results of the three-way ANOVA for ratings of ‘happy to have on a listening test’ were similar to those of ‘good pronunciation’. That is, there were significant main effects of passage type, $F(1, 629) = 11.71, p < .001, \omega^2 = .01$, and speaker, $F(1, 629) = 5.08, p < .05, \omega^2 = .01$. However, no significant main effect of exposure condition was detected. Together, these findings seem to suggest that more positive attitudes were associated with conversations and the Ukrainian speaker than with lectures and the Turkish speaker, and that the three exposure conditions did not seem to make a difference.

Table 5.17 ANOVA Descriptive Statistics for L2 Listeners' Attitudes Towards L2 Speakers on the Experimental Test

Exposure condition	Passage type	Accent	Good pronunciation		Happy to have as my instructor		Happy to have on a listening test	
			Mean	SD	Mean	SD	Mean	SD
Audio script	Conversation	L2S1	2.94	1.60	2.72	1.49	2.40	1.57
Audio script	Lecture	L2S1	2.67	1.34	2.45	1.46	2.12	1.44
Audio only	Conversation	L2S1	3.05	1.52	3.00	1.67	2.45	1.64
Audio only	Lecture	L2S1	2.22	1.28	2.10	1.28	1.71	1.12
No exposure	Conversation	L2S1	2.98	1.71	2.62	1.62	2.34	1.69
No exposure	Lecture	L2S1	2.74	1.88	2.54	1.91	2.17	1.80
Audio script	Conversation	L2S2	3.67	1.30	3.43	1.51	2.80	1.44
Audio script	Lecture	L2S2	2.75	1.67	2.49	1.42	2.25	1.49
Audio only	Conversation	L2S2	3.49	1.52	3.00	1.53	2.67	1.57
Audio only	Lecture	L2S2	2.73	1.64	2.90	1.87	2.35	1.63
No exposure	Conversation	L2S2	3.37	1.94	3.22	2.03	2.72	2.05
No exposure	Lecture	L2S2	2.36	1.87	2.45	1.82	2.14	1.76

6 DISCUSSION

This chapter begins by discussing the findings of RQ1, in the order of main measurements of exposure conditions, listening passages, and test items, followed by Rasch interaction/bias analyses at passage and item level. It then discusses the key findings in L2 listeners' perceived efficacy of the 60-second exposure (RQ2). It closes with main takeaway points in L2 listeners' attitudes towards speakers on the experimental test (RQ3).

6.1 RQ1

RQ 1 focused on how passage-level listening comprehension scores and item-level response would vary across four exposure conditions (i.e., audio-with-script, audio-only, no exposure, and baseline), among L2 listeners at three different proficiency levels (i.e., high, medium and low), and between two passage types (i.e., academic monologue lectures and conversations)/ three item types (i.e., main idea, explicit detail, implicit detail). In the subsections to follow, I discuss the relative difficulty of exposure condition at both passage and item level, followed by relative difficulty of listening passage and passage type, and relative difficulty of test item and item type. I then discuss Rasch interaction/bias analyses at passage level, and Rasch interaction/bias analyses at item level.

6.1.1 Relative Difficulty of Exposure Conditions at Passage and Item Level

The results of the study show that among the four conditions at both passage and item level the baseline was the easiest; the audio-with-script and audio-only exposure conditions helped somewhat, though are still more difficult than the baseline; and the no exposure condition was the most difficult. There are three points of interest in these findings. Firstly, the baseline condition was significantly easier than all three experimental conditions. This is, in fact, consistent with the trend observed in studies looking into the effects of unfamiliar accents on L2

listening comprehension. That is, L2 listeners often attained highest scores on listening passages (lectures mostly) delivered by speakers of native prestige accents, such as L1 American English speakers in Shin et al., 2021, and L1 American and British English speakers in Kang et al., (2019).

The most plausible explanation for such a finding would be that a 60-second exposure was simply not enough to erase the difficulty gap, or it is long-term exposure to a specific L2 speaker that is needed for test takers to score as well in the experimental conditions as in the baseline condition. However, the group of test takers who participated in the experiment had sparse exposure to non-prestige English varieties in general, L2 varieties included. This stands in stark contrast to their abundant and constant exposure to prestige English, most likely since the first day they started to learn English at school. Exposure of such goes beyond the walls of classroom, as it also the variety of English that dominate news, movies, TV shows, etc. in everyday life. More critically, L2 listeners, particularly those in EFL contexts have never been given a reason to listen to non-prestige varieties, let alone make effort to understand them. Testing could be such a reason shared by L2 listeners, in the sense that achieving a higher score on a test, a high-stakes proficiency test in particular, would certainly be likely to be a strong motivation for L2 learners as a whole. With a strong motivation and sufficient long-term exposure, the gap between performance on listening passages featuring prestige accents and that on the same listening passages featuring non-prestige accents would be closed. This is because L2 listeners in general were not born to be familiar with any variety of English, rather, they have been trained to, mainly through education, and then mass media. If they can be trained to understand prestige accents without much difficulty, there is no reason why they cannot be

trained to understand any non-prestige accents, such as the Turkish and Ukrainian accents used in the experiment.

Further, it is important to go beyond the test scores and to consider the inferences that can be made based on the scores obtained on the baseline condition versus the experimental conditions. Basically, this group of L2 listeners should be able to understand professors, TAs, and their fellow classmates who speak a prestige accent in the real world but are likely to have at least initial trouble understanding those who do not speak such a prestige accent. However, being able to shuttle between prestige and non-prestige accents is the ability that test takers need to be equipped with in the multidialectal context of higher education in English-speaking countries. This brings home the validity of current high-stakes proficiency tests, as test scores obtained from these high-stakes tests are supposed to allow valid inferences made about a test taker's ability to understand not just prestige accents but also diverse non-prestige accents that co-exist in the TLU domain, where these tests are specifically sampled from.

Secondly, the no exposure condition found to be the most challenging condition of all is in line with findings from L1 and L2 perceptual adaptation studies. In Witteman et al., (2013), for example, while completing a cross-modal priming task, native Dutch listeners with limited prior experience with German-accented Dutch was primed by the medium and weakly accented words, but not by the strongly accented words. As for L2 perceptual adaptation studies, L2 listeners who were randomly assigned to the no exposure condition in the experiment were almost under the same situation as L2 listeners who participated in Harding (2018), where they went into a listening passage featuring an unfamiliar and less intelligible accent without 'warning' or 'assistant' in any form. Hence, it is likely that L2 listeners on the no exposure condition experienced similar listening difficulties as L2 listeners in Harding (2018) did,

including understanding speaker's pronunciation, input coding, word recognition/segmentation, and attention being taken away from the listening passage (and onto accent itself). It is also possible that L2 listeners on the no exposure condition only adjusted to the unfamiliar accent after hearing the first half of the assigned listening passage, just as L2 listeners in Harding (2018). It is worth pointing out that it is not possible to make comparison with Harding (2018) in terms of L2 listeners' test performance, for this study did not report on this and it is a qualitative investigation, focusing on listening difficulties, strategy use, and adaptive behavior when L2 listeners encountering an unfamiliar L2 accent. Very importantly, the finding that no exposure was the most difficult, or more difficult than audio-with-script and audio-only conditions, provided further evidence that a 60-second exposure helped and that L2 perceptual adaptation took place, although it was not sufficient to bring scores into alignment with those for highly familiar speakers.

Last but not least, the audio-only condition being easier than the audio-with-script was somewhat surprising, given that the provision of lexical information in the form of shadowing-with-script in Hamada and Suzuki (2021) and in the form of written subtitles in Mitterer and McQueen (2009) was found more effective in facilitating L2 perceptual adaptation. Nonetheless, this result is probably not that unreasonable if digging deeper into the discrepancies between the two prior studies and the present one, in terms of operationalization of shadow-with-script in Hamada and Suzuki (2021) and tasks employed in these two studies versus in the current study.

Regarding the operationalization of the exposure, in Hamada and Suzuki (2021) the shadowing-with-script group was given a few minutes to study the script between the second and the third rounds (three consecutive rounds of shadowing practice in total). The primary purpose of this was to allow listeners to see the mismatch between what they thought they had heard and

what had actually been spoken, and thus activating their lexical analysis. By contrast, listeners in the present study were only allowed to listen to the exposure clip with the script on their computer screen once. In other words, they did not have the ‘luxury’ of studying the script between rounds and noticing the discrepancies. It is therefore possible that having an audio-with-script exposure may have the opposite effect on L2 listeners in the present work. On the one hand, the audio-with-script clip facilitated adaptation to the two L2 speakers’ speaking styles and accents, otherwise, some of the words or information in the exposure clip may be misunderstood or missed. On the other hand, having a script may result in over-dependence on the script and gave L2 listeners false confidence that they were able to understand every single word that the speaker uttered, though in reality, they did not really listen to or adapt to the speaker.

As for audio-only exposure, perhaps it gave listeners an opportunity to just focus on the characteristics of the speaker without being distracted by other modalities, printed text in this case. Therefore, in a way audio-only exposure forced L2 listeners to be concentrated on decoding words and making sense of the entire discourse on their own. Further, let us also not forget the fact that the ensuing listening passage was not accompanied with a script on listeners’ computer screen, which is in the exact same format as the audio-only exposure clip. Hence, it is likely that audio-with-script listeners might need extra time adjusting from with a script to without one, whereas audio-only listeners may have a smooth transition to following listening passage.

Turning now to how the tasks differed, Hamada and Suzuki (2021) used a 75-word dictation test, with the same items in the pre- and post-tests to evaluate progress, and Mitterer and McQueen (2009) asked participants to repeat back 160 audio excerpts, with half excerpts taken from the exposure material, while the other half being completely new. These two tasks, while different in modality (written versus oral), are essentially targeting word recognition only,

and therefore it makes perfect sense for the provision of lexical information along with audio to outperform its counterpart without such a provision. By contrast, it was TOEFL iBT listening comprehension tasks that were used in the present study. Listening comprehension, as described by Buck (2001), “is an inferential process, an ongoing process of constructing and modifying an interpretation of what the text is about, based on whatever information seems relevant at the time” (p. 29). In this sense, it is safe to say that tasks employed in the two prior studies are vastly different from the ones used in the present research. Further, what makes a listening passage difficult is not limited to unfamiliar and less intelligible accents, but also factors such as familiarity with the passage topic (or background knowledge), specialized vocabulary, L2 listening proficiency, and metacognitive strategies (see Bloomfield et al., 2010 for a detailed review). Together, the efficacy of the provision of a script was probably severely restricted in the context of TOEFL iBT listening comprehension tasks.

6.1.2 Relative Difficulty of Listening passage and Passage Type

Regarding the relative difficulty level of the four listening passages, Lecture 2 turned out to be the easiest while Conversation 1 was the most challenging. And comparing the difficulty level of the two passage types, conversation was .25 logits more difficult than lecture (-.37 versus -.12 logits). Such a result adds to the already mixed findings to studies conducted in the field of L2 assessment on the influence of passage type, and more importantly, contradicts the idea that conversations, which tend to have higher degree of orality than lectures, are presumably easier than lectures. This probably can be explained by the nature of conversation listening passages used in the present research, along with their associated test items. About the conversations, recall that both conversations cover academic topics such as speciation (a topic in the field of anthropology) in Conversation 1, and breathing and respiration at high altitudes in

Conversation 2, and they were purposefully chosen to allow easier comparison with monologic lectures. That is to say, although the two conversations inherently have higher degree of orality than the two lectures, since they are conversations after all, the topics involved in the two conversations are not at all ‘oral’, but rather, very ‘academic’.

As for the test items associated with the two conversation listening passages, there were three multi-select test items in conversations, whereas only one of such in lectures. This type of test items was found to be more difficult than test items that required mono select, as explained in Section 5.3.2. For these reasons, I would argue that using degree of ‘discourse orality’ only to gauge the difficulty level of conversations versus lectures might be too limited, as the degree of ‘topic orality’ and the difficulty level of associated test items are also likely to play an equally, if not more, important role in determining the listening comprehension difficulty of monologic versus dialogic discourse. This might in part explain why findings in this line of research in L2 assessment have been inconclusive. It is also important to keep in mind that in the conversations recorded by L2 speakers, listeners had to switch between two different accents (i.e., either Turkish or Ukrainian accent and GA accent). For some test takers this turned out to be more challenging than just listening to one L2 accent all along as in lectures, as they reported in their post-task interview. The main reason given was that listening to one L2 accent would, at least, allow longer time to adjust.

6.1.3 Relative Difficulty of Test Item and Item Type

The distribution of relatively more difficult (with negative logit values or measures) and easier items (with positive logit values or measures) was almost equal across the four listening passages, though three out of five items from Conversation 1 were among the top five most difficult test items. If grouping test items by item type, 60% (six out of 10 items) of explicit-

detail-oriented items were among the more difficult items, whereas about 63% (five out of eight items) of implicit-detail-oriented items was among the easier items. Items requiring understanding of gist was evenly split between the two categories. This pattern is reflected in item type measurement report, which shows that explicit detail and main idea test items were almost at the same difficulty level, which was about .20 logits more difficult than implicit detail items. Such results corroborate and contradict the previous studies' findings at the same time. Specifically, for explicit detail test items, the finding echoes the preceding studies which have found that items tapping into retrieval of specific details are likely to cause more differential performance with unfamiliar L2 accents (e.g., Harding, 2012; 2018; Shin et al., 2021). On the other hand, implicit detail items, which have frequently been found to be of greater difficulty than explicit detail items in listening comprehension tests in general (Bloomfield et al., 2010), were found to be easier than explicit detail items in the present research. This discrepancy could be attributed to the fact that the key information in five out of eight implicit detail items were listened to twice or three times, whereas that in both explicit detail and main idea items were listened to once only (see Table 4.8 for listen times of each test item). Taken together, therefore, it seems that being able to listen to the key information more than once outweighs the inherent difficulty level intended by test developers. This might have consequences for item design. If implicit test items are written with the expectation that they will be the more challenging items on the test, but when the key information is allowed to be listened to twice or more, tests may be mis-targeted to the population.

Another interesting pattern observed for explicit test items was that four out of six items require two answers, and one asks for the correct order of a process with four steps. It can thus be suggested that items requiring multi-select is more difficult. Given that all test items were

scored dichotomously, those who made one correct choice were treated the same way as those who made two incorrect choices. A closer look at one such item, item 1321 (i.e., Question 3 in Conversation 1), shows that about 19% of test takers either chose completely wrong answers or left it unanswered, 31% chose two correct answers, and the remaining 50% chose one correct answer. This is the pattern observed for all four multi-select items. This raises the question whether or not partial understanding should be rewarded. If not, is it fair to those who demonstrate at least partial understanding?

6.1.4 Rasch Interaction/Bias Analyses at Passage level

Three points of interest in the findings revealed by Rasch interaction/bias analyses conducted at passage level are, first, the bias detected, be it favoring the baseline condition or against it, was not prevalent, second, a 60-second exposure was most helpful for low-proficiency test takers, and also necessary for high-proficiency test takers, and third, L2 listening proficiency came into play in perceptual adaptation. On the first point, the instances of bias detected were mostly connected with lectures, Lecture 2 in particular, which happened to be the easiest of the listening passages. A possible explanation for this would be when a listening passage and test items are relatively easy, listening to a speaker with a highly familiar prestige accent delivering that listening passage would allow easier comprehension than listening to a speaker with an unfamiliar L2 accent, thus leading to higher scores. On the other hand, when a passage and its associated test items are relatively more challenging, listening to a speaker with a familiar prestige accent may not facilitate listening comprehension to the same degree as with a relatively easy listening passage. However, this interpretation should be taken with caution, as only four listening passages were used in the experiment.

On the second point, a 60-second exposure, particularly in the form of audio only, was most useful for low-proficiency L2 listeners, which was found at both passage and item level interaction between exposure condition and L2 listening proficiency. A 60-second exposure, regardless of the format, was also necessary for high-proficiency L2 listeners, again at both passage and item level. By contrast, such an exposure phase did not seem to make any difference to medium-level L2 listeners. It is, in fact, very difficult to explain these mixed findings, which merits further research, preferably replication studies. However, such mixed findings might support Harding's (2018) assumption that L2 listener proficiency has a critical role to play in L2 perceptual adaptation, although his operationalization of exposure is not directly comparable to that of the present study.

On the third point, the facts that the direction of bias associated with lectures was different depending on L2 listener proficiency, and that no bias was detected for conversation (Conversation 2 only) until L2 listening proficiency was factored in the interaction analyses seem to provide further support for Harding's (2018) assumption. Regarding the direction of bias, L2 listeners in the current study could be collapsed into two major camps: high- and medium-proficiency listeners versus low-proficiency listeners, in the sense that Lecture 2 was significantly easier under the baseline condition than the experimental conditions for both high- and medium-proficiency listeners, whereas Lecture 1 was significantly easier under the experimental conditions than the baseline condition for low-proficiency listeners. Similar pattern emerged for interactions among exposure condition, passage type and listening proficiency. That is, while all bias were related to lecture, the direction of bias was opposite for high- and medium-level test takers versus low-level test takers, with the baseline condition easier than the experimental conditions for high- and medium-level test takers, whereas the no exposure

condition easier than the baseline condition for low-level test takers. This, again, is hard to explain, given that the issues on how and to what extent L2 listeners' proficiency would affect their listening comprehension of L2 varieties of English have not been well-addressed by the literature, let alone when an exposure phase is factored in, but it might be related to what Kang et al., (2019) suggested that compared to low-proficiency L2 listeners, their high-and medium-proficiency counterparts exhibit more sensitivity towards English varieties when responding to less comprehensible speech or L2 speaker with lower intelligibility, although this does not explain why low-proficiency test takers did better when listening to the two L2 speakers in comparison to speakers of General American accents. Note that low-comprehensibility speakers were rated between 4.5 to 5 on a 7-point Likert scale in Kang et al., (2019). If roughly translating to the scale used in the current study, 3 on a 5-point scale, the two L2 speakers could be classified as low- comprehensibility speakers in their study.

6.1.5 Rasch Interaction/Bias Analyses at Item Level

The findings yielded by interaction analyses conducted between exposure condition and test item almost aligned with what were uncovered by interaction analyses conducted between exposure condition and listening passage. That is, amongst the six test items that were detected with significant bias, the two items that were easier under the baseline were both from Lecture 2, whereas the other four test item that were easier under one of the exposure conditions came from the other three listening passages. Further, it seems that more difficult test items are more likely to show significant bias, considering that five out of these six items belonged to the more difficult ones, although this does not mean that these items were easier under the baseline. In fact, four of them were easier under either the audio-with-script or audio-only condition than under the baseline condition. Moreover, contrary to previous studies (e.g., Harding, 2012; Shin et

al., 2021) which found that detail-oriented items are more vulnerable to unfamiliar L2 accents, explicit items in the current study turned out to be easier under L2 accent conditions when provided with an exposure. These findings together seem to provide further evidence of some perceptual adaptation taking place.

In a similar vein, item-level Rasch analyses among exposure condition, test item and proficiency level also to a large extent reflected those among exposure condition, listening passage and proficiency level, in that half of the items (i.e., three) showing bias favoring the baseline for high-proficiency listeners and all three items that biased in favor of the baseline for medium-level listeners were from Lecture 2 (with two shared items, 4111 and 4521), whereas four out of seven items showed bias in the direction of experimental exposure conditions for low-proficiency listeners were from Lecture 1. This again demonstrated that instances of significant bias in the direction of baseline were, in most cases, connected with Lecture 2 and high- and medium-proficiency test takers. By contrast, instances of significant bias in the direction of experimental exposure condition were mostly from the other three listening passages and associated with low-proficiency test takers.

The most interesting finding related to the interactions among exposure condition, item type and proficiency was that significant biases/interactions between the baseline condition and the experimental conditions were only identified for high- and low-proficiency L2 listeners, and, again, the direction is the opposite, particularly for explicit detail items, which was easier under the baseline for high-proficiency test takers while easier under the audio-only condition for low-proficiency test takers. This lends further support to Harding (2018)'s assumption that L2 listening proficiency has a role to play in L2 adaptation.

6.2 RQ2

RQ2 addressed L2 listeners' perceived efficacy of the 60-second exposure in four aspects, namely, 'helped adapt to speaking style', 'helped adapt to accent', 'made me less anxious', and 'was long enough'. Three points of interest exist in the findings. Firstly, passage type made a difference to the ratings of all four aspects, with having an exposure considered to be more useful for listening to conversations than to lectures. There are two possible explanations for this result. Firstly, conversations were harder than lectures as shown by results from RQ1, and therefore an exposure was more necessary for listening to conversations in comparison to lectures. Then, there is less chance for L2 listeners to adapt to an unfamiliar L2 accent in a conversation than in a lecture, because they are not hearing it all of the time. This again, makes the extra exposure crucial. Nonetheless, this finding is reassuring in the sense that the exposure clip was perceived to help L2 listeners more while listening to conversations, especially in terms of adjusting to L2 speakers' speaking style and accent, and reducing anxiety.

Then, neither exposure condition nor speaker (i.e., the two L2 speakers) was found to have a significant effect on the ratings. For exposure condition, again, the result was somewhat unexpected, as the audio-with-script exposure was expected to be rated more useful than the audio-only exposure, based on the findings from L2 perceptual adaptation studies, but it is in line with the result regarding the difficulty level of exposure conditions found for RQ1, with audio-only being the easiest among the three experimental conditions at both passage and item level. With respect to the two speakers, the fact that only two significant interaction effects were related to speaker, in effect, was much welcomed, since these two speakers were selected precisely because of their comparability. In addition, such a result also aligned with their comparability in terms of difficulty level revealed by Rasch analysis (see Section 4.9.4.1.3).

Finally, ‘the 60-second recording was long enough for me to get used to the professor's accent’ received the lowest ratings across the board. This result is not unanticipated. From the perspective of a test taker, the longer the exposure, the better. However, a 60-second exposure was chosen by taking into consideration both the finding from the pilot study and practicality (or feasibility) in a high-stakes testing situation, although it was found to be most useful for low-proficiency L2 listeners.

6.3 RQ3

The last RQ dealt with L2 listeners’ attitudes towards the speakers involved in the experimental test, including both TOEFL iBT speakers and L2 speakers. Regarding the attitudes towards two accents, NS accent being rated much higher than L2 accent was consistent with findings from studies investigating listeners’ attitudes towards pronunciation. That is, prestige native accents are normally rated more positively on language-focused traits, such as ‘good pronunciation’ (Lindemann & Campbell, 2018). However, this finding is contrary to that of Harding (2011) who found that L2 listeners generally held positive attitude towards L2 speakers. This discrepancy may be explained in part by the speakers and listeners used in Harding (2011). For the former, L2 speakers in his study were highly intelligible speakers; for the latter, listeners were studying in Australia at the time of data collection so that presumably they already had experience with speakers from diverse L1 backgrounds or sociolinguistic characteristics of the academic context.

With regard to attitudes towards L2 speaker, what is interesting is that no significant differences were found among the three exposure conditions. That is to say, with or without a 60-second exposure made no difference to L2 listeners’ attitudes towards L2 speakers. However, this really is not surprising, and it is probably not reasonable to expect a 60-second exposure to

work its magic. After all, the deep-rooted ‘standard’ language ideology has been perpetuated by generations of persistent reinforcement of achieving so-called ‘native-ness’ in English language teaching practices in both English as a foreign (EFL) and English as a second language (ESL), as well as by the “the absence of near absence of so-called non-standard varieties of English in teaching materials (e.g., coursebooks) as well as on tests of English proficiency, which privilege American or British English varieties” (Kang et al., 2018, p.2). Regarding more positive attitudes towards the Ukrainian speaker than the Turkish speaker, it might be related to the study design, where the Ukrainian speaker was placed after the Turkish speaker under all 12 conditions. After the initial ‘shock’ hearing the Turkish speaker, L2 listeners seemed to ‘calm down’ or accept the ‘cruel fact’ when it comes to the Ukrainian speaker. However, without a closer inspection of interview data and L2 listeners’ comments on these two L2 speakers, caution must be applied in this interpretation.

7 CONCLUSIONS

This chapter presents the conclusions drawn from the study. It begins with a summary of the key findings, then moves to a discussion of the implications of these findings. Following the implications is the limitations of the current study and recommendation for future work to address the limitations laid out and explore the topics of this research in more depth.

7.1 Summary and Implications

This dissertation project contributes to a small but growing body of research dedicated to diversifying English accents represented on high-stakes listening assessment. Instead of using L2 speakers with high I/C as recommended by scholars and researchers in the field of L2 listening assessment, the present work took a different approach by exploring the possibility of using L2 speakers with low familiarity and relatively lower I/C to the targeted L2 listeners, because this approach would better represent accent strengths in the TLU domain and expand the construct of listening comprehension to become more multidialectal, when compared with using L2 speakers with high I/C. Taking inspiration from both L1 and L2 perceptual adaptation studies, I conducted an L2 perceptual adaptation study in the context of a simulated TOEFL iBT listening test, using one Turkish speaker and one Ukrainian speaker, whose accents were generally unfamiliar to the targeted L1 Chinese listeners and whose rated intelligibility at intermediate level. I then investigated whether the assistance of a 60-second exposure to the two L2 speakers' accents would improve comprehension of texts delivered by them. The key findings are: (1) the baseline condition (or listening passages recorded by speakers with prestige native accents (was significantly easier than all the experimental conditions (or the same listening passages recorded by the Turkish and Ukrainian speakers) across all listening passages, although bias detected between the native and L2 versions was particularly pronounced on lectures, Lecture 2 in

particular and test items associated with Lecture 2; a 60-second exposure was most useful for low-proficiency L2 listeners, and it was also necessary for high-proficiency L2 listeners; (2) a 60-second exposure was considered to be more useful for listening to L2 speakers participating in conversations than delivering lectures; (3) L2 listeners' attitudes were overwhelmingly negative towards L2 speakers, whose average favorability ratings were only half of those received by speakers with 'standard' American accents.

It is important to keep in mind that these findings were elicited from a group of L2 listeners, 83% of whom reported having no prior experience either preparing for or sitting for a TOEFL iBT test. Further, no participants were informed beforehand of what was in store for them. In other words, they had no idea that they would be tested on unfamiliar and less intelligible L2 English accents. Moreover, while taking the experimental test, this group of participants were not under the kind of pressure of real-world TOEFL iBT test takers nor did they have same level of desire to achieve a higher score. Imagine what the results would be when real-world TOEFL iBT test takers know ahead of time that TOEFL iBT listening test incorporates unfamiliar and less intelligible L2 English accents.

There are significant implications of these findings for both test developers of high-stakes listening tests and TESOL practitioners. For test developer the result that the baseline condition was the easiest, and the 60-second exposure helped somewhat, but not enough to make it equivalent to the baseline condition could mean two opposite implications. On the one hand, they can use this as one more added piece of evidence not to include L2 accents as used in the current study in their tests. However, this argument is only circular, in that L2 listeners' poorer performances under the experimental conditions (or listening passages delivered by the Turkish and Ukrainian speakers) than the baseline condition (or listening passages recorded by native

speakers with prestige accents) was mostly likely to be the result of them never being tested on L2 accents akin to these two L2 speakers. That is to say, if test developers of high-stakes listening tests decided not to incorporate unfamiliar and less intelligible L2 speakers, test takers performance on listening passage featuring these speakers would not likely to equal their performance on listening passage featuring speakers with familiar and prestige accents. An important question arises: Is it a reasonable concern on the part of test developers when test takers' worse performance on listening passages featuring unfamiliar and less intelligible L2 speakers is likely to be test takers' lack of ability to adapt to those speakers, whom they are likely to encounter in the TLU domain?

On the other hand, such a finding could be the exact reason why it is critical to for test developers to incorporate L2 accents into their tests, given the inferences that could be made about this group of L2 listeners vastly different performance under the three experimental conditions versus the baseline condition, that is, their performance on the same listening passage featuring the Turkish and Ukrainian L2 speakers, regardless of exposure conditions, was much poorer than their performance on that listening passage featuring speakers with prestige accents. In relation to the current practice of using prestige accents only on high-stakes proficiency test, test takers, regardless of proficiency levels, are likely to have difficulty understanding non-prestige accent in the TLU domain, higher education in English-speaking countries, where they, in fact, need to cope with, rapidly attune to unfamiliar and less intelligible English accents, ultimately to successfully negotiate meanings. This points directly to the validity to the current high-stakes listening proficiency test.

For these reasons, I would argue that it would be a wise decision for test developers to include L2 accents used in the current study in their tests. One promising and practical way

would be to incorporate L2 speakers of relatively low I/C and general unfamiliarity to the targeted test takers into conversations, where they record one side, along with a provision of a 60-second audio-only exposure to those L2 speakers. This is because very few bias was detected for conversations between the baseline and experimental conditions. This approach could help enhance the interactiveness and bolster the construct validity of high-stakes listening tests whose TLU domain requires L2 listeners to deal with accents that they are not familiar with and that are not highly intelligible, considering the increasing number of international professors, teaching assistant, and students in academic settings. Specifically, using unfamiliar accents might have the potential to solve one thorny issue facing high-stakes test developers when it comes to accent or speaker selection, that is, background information about the test-taker population is normally unknown in advance (Harding, 2012, p. 177). Fortunately, though, the vast majority of test takers of TOEFL iBT or any other high-stakes English proficiency test are likely to take the test in their home country, meaning that in each region or country, test taker population is relatively homogeneous, and that by surveying the potential test-taker population, test developers would be able to gather information regarding accent familiarity. With this data, they could select unfamiliar accent(s) depending on different regions/countries, while keeping listening passages and test items the same across regions. That way, the cost of developing materials is not likely to be driven up too much, despite the fact that speakers from different L1 backgrounds would be needed to cater the needs of different regions/countries in terms of unfamiliar accent(s).

Then, many testing agencies claim that part of the purpose of their test is to have a positive impact on educational systems, and on classroom teaching and learning. One way to promote positive washback on educational systems would be to hire L2 speakers with lower intelligibility and familiarity to the targeted L2 listeners alongside speakers with ‘standard’

accents (who can deliver monologue or interactive lectures) on high-stakes listening tests. That way, test takers would be motivated to enhance their multidialectal skills, even if their sole incentive is to obtain higher scores from high-stakes English proficiency tests. Over time, more diverse accents can be used without advantaging or disadvantaging some test takers, which in turn better ascertains that the test score inferences made about a test taker's listening ability could be representative of test takers' performance in English-medium institutions of higher education. Of course, it is also hoped that positive washback could go beyond the realm of language testing such that test takers would develop familiarity with a wide variety of English accents in their day-to-day life and appreciate the diversity of English, which would hopefully break down linguistic racism, defined as "the ideologies and practices that are utilized to conform, normalize and reformulate an unequal and uneven linguistic power between language users" (Skutnabb-Kangas, 2015, p.). No variety of a language is inherently better than others, and standard varieties of a language are 'better' only in a social sense. This is known and agreed by linguists, not the general public. By incorporating L2 speakers into high-stakes tests, test developers may help to communicate this with different stake-holder groups, test takers included.

Turning now to TESOL practitioners, while it is easy to defend language teachers for not including listening materials with non-prestige accents into curriculum on the ground that teachers will not teach what will not be tested, in this case, the ability to comprehend L2 varieties of English, language teaching and learning should not be limited to, nor should it be tied to language testing only. Learning a language other than our mother tongue, at its core, is about communicating with speakers of that language or even the world, and yet common wisdom tells us that textbook English, French, or Spanish is clearly not enough. Also, communication is not just about making ourselves understood, but also about being able to understand others. In

relation to English, the reality is that it is spoken far more by non-native speakers than by native speakers (e.g., Jenkins, 2006). Also being a native speaker does not automatically mean that person speaks a 'standard' accent. In this sense, while learning 'standard' accent would allow learners to communicate themselves with most people, as this is the most learned variety of English, it may not necessarily allow them to be able to understand most English users. In other words, teaching 'standard' English is a necessary evil, especially considering the limited class time, but learners would benefit more if exposed to non-prestige varieties of English. That way, they would be made aware of the existence of a vast array of English varieties, and most importantly, the legitimacy of every single variety of English, since no variety is inherently better than others in linguistic terms.

One concern that English language teachers may have would be that learners might be led astray in terms of pronunciation if exposed to non-prestige accents. Whether this is true or not is still an empirical question, though an obvious anecdotal counter argument would be if it is so easy for L2 learners to pick up non-standard accent features, then it would be equally easy for them to pick up 'standard' ones, meaning that L2 speakers should generally speak a 'standard' accent that they have been taught and exposed to, and yet this is hardly the case, as we all know. L2 accents are mostly influenced by L1, particularly when learners past the prime age in terms of mastering pronunciation. No one should be blamed for being able to speak their L1 too well. Knowing the difficulties facing different L1 groups when learning English should also help learners to be prepared with the pronunciation features that they may hear from L2 learners from a specific L1 background, so that they would have less trouble understanding those learners. Further, knowing that we all have difficulties learning English, unique and shared ones, may help learners sympathize with each other, which, in turn, could build solidarity. In short, it is highly

recommended that English language teachers bring in ‘organic’ English, or more precisely, Englishes, into their classroom by systematically incorporating listening materials with non-standard accents into their curriculum, just for the sake of teaching and learning English itself.

7.2 Limitations and Future Research

There are several limitations to the present study. To start with, I did not fully analyze post-task interview data elicited from 160 of the participants. This set of data focused on L2 listeners’ test-taking experiences, along with all their perceptions of the 60-second adaptation phase, and attitudes towards all speakers. A natural progression of the present work is therefore to analyze these qualitative data, interview data in particular. This would not only add to Harding’ (2018) qualitative analysis of processing difficulties, strategy use, perceptions of difficulty, with a much large sample size, but also provides a deeper insight into how a 60-second exposure clip help or not help L2 listeners to adjust to unfamiliar and less intelligible accents, more importantly, how to better operationalize the exposure clip to achieve a better result. A related step forward would be carrying out qualitative item analysis, which would dive deeper into possible factors related to both listening passages and test items (including both stem and choices) that contribute to bias.

Then, while one contribution of this study has been to confirm Harding’s (2018) assumption that L2 listening proficiency did play a role in adaption and adds to our understanding of the role of L2 listening proficiency level in listening to unfamiliar and less intelligible accents, some of the relevant findings in the present work were difficult to explain, in part due to limited existing literature on the potential impacts of listening proficiency on L2 comprehension of L2 accents. As such, more studies looking into the role of listening proficiency are needed before more solid conclusions can be drawn. Further, often times how participants

were classified into proficiency levels differs from study to study, thus making it difficult to compare results across studies. This could be attributed to the fact that very few free and high-quality listening proficiency tests have been made available to researchers in our field, leaving researchers having to resort to whatever is at disposal. Hence, as a field, we need to make an effort to provide such resources to researchers, which, in turn, benefit our field as a whole.

Moreover, listeners who participated in the current study were limited to L1 Chinese listeners, who had very limited exposure to non-prestige varieties. As such, the findings did preclude from offering insights into L2 listeners who have been exposed to various accents. In this sense, it would be interesting and important to carry out research working with such L2 listeners to see whether and how their performance and attitudes towards L2 speakers would differ from the group of listeners participating in the present work, the vast majority of whom almost had no experience with non-prestige accents, and whose attitudes towards both Turkish and Ukrainian speakers were overwhelmingly negative. The assumption tested is the potential positive washback followed by the inclusion of L2 accents on high-stakes English proficiency test, which presumably would encourage test takers to develop familiarity with various English accents in their daily life.

Finally, to allow direct comparison between the results of the present work with TOEFL iBT test, all efforts were made to simulate the TOEFL iBT test in terms of test administration and delivery. One such effort was to employ the recordings of the spoken input on TOEFL iBT test, which are audio-only, although still pictures of the speakers are presented visually to serve as context visuals. While using audio-only stimuli is a long-standing tradition in L2 listening assessment, this practice would in fact result in a threat to the construct validity of this test, due to construct underrepresentation (Messick, 1996), in the sense that virtually all the speaking

contexts presented in TOEFL iBT test concerns the listener being able to see the speaker in a real-life situation, and the TLU domains of this test almost always are lecture halls and classrooms, service encounters, or two-person conversations in which the speakers are in close contact, and hence would be able to see each other (as opposed to listening to the radio or listening to a phone call). If video input is decided to use for a given test, and if L2 speakers are also decided to be included in this test, it is then necessary to conduct interdisciplinary studies, drawing on research methods from two lines of research: the use of audio-visual texts on L2 listening tests and language attitudes. This type of interdisciplinary study is important, in that while a speaker's personal attributes such as gender, ethnicity, and social class have been addressed extensively in language attitudes literature and found to make a whole difference to speech comprehension and accent ratings, these personal attributes have not yet been controlled for nor investigated as a variable in language testing studies.

REFERENCES

- Abeywickrama, P. (2013). Why not non-native varieties of English as listening comprehension test input?. *RELC Journal*, 44(1), 59-74.
- Alderson, J. C., & Wall, D. (1993). Does washback exist?. *Applied linguistics*, 14(2), 115-129.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3), EL174-EL180.
- Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 114(3), 1600-1610.
- Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension*. MARYLAND UNIV COLLEGE PARK.
- Borg, S., & Liu, Y. (2013). Chinese college English teachers' research engagement. *Tesol Quarterly*, 47(2), 270-299.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729.
- Brown, J. D. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly*, 11(1), 5-26.
- Buck, G. (2001). *Assessing listening*. Cambridge University Press.

- Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly: An International Journal*, 3(3), 229-242.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647-3658.
- Crystal, D. (2000). Emerging Englishes. *English Teaching Professional*, 14(1), 3-6.
- Dai, D. W., & Roever, C. (2019). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Language Assessment Quarterly*, 16(1), 64-86.
- Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes*, 22(4), 571-584.
- Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*.
- Eger, N. A., & Reinisch, E. (2019). The role of acoustic cues and listener proficiency in the perception of accent in nonnative sounds. *Studies in Second Language Acquisition*, 41(1), 179-200.
- Elder, C., & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics*, 31(3), 34-1.
- Field, J. (2013). Cognitive validity. In A. Geranpayeh, & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77-151). Cambridge University Press.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: Sage.

- Galloway, N., & Numajiri, T. (2020). Global Englishes language teaching: Bottom-up curriculum implementation. *Tesol Quarterly*, 54(1), 118-145.
- Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, 34(1), 65-87.
- Gu, L., & So, Y. (2015). Voices from stakeholders: What makes an academic English test ‘international’?. *Journal of English for Academic Purposes*, 18, 9-24.
<https://doi.org/10.1016/j.jeap.2014.10.002>
- Hamada, Y., & Suzuki, S. (2021). Listening to Global Englishes: Script-assisted shadowing. *International Journal of Applied Linguistics*, 31(1), 31-47.
<https://doi.org/10.1111/ijal.12318>
- Hamid, M. O. (2014). World Englishes in international proficiency tests. *World Englishes*, 33(2), 263-277.
- Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Peter Lang.
- Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing*, 29(2), 163-180.
- Harding, L. (2014). Communicative language testing: Current issues and future research. *Language assessment quarterly*, 11(2), 186-197.
- Harding, L. (2018). Listening to an unfamiliar accent: Exploring difficulty, strategy use, and evidence of adaptation on listening assessment tasks. In G.J. Ockey & E.Wanger (Eds.), *Assessing L2 Listening: Moving towards Authenticity*. (pp.97-112). Amsterdam: John Benjamins.

- Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of phonetics*, 36(4), 664-679.
- Isaacs, T., & Rose, H. (2022). Redressing the balance in the native speaker debate: Assessment standards, standard language, and exposing double standards. *TESOL Quarterly*, 56(1), 401-412.
- Lindemann, S., & Campbell, M. A. (2018). Attitudes towards non-native pronunciation. In *The Routledge handbook of contemporary English pronunciation* (pp. 399-412). Routledge.
- Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). TOEFL 2000 framework. *Princeton, NJ: Educational Testing Service*.
- Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT journal*, 60(1), 42-50.
- Kang, O., Moran, M., Ahn, H., & Park, S. (2019). Proficiency as a Mediating Variable of Intelligibility for Different Varieties of Accents. *Studies in Second Language Acquisition*, 1-17.
- Kang, O., Thomson, R. I., & Moran, M. (2018). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*.
- Kang, O., Thomson, R.I., & Moran, M. (2019). The effects of international accents and shared first language on listening comprehension tests. *TESOL Quarterly*, 53(1), 56-81.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.
- Lambert, S. (1992). Shadowing. *Meta*: 37(2), 263-273.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press

- Linacre, J. M. (2022) Facets computer program for many-facet Rasch measurement, version 3.84.0. Beaverton, Oregon: Winsteps.com
- Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, 63(3), 567-594.
- Lowenberg, P. H. (2002). Assessing English proficiency in the expanding circle. *World Englishes*, 21(3), 431-435.
- Llurda, E. (2004). Non-native-speaker teachers and English as an International Language. *International Journal of Applied Linguistics*, 14(3), 314-323.
- Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL quarterly*, 36(2), 173-190.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.
- Messick, S. (1996). Validity and washback in language testing. *Language testing*, 13(3), 241-256.
- McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment*. Oxford University Press.
- Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*, 4(11), e7785.
- Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds). *Phonology and second language acquisition*, 193-218. Amsterdam: John Benjamins.

- Munro, M. J., & Derwing, T. M. (1995a). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, 38(3), 289-306.
- Munro, M. J., & Derwing, T. M. (1995b). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 45(1), 73-97.
- Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in second language acquisition*, 28(1), 111-131.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2), 204-238.
- Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, 37(5), 693-715.
- Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening*, 30(1-2), 84-98.
- Ockey, G. J., & Wagner, W. (2018). *Assessing L2 listening: Moving toward authenticity*. Philadelphia, PA: John Benjamin.
- Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *Tesol Quarterly*, 35(2), 233-255.
- Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, 58, 1-21.
- Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, 26(1), 87-98.
- Rajagopalan, K. (2010). The soft ideological underbelly of the notion of intelligibility in discussions about 'World Englishes'. *Applied Linguistics*, 31(3), 465-470.

- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2), 539.
- Reinisch, E., & Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *The Journal of the Acoustical Society of America*, 132(2), 1165-1176.
- Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York, NY: McGraw-Hill.
- Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6), 1207-1218.
- Shin, S. Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language Testing*, 38(4), 580-601.
- Shohamy, E. G. (2006). *Language policy: Hidden agendas and new approaches*. Psychology Press.
- Sidasas, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, 125(5), 3306-3316.
- Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, 4(3), 333-342.

- Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of memory and language*, 60(4), 487-501.
<https://doi.org/10.1016/j.jml.2009.01.001>
- Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 58(1), 1–21.
- Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *ELT journal*, 60(1), 51-60.
- Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10(2), 89-101.
- Wagner, E. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 103–123). London, UK: Bloomsbury.
- Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.
- Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, 75(3), 537-556.
- Wu, Y. A. (2001). TESOL in China: Current challenges. *TESOL Quarterly*, 35(1), 191-198.
- Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, 143(4), 2013-2031.

APPENDICES

Appendix A

Appendix A.1 Speaker Audition Survey

https://gsu.qualtrics.com/jfe/form/SV_5mvKomjm6kDXt0a

Appendix A.2 Pre-test

https://gsu.qualtrics.com/jfe/form/SV_bwOMgYCwLndXBno

Appendix A.3 Experimental Test

https://gsu.qualtrics.com/jfe/form/SV_bEkwXm7OUCmgB9Q

Appendix B

Appendix B.1 Pairwise Bias Report for Exposure Condition and Listening Passage

Target listening passage	Target Measure +	S.E.	Exposure condition	Target Measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Conversation 1	-0.50	0.17	audio script	-0.60	0.17	audio only	0.11	0.24	0.45	104	0.655	0.09
Conversation 1	-0.50	0.17	audio script	-0.74	0.17	no exposure	0.24	0.24	1.00	102	0.319	0.20
Conversation 1	-0.50	0.17	audio script	-0.70	0.10	baseline	0.21	0.19	1.09	87	0.280	0.23
Conversation 1	-0.60	0.17	audio only	-0.74	0.17	no exposure	0.13	0.24	0.56	105	0.579	0.11
Conversation 1	-0.60	0.17	audio only	-0.70	0.10	baseline	0.10	0.19	0.53	91	0.597	0.11
Conversation 1	-0.74	0.17	no exposure	-0.70	0.10	baseline	-0.03	0.20	-0.16	86	0.875	-0.03
Lecture 1	-0.11	0.14	audio script	-0.14	0.14	audio only	0.02	0.20	0.11	104	0.910	0.02
Lecture 1	-0.11	0.14	audio script	-0.09	0.14	no exposure	-0.03	0.20	-0.13	102	0.898	-0.03
Lecture 1	-0.11	0.14	audio script	-0.41	0.08	baseline	0.29	0.16	1.80	86	0.075	0.39
Lecture 1	-0.14	0.14	audio only	-0.09	0.14	no exposure	-0.05	0.20	-0.24	105	0.809	-0.05
Lecture 1	-0.14	0.14	audio only	-0.41	0.08	baseline	0.27	0.16	1.68	92	0.097	0.35
Lecture 1	-0.09	0.14	no exposure	-0.41	0.08	baseline	0.32	0.16	1.96	88	0.053	0.42
Conversation 2	0.10	0.16	audio script	0.12	0.16	audio only	-0.02	0.23	-0.08	102	0.937	-0.02

Target listening passage	Target Measure +	S.E.	Exposure condition	Target Measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Conversation 2	0.10	0.16	audio script	-0.12	0.16	no exposure	0.22	0.23	0.95	98	0.346	0.19
Conversation 2	0.10	0.16	audio script	-0.21	0.09	baseline	0.31	0.19	1.64	83	0.105	0.36
Conversation 2	0.12	0.16	audio only	-0.12	0.16	no exposure	0.24	0.23	1.05	103	0.297	0.21
Conversation 2	0.12	0.16	audio only	-0.21	0.09	baseline	0.33	0.18	1.80	97	0.075	0.37
Conversation 2	-0.12	0.16	no exposure	-0.21	0.09	baseline	0.09	0.19	0.47	84	0.638	0.10
Lecture 2	-0.37	0.14	audio script	-0.27	0.13	audio only	-0.10	0.20	-0.51	102	0.609	-0.10
Lecture 2	-0.37	0.14	audio script	-0.07	0.14	no exposure	-0.30	0.20	-1.51	98	0.135	-0.31
Lecture 2	-0.37	0.14	audio script	0.30	0.08	baseline	-0.67	0.16	-4.11	82	0.000	-0.91
Lecture 2	-0.27	0.13	audio only	-0.07	0.14	no exposure	-0.20	0.20	-1.03	104	0.304	-0.20
Lecture 2	-0.27	0.13	audio only	0.30	0.08	baseline	-0.57	0.16	-3.63	96	0.000	-0.74
Lecture 2	-0.07	0.14	no exposure	0.30	0.08	baseline	-0.37	0.16	-2.27	84	0.026	-0.50

Appendix B.2 Pairwise Bias Report for Exposure Condition and Proficiency Level at Passage Level

Target proficiency	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	0.05	0.13	audio script	-0.12	0.13	audio only	0.17	0.19	0.90	127	0.368	0.16
High	0.05	0.13	audio script	-0.21	0.12	no exposure	0.26	0.18	1.43	133	0.155	0.25
High	0.05	0.13	audio script	0.15	0.08	baseline	-0.11	0.15	-0.69	123	0.491	-0.12
High	-0.12	0.13	audio only	-0.21	0.12	no exposure	0.09	0.18	0.48	130	0.632	0.08
High	-0.12	0.13	audio only	0.15	0.08	baseline	-0.27	-0.16	1.74	116	0.085	0.32
High	-0.21	0.12	no exposure	0.15	0.08	baseline	-0.36	-0.15	-2.42	136	0.017	-0.42
Medium	-0.17	0.12	audio script	-0.10	0.12	audio only	-0.07	0.17	-0.42	153	0.675	-0.07
Medium	-0.17	0.12	audio script	0.18	0.13	no exposure	-0.35	-0.18	-1.98	140	0.050	-0.33
Medium	-0.17	0.12	audio script	0.02	0.07	baseline	-0.19	-0.14	-1.33	126	0.186	-0.24
Medium	-0.10	0.12	audio only	0.18	0.13	no exposure	-0.28	-0.17	-1.62	140	0.108	-0.27
Medium	-0.10	0.12	audio only	0.02	0.07	baseline	-0.11	0.13	-0.85	138	0.394	-0.14
Medium	0.18	0.13	no exposure	0.02	0.07	baseline	0.17	0.15	1.13	107	0.260	0.22
Low	0.18	0.14	audio script	0.24	0.13	audio only	-0.07	0.19	-0.34	131	0.731	-0.06
Low	0.18	0.14	audio script	0.04	0.14	no exposure	0.14	0.20	0.67	129	0.503	0.12

Target proficiency	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	0.18	0.14	audio script	-0.15	0.08	baseline	0.33	0.16	2.03	98	0.045	0.41
Low	0.24	0.13	audio only	0.04	0.14	no exposure	0.20	0.19	1.05	143	0.294	0.18
Low	0.24	0.13	audio only	-0.15	0.08	baseline	0.40	0.15	2.63	134	0.010	0.45
Low	0.04	0.14	no exposure	-0.15	0.08	baseline	0.20	0.16	1.23	113	0.223	0.23

Appendix B.3 Pairwise Bias Report for Exposure Condition, Listening Passage, and Proficiency Level

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	Conversation 1	-0.56	0.24	audio script	-0.50	0.29	audio only	-0.06	0.38	-0.16	31	0.874	-0.06
High	Conversation 1	-0.56	0.24	audio script	-0.55	0.25	no exposure	-0.01	0.35	-0.04	38	0.967	-0.01
High	Conversation 1	-0.56	0.24	audio script	-0.46	0.17	baseline	-0.10	0.30	-0.34	40	0.739	-0.11
High	Conversation 1	-0.50	0.29	audio only	-0.55	0.25	no exposure	0.05	0.38	0.12	31	0.905	0.04
High	Conversation 1	-0.50	0.29	audio only	-0.46	0.17	baseline	-0.04	0.34	-0.12	26	0.906	-0.05
High	Conversation 1	-0.55	0.25	no exposure	-0.46	0.17	baseline	-0.09	0.31	-0.28	37	0.780	-0.09
High	Lecture 1	-0.14	0.21	audio script	-0.43	0.23	audio only	0.28	0.31	0.90	32	0.373	0.32
High	Lecture 1	-0.14	0.21	audio script	-0.47	0.20	no exposure	0.33	0.29	1.14	38	0.262	0.37
High	Lecture 1	-0.14	0.21	audio script	-0.34	0.16	baseline	0.19	0.26	0.74	42	0.462	0.23
High	Lecture 1	-0.43	0.23	audio only	-0.47	0.20	no exposure	0.05	0.31	0.16	31	0.877	0.06
High	Lecture 1	-0.43	0.23	audio only	-0.34	0.16	baseline	-0.09	0.28	-0.32	29	0.751	-0.12
High	Lecture 1	-0.47	0.20	no exposure	-0.34	0.16	baseline	-0.14	0.26	-0.54	41	0.593	-0.17
High	Conversation 2	0.97	0.41	audio script	-0.48	0.30	audio only	1.45	0.51	2.84	21	0.010	1.24
High	Conversation 2	0.97	0.41	audio script	-0.37	0.31	no exposure	1.35	0.52	2.61	21	0.016	1.14

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	Conversation 2	0.97	0.41	audio script	0.06	0.18	baseline	0.92	0.45	2.03	15	0.060	1.05
High	Conversation 2	-0.48	0.30	audio only	-0.37	0.31	no exposure	-0.10	0.43	-0.24	28	0.809	-0.09
High	Conversation 2	-0.48	0.30	audio only	0.06	0.18	baseline	-0.53	0.35	-1.53	26	0.138	-0.60
High	Conversation 2	-0.37	0.31	no exposure	0.06	0.18	baseline	-0.43	0.36	-1.21	24	0.240	-0.49
High	Lecture 2	-0.64	0.27	audio script	-0.06	0.26	audio only	-0.58	0.37	-1.57	24	0.129	-0.64
High	Lecture 2	-0.64	0.27	audio script	-0.42	0.24	no exposure	-0.22	0.36	-0.61	23	0.545	-0.25
High	Lecture 2	-0.64	0.27	audio script	0.39	0.16	baseline	-1.03	0.31	-3.28	20	0.004	-1.47
High	Lecture 2	-0.06	0.26	audio only	-0.42	0.24	no exposure	0.36	0.35	1.02	28	0.317	0.39
High	Lecture 2	-0.06	0.26	audio only	0.39	0.16	baseline	-0.45	0.30	-1.47	28	0.152	-0.56
High	Lecture 2	-0.42	0.24	no exposure	0.39	0.16	baseline	-0.81	0.29	-2.75	27	0.011	-1.06
Medium	Conversation 1	-0.52	0.30	audio script	-0.88	0.29	audio only	0.36	0.42	0.86	33	0.396	0.30
Medium	Conversation 1	-0.52	0.30	audio script	-0.79	0.31	no exposure	0.27	0.43	0.62	31	0.540	0.22
Medium	Conversation 1	-0.52	0.30	audio script	-0.88	0.15	baseline	0.35	0.34	1.05	25	0.303	0.42
Medium	Conversation 1	-0.88	0.29	audio only	-0.79	0.31	no exposure	-0.09	0.42	-0.21	33	0.835	-0.07
Medium	Conversation 1	-0.88	0.29	audio only	-0.88	0.15	baseline	0.00	0.33	-0.01	29	0.995	0.00
Medium	Conversation 1	-0.79	0.31	no exposure	-0.88	0.15	baseline	0.09	0.35	0.25	24	0.806	0.10

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	Lecture 1	-0.16	0.25	audio script	-0.17	0.23	audio only	0.01	0.33	0.03	33	0.973	0.01
Medium	Lecture 1	-0.16	0.25	audio script	0.29	0.24	no exposure baseline	-0.45	0.34	-1.31	31	0.200	-0.47
Medium	Lecture 1	-0.16	0.25	audio script	-0.21	0.12	baseline	0.05	0.27	0.19	24	0.852	0.08
Medium	Lecture 1	-0.17	0.23	audio only	0.29	0.24	no exposure baseline	-0.46	0.33	-1.41	33	0.169	-0.49
Medium	Lecture 1	-0.17	0.23	audio only	-0.21	0.12	baseline	0.04	0.26	0.16	28	0.875	0.06
Medium	Lecture 1	0.29	0.24	no exposure	-0.21	0.12	baseline	0.50	0.27	1.88	24	0.072	0.77
Medium	Conversation 2	-0.17	0.24	audio script	0.43	0.24	audio only	-0.59	0.34	-1.73	39	0.091	-0.55
Medium	Conversation 2	-0.17	0.24	audio script	0.26	0.28	no exposure baseline	-0.43	0.37	-1.17	33	0.252	-0.41
Medium	Conversation 2	-0.17	0.24	audio script	-0.29	0.16	baseline	0.13	0.29	0.44	37	0.663	0.14
Medium	Conversation 2	0.43	0.24	audio only	0.26	0.28	no exposure baseline	0.17	0.37	0.45	34	0.652	0.15
Medium	Conversation 2	0.43	0.24	audio only	-0.29	0.16	baseline	0.72	0.29	2.49	37	0.017	0.82
Medium	Conversation 2	0.26	0.28	no exposure	-0.29	0.16	baseline	0.55	0.32	1.75	26	0.093	0.69
Medium	Lecture 2	-0.50	0.21	audio script	-0.57	0.21	audio only	0.07	0.29	0.24	39	0.813	0.08
Medium	Lecture 2	-0.50	0.21	audio script	-0.05	0.23	no exposure baseline	-0.45	0.31	-1.47	34	0.149	-0.50
Medium	Lecture 2	-0.50	0.21	audio script	0.34	0.13	baseline	-0.84	0.24	-3.47	36	0.001	-1.16
Medium	Lecture 2	-0.57	0.21	audio only	-0.05	0.23	no exposure	-0.52	0.31	-1.70	34	0.099	-0.58

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	Lecture 2	-0.57	0.21	audio only	0.34	0.13	baseline	-0.91	0.24	-3.74	36	0.001	-1.25
Medium	Lecture 2	-0.05	0.23	no exposure	0.34	0.13	baseline	-0.39	0.26	-1.48	27	0.150	-0.57
Low	Conversation 1	-0.33	0.34	audio script	-0.40	0.29	audio only	0.07	0.45	0.16	28	0.876	0.06
Low	Conversation 1	-0.33	0.34	audio script	-1.04	0.37	no exposure	0.70	0.50	1.41	27	0.170	0.54
Low	Conversation 1	-0.33	0.34	audio script	-0.73	0.17	baseline	0.39	0.38	1.03	20	0.314	0.46
Low	Conversation 1	-0.40	0.29	audio only	-1.04	0.37	no exposure	0.63	0.47	1.35	30	0.187	0.49
Low	Conversation 1	-0.40	0.29	audio only	-0.73	0.17	baseline	0.32	0.34	0.95	32	0.351	0.34
Low	Conversation 1	-1.04	0.37	no exposure	-0.73	0.17	baseline	-0.31	0.40	-0.77	22	0.451	-0.33
Low	Lecture 1	0.01	0.29	audio script	0.24	0.24	audio only	-0.24	0.38	-0.63	28	0.537	-0.24
Low	Lecture 1	0.01	0.29	audio script	0.12	0.28	no exposure	-0.11	0.40	-0.27	27	0.787	-0.10
Low	Lecture 1	0.01	0.29	audio script	-0.78	0.16	baseline	0.79	0.33	2.39	21	0.026	1.04
Low	Lecture 1	0.24	0.24	audio only	0.12	0.28	no exposure	0.13	0.37	0.34	31	0.735	0.12
Low	Lecture 1	0.24	0.24	audio only	-0.78	0.16	baseline	1.03	0.29	3.53	36	0.001	1.18
Low	Lecture 1	0.12	0.28	no exposure	-0.78	0.16	baseline	0.90	0.32	2.79	25	0.010	1.12
Low	Conversation 2	0.01	0.28	audio script	0.18	0.26	audio only	-0.18	0.38	-0.46	33	0.648	-0.16
Low	Conversation 2	0.01	0.28	audio script	-0.29	0.27	no exposure	0.29	0.39	0.75	33	0.457	0.26

Target proficiency level	Listening Passage	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	Conversation 2	0.01	0.28	audio script	-0.33	0.16	baseline	0.33	0.32	1.04	27	0.308	0.40
Low	Conversation 2	0.18	0.26	audio only	-0.29	0.27	no exposure	0.47	0.38	1.25	35	0.221	0.42
Low	Conversation 2	0.18	0.26	audio only	-0.33	0.16	baseline	0.51	0.30	1.67	32	0.104	0.59
Low	Conversation 2	-0.29	0.27	no exposure	-0.33	0.16	baseline	0.04	0.31	0.13	31	0.896	0.05
Low	Lecture 2	0.13	0.26	audio script	0.00	0.25	audio only	0.13	0.36	0.37	33	0.714	0.13
Low	Lecture 2	0.13	0.26	audio script	0.24	0.24	no exposure	-0.11	0.35	-0.30	33	0.765	-0.10
Low	Lecture 2	0.13	0.26	audio script	0.19	0.13	baseline	-0.06	0.29	-0.20	25	0.843	-0.08
Low	Lecture 2	0.00	0.25	audio only	0.24	0.24	no exposure	-0.24	0.35	-0.69	35	0.496	-0.23
Low	Lecture 2	0.00	0.25	audio only	0.19	0.13	baseline	-0.19	0.29	-0.67	28	0.511	-0.25
Low	Lecture 2	0.24	0.24	no exposure	0.19	0.13	baseline	0.05	0.27	0.17	30	0.862	0.06

Appendix B.4 Pairwise Bias Report for Exposure Condition and Passage Type

Target passage type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Conversation	0.17	0.12	audio script	0.13	0.11	audio only	0.04	0.16	0.24	210	0.812	0.03
Conversation	0.17	0.12	audio script	-0.06	0.12	no exposure	0.23	0.17	1.38	203	0.170	0.19
Conversation	0.17	0.12	audio script	-0.09	0.07	baseline	0.26	0.13	1.93	172	0.055	0.29
Conversation	0.13	0.11	audio only	-0.06	0.12	no exposure	0.19	0.16	1.15	211	0.249	0.16
Conversation	0.13	0.11	audio only	-0.09	0.07	baseline	0.22	0.13	1.67	191	0.096	0.24
Conversation	-0.06	0.12	no exposure	-0.09	0.07	baseline	0.03	0.14	0.23	173	0.821	0.03
Lecture	-0.13	0.10	audio script	-0.10	0.10	audio only	-0.03	0.14	-0.22	210	0.829	-0.03
Lecture	-0.13	0.10	audio script	0.04	0.10	no exposure	-0.16	-0.14	1.17	203	0.245	0.16
Lecture	-0.13	0.10	audio script	0.06	0.06	baseline	-0.19	-0.11	1.66	170	0.099	0.25
Lecture	-0.10	0.10	audio only	0.04	0.10	no exposure	-0.13	0.14	-0.96	212	0.336	-0.13
Lecture	-0.10	0.10	audio only	0.06	0.06	baseline	-0.16	-0.11	1.42	189	0.157	0.21
Lecture	0.04	0.10	no exposure	0.06	0.06	baseline	-0.03	0.11	-0.22	173	0.824	-0.03

Appendix B.5 Pairwise Bias Report for Exposure Condition, Passage Type, and Proficiency Level

Target proficiency level	Passage type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	conversation	0.36	0.20	audio script	-0.11	0.21	audio only	0.47	0.29	1.59	62	0.117	0.40
High	conversation	0.36	0.20	audio script	-0.05	0.20	no exposure	0.41	0.28	1.45	65	0.152	0.36
High	conversation	0.36	0.20	audio script	0.17	0.12	baseline	0.19	0.24	0.82	58	0.418	0.22
High	conversation	-0.11	0.21	audio script	-0.05	0.20	no exposure	-0.06	0.29	-	64	0.847	-0.05
High	conversation	-0.11	0.21	audio only	0.17	0.12	baseline	-0.27	-	0.19	54	0.273	-0.30
High	conversation	-0.05	0.20	no exposure	0.17	0.12	baseline	-0.22	0.23	-	63	0.359	-0.23
High	lecture	-0.17	0.16	audio script	-0.13	0.17	audio only	-0.03	0.24	-	62	0.885	-0.04
High	lecture	-0.17	0.16	audio script	-0.31	0.16	no exposure	0.14	0.23	0.64	65	0.527	0.16
High	lecture	-0.17	0.16	audio script	0.14	0.11	baseline	-0.31	-	-	64	0.127	-0.39
High	lecture	-0.13	0.17	audio script	-0.31	0.16	no exposure	0.18	0.23	0.77	63	0.445	0.19
High	lecture	-0.13	0.17	audio only	0.14	0.11	baseline	-0.27	-	-	59	0.189	-0.35
High	lecture	-0.31	0.16	no exposure	0.14	0.11	baseline	-0.45	-	-	70	0.022	-0.56
Medium	conversation	0.00	0.19	audio script	0.20	0.18	audio only	-0.20	0.26	-	75	0.451	-0.18
Medium	conversation	0.00	0.19	audio script	0.13	0.20	no exposure	-0.13	0.28	-	68	0.641	-0.11

Target proficiency level	Passage type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	conversation	0.00	0.19	audio script	-0.22	0.11	baseline	0.22	0.22	0.99	63	0.326	0.25
Medium	conversation	0.20	0.18	audio only	0.13	0.20	no exposure	0.07	0.27	0.25	69	0.803	0.06
Medium	conversation	0.20	0.18	audio only	-0.22	0.11	baseline	0.41	0.21	1.96	69	0.055	0.47
Medium	conversation	0.13	0.20	no exposure	-0.22	0.11	baseline	0.35	0.23	1.50	53	0.141	0.41
Medium	lecture	-0.28	0.16	audio script	-0.30	0.15	audio only	0.02	0.22	0.07	75	0.945	0.02
Medium	lecture	-0.28	0.16	audio script	0.22	0.17	no exposure	-0.50	-	-	69	0.032	-0.53
Medium	lecture	-0.28	0.16	audio script	0.17	0.09	baseline	-0.45	-	-	61	0.015	-0.64
Medium	lecture	-0.30	0.15	audio script	0.22	0.17	no exposure	-0.52	-	-	70	0.025	-0.55
Medium	lecture	-0.30	0.15	audio only	0.17	0.09	baseline	-0.47	-	-	66	0.010	-0.66
Medium	lecture	0.22	0.17	no exposure	0.17	0.09	baseline	0.05	0.19	0.26	52	0.798	0.07
Low	conversation	0.18	0.22	audio script	0.26	0.20	audio only	-0.08	0.29	-	64	0.791	-0.07
Low	conversation	0.18	0.22	audio script	-0.27	0.22	no exposure	0.45	0.31	1.47	63	0.147	0.37
Low	conversation	0.18	0.22	audio script	-0.16	0.12	baseline	0.35	0.25	1.42	49	0.163	0.41
Low	conversation	0.26	0.20	audio only	-0.27	0.22	no exposure	0.53	0.29	1.81	70	0.075	0.43
Low	conversation	0.26	0.20	audio only	-0.16	0.12	baseline	0.43	0.23	1.87	67	0.066	0.46

Target proficiency level	Passage type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	conversation	-0.27	0.22	no exposure	-0.16	0.12	baseline	-0.10	0.25	-	55	0.682	-0.11
Low	lecture	0.17	0.19	audio script	0.23	0.17	audio only	-0.06	0.26	-	65	0.824	-0.05
Low	lecture	0.17	0.19	audio script	0.28	0.18	no exposure	-0.10	0.26	-	63	0.696	-0.10
Low	lecture	0.17	0.19	audio script	-0.14	0.10	baseline	0.32	0.22	1.46	48	0.150	0.42
Low	lecture	0.23	0.17	audio only	0.28	0.18	no exposure	-0.05	0.25	-	71	0.857	-0.04
Low	lecture	0.23	0.17	audio only	-0.14	0.10	baseline	0.38	0.20	1.86	65	0.068	0.46
Low	lecture	0.28	0.18	no exposure	-0.14	0.10	baseline	0.42	0.21	2.03	57	0.047	0.54

Appendix B.6 Pairwise Bias Report for Exposure Condition and Test Item

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
1111	-2.03	0.40	audio script	-1.50	0.38	audio only	-0.54	0.55	-0.97	104	0.336	-0.19
1111	-2.03	0.40	audio script	-2.29	0.44	no exposure	0.26	0.60	0.43	102	0.671	0.09
1111	-2.03	0.40	audio script	-1.62	0.20	baseline	-0.42	0.45	-0.92	78	0.360	-0.21
1111	-1.50	0.38	audio only	-2.29	0.44	no exposure	0.79	0.58	1.35	102	0.179	0.27
1111	-1.50	0.38	audio only	-1.62	0.20	baseline	0.12	0.43	0.28	86	0.782	0.06
1111	-2.29	0.44	no exposure	-1.62	0.20	baseline	-0.67	-0.49	-1.37	75	0.173	-0.32
1221	1.29	0.35	audio script	0.73	0.31	audio only	0.56	0.46	1.20	102	0.232	0.24
1221	1.29	0.35	audio script	0.95	0.33	no exposure	0.33	0.48	0.69	102	0.490	0.14
1221	1.29	0.35	audio script	0.56	0.19	baseline	0.72	0.40	1.82	83	0.072	0.40
1221	0.73	0.31	audio only	0.95	0.33	no exposure	-0.23	0.45	-0.50	105	0.618	-0.10
1221	0.73	0.31	audio only	0.56	0.19	baseline	0.16	0.36	0.45	97	0.655	0.09
1221	0.95	0.33	no exposure	0.56	0.19	baseline	0.39	0.38	1.02	88	0.312	0.22
1321	-0.64	0.33	audio script	-0.86	0.34	audio only	0.22	0.47	0.48	104	0.636	0.09
1321	-0.64	0.33	audio script	-0.86	0.34	no exposure	0.22	0.47	0.47	102	0.637	0.09

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
1321	-0.64	0.33	audio script	-1.58	0.20	baseline	0.94	0.39	2.43	92	0.017	0.51
1321	-0.86	0.34	audio only	-0.86	0.34	no exposure	0.00	0.48	0.00	105	0.999	0.00
1321	-0.86	0.34	audio only	-1.58	0.20	baseline	0.72	0.39	1.81	95	0.073	0.37
1321	-0.86	0.34	no exposure	-1.58	0.20	baseline	0.72	0.40	1.81	91	0.073	0.38
1431	-1.73	0.38	audio script	-1.50	0.38	audio only	-0.23	0.54	-0.43	104	0.667	-0.08
1431	-1.73	0.38	audio script	-1.62	0.38	no exposure	-0.11	0.54	-0.20	102	0.843	-0.04
1431	-1.73	0.38	audio script	-1.38	0.20	baseline	-0.35	0.43	-0.83	80	0.412	-0.19
1431	-1.50	0.38	audio only	-1.62	0.38	no exposure	0.12	0.54	0.23	105	0.817	0.04
1431	-1.50	0.38	audio only	-1.38	0.20	baseline	-0.12	0.43	-0.28	84	0.779	-0.06
1431	-1.62	0.38	no exposure	-1.38	0.20	baseline	-0.25	0.43	-0.57	81	0.567	-0.13
1533	0.51	0.32	audio script	0.44	0.31	audio only	0.06	0.45	0.15	104	0.885	0.03
1533	0.51	0.32	audio script	0.21	0.32	no exposure	0.30	0.46	0.65	102	0.517	0.13
1533	0.51	0.32	audio script	0.64	0.19	baseline	-0.13	0.38	-0.34	89	0.733	-0.07
1533	0.44	0.31	audio only	0.21	0.32	no exposure	0.23	0.45	0.52	105	0.603	0.10
1533	0.44	0.31	audio only	0.64	0.19	baseline	-0.19	0.36	-0.54	99	0.594	-0.11
1533	0.21	0.32	no exposure	0.64	0.19	baseline	-0.43	-0.38	-1.14	91	0.259	-0.24

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
2111	0.61	0.33	audio script	0.44	0.31	audio only	0.17	0.45	0.38	104	0.705	0.07
2111	0.61	0.33	audio script	0.21	0.32	no exposure	0.40	0.46	0.88	102	0.382	0.17
2111	0.61	0.33	audio script	0.25	0.19	baseline	0.37	0.37	0.98	86	0.332	0.21
2111	0.44	0.31	audio only	0.21	0.32	no exposure	0.23	0.45	0.52	105	0.603	0.10
2111	0.44	0.31	audio only	0.25	0.19	baseline	0.20	0.36	0.54	96	0.587	0.11
2111	0.21	0.32	no exposure	0.25	0.19	baseline	-0.04	0.37	-0.10	88	0.920	-0.02
2221	-0.86	0.33	audio script	-0.98	0.34	audio only	0.12	0.48	0.25	104	0.802	0.05
2221	-0.86	0.33	audio script	-1.10	0.35	no exposure	0.24	0.48	0.50	102	0.621	0.10
2221	-0.86	0.33	audio script	-1.70	0.21	baseline	0.84	0.39	2.15	92	0.034	0.45
2221	-0.98	0.34	audio only	-1.10	0.35	no exposure	0.12	0.49	0.24	105	0.808	0.05
2221	-0.98	0.34	audio only	-1.70	0.21	baseline	0.72	0.40	1.80	95	0.075	0.37
2221	-1.10	0.35	no exposure	-1.70	0.21	baseline	0.60	0.41	1.49	91	0.139	0.31
2321	-0.32	0.32	audio script	-0.43	0.32	audio only	0.11	0.46	0.24	104	0.809	0.05
2321	-0.32	0.32	audio script	-0.10	0.32	no exposure	-0.22	0.46	-0.48	102	0.634	-0.10
2321	-0.32	0.32	audio script	-0.19	0.18	baseline	-0.13	0.37	-0.35	85	0.727	-0.08
2321	-0.43	0.32	audio only	-0.10	0.32	no exposure	-0.33	0.46	-0.72	105	0.473	-0.14

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
2321	-0.43	0.32	audio only	-0.19	0.18	baseline	-0.24	0.37	-0.65	91	0.517	-0.14
2321	-0.10	0.32	no exposure	-0.19	0.18	baseline	0.09	0.37	0.24	87	0.813	0.05
2421	-0.11	0.32	audio script	0.54	0.31	audio only	-0.65	-0.44	-1.46	104	0.147	-0.29
2421	-0.11	0.32	audio script	0.52	0.32	no exposure	-0.64	-0.46	-1.39	102	0.167	-0.28
2421	-0.11	0.32	audio script	-0.09	0.18	baseline	-0.02	0.37	-0.06	86	0.951	-0.01
2421	0.54	0.31	audio only	0.52	0.32	no exposure	0.01	0.45	0.03	105	0.976	0.01
2421	0.54	0.31	audio only	-0.09	0.18	baseline	0.63	0.36	1.76	94	0.083	0.36
2421	0.52	0.32	no exposure	-0.09	0.18	baseline	0.61	0.37	1.65	87	0.103	0.35
2532	0.20	0.32	audio script	0.06	0.31	audio only	0.13	0.45	0.30	104	0.767	0.06
2532	0.20	0.32	audio script	0.52	0.32	no exposure	-0.33	0.46	-0.72	103	0.475	-0.14
2532	0.20	0.32	audio script	-0.12	0.18	baseline	0.32	0.37	0.87	86	0.389	0.19
2532	0.06	0.31	audio only	0.52	0.32	no exposure	-0.46	-0.45	-1.03	105	0.308	-0.20
2532	0.06	0.31	audio only	-0.12	0.18	baseline	0.19	0.36	0.52	94	0.604	0.11
2532	0.52	0.32	no exposure	-0.12	0.18	baseline	0.65	0.37	1.74	87	0.086	0.37
2632	-0.53	0.33	audio script	-0.86	0.34	audio only	0.33	0.47	0.71	104	0.482	0.14
2632	-0.53	0.33	audio script	-0.98	0.34	no exposure	0.45	0.47	0.94	102	0.347	0.19

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
2632	-0.53	0.33	audio script	-0.83	0.19	baseline	0.30	0.38	0.80	86	0.427	0.17
2632	-0.86	0.34	audio only	-0.98	0.34	no exposure	0.12	0.48	0.24	105	0.811	0.05
2632	-0.86	0.34	audio only	-0.83	0.19	baseline	-0.03	0.39	-0.08	88	0.936	-0.02
2632	-0.98	0.34	no exposure	-0.83	0.19	baseline	-0.15	0.39	-0.38	84	0.707	-0.08
3111	0.80	0.33	audio script	1.50	0.34	audio only	-0.70	-0.47	-1.49	103	0.140	-0.29
3111	0.80	0.33	audio script	0.87	0.33	no exposure	-0.07	0.46	-0.15	98	0.881	-0.03
3111	0.80	0.33	audio script	0.58	0.19	baseline	0.22	0.38	0.57	84	0.567	0.12
3111	1.50	0.34	audio only	0.87	0.33	no exposure	0.63	0.47	1.33	104	0.186	0.26
3111	1.50	0.34	audio only	0.58	0.19	baseline	0.92	0.39	2.36	91	0.021	0.49
3111	0.87	0.33	no exposure	0.58	0.19	baseline	0.29	0.38	0.75	85	0.455	0.16
3221	-0.45	0.33	audio script	-0.88	0.33	audio only	0.43	0.47	0.92	103	0.358	0.18
3221	-0.45	0.33	audio script	-1.26	0.37	no exposure	0.81	0.49	1.64	98	0.105	0.33
3221	-0.45	0.33	audio script	-0.82	0.19	baseline	0.37	0.38	0.98	82	0.331	0.22
3221	-0.88	0.33	audio only	-1.26	0.37	no exposure	0.38	0.49	0.77	102	0.444	0.15
3221	-0.88	0.33	audio only	-0.82	0.19	baseline	-0.06	0.38	-0.16	92	0.876	-0.03
3221	-1.26	0.37	no exposure	-0.82	0.19	baseline	-0.44	-0.41	-1.06	77	0.291	-0.24

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
3331	0.49	0.32	audio script	0.11	0.31	audio only	0.38	0.44	0.85	102	0.395	0.17
3331	0.49	0.32	audio script	0.44	0.32	no exposure	0.05	0.46	0.10	98	0.920	0.02
3331	0.49	0.32	audio script	-0.04	0.18	baseline	0.53	0.37	1.43	83	0.158	0.31
3331	0.11	0.31	audio only	0.44	0.32	no exposure	-0.33	0.45	-0.75	103	0.458	-0.15
3331	0.11	0.31	audio only	-0.04	0.18	baseline	0.15	0.36	0.42	97	0.679	0.09
3331	0.44	0.32	no exposure	-0.04	0.18	baseline	0.48	0.37	1.29	84	0.201	0.28
3421	-1.03	0.35	audio script	-0.88	0.33	audio only	-0.15	0.48	-0.31	102	0.757	-0.06
3421	-1.03	0.35	audio script	-1.54	0.39	no exposure	0.51	0.52	0.98	98	0.330	0.20
3421	-1.03	0.35	audio script	-1.89	0.21	baseline	0.86	0.41	2.10	86	0.039	0.45
3421	-0.88	0.33	audio only	-1.54	0.39	no exposure	0.66	0.51	1.31	100	0.194	0.26
3421	-0.88	0.33	audio only	-1.89	0.21	baseline	1.01	0.39	2.60	103	0.011	0.51
3421	-1.54	0.39	no exposure	-1.89	0.21	baseline	0.35	0.44	0.80	81	0.429	0.18
3533	1.02	0.33	audio script	1.07	0.32	audio only	-0.05	0.46	-0.11	103	0.910	-0.02
3533	1.02	0.33	audio script	1.21	0.34	no exposure	-0.19	0.48	-0.40	98	0.689	-0.08
3533	1.02	0.33	audio script	1.57	0.22	baseline	-0.55	-0.40	-1.39	96	0.169	-0.28
3533	1.07	0.32	audio only	1.21	0.34	no exposure	-0.14	0.47	-0.30	103	0.767	-0.06

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
3533	1.07	0.32	audio only	1.57	0.22	baseline	-0.50	-0.39	-1.29	111	0.200	-0.24
3533	1.21	0.34	no exposure	1.57	0.22	baseline	-0.36	0.41	-0.89	96	0.375	-0.18
4111	-1.15	0.36	audio script	-0.88	0.33	audio only	-0.28	0.49	-0.57	101	0.571	-0.11
4111	-1.15	0.36	audio script	-1.26	0.37	no exposure	0.10	0.52	0.20	98	0.845	0.04
4111	-1.15	0.36	audio script	-0.17	0.18	baseline	-0.98	-0.41	-2.42	75	0.018	-0.56
4111	-0.88	0.33	audio only	-1.26	0.37	no exposure	0.38	0.49	0.77	102	0.444	0.15
4111	-0.88	0.33	audio only	-0.17	0.18	baseline	-0.70	-0.38	-1.87	91	0.065	-0.39
4111	-1.26	0.37	no exposure	-0.17	0.18	baseline	-1.08	-0.41	-2.63	76	0.010	-0.60
4231	-0.02	0.32	audio script	0.11	0.31	audio only	-0.13	0.44	-0.30	102	0.764	-0.06
4231	-0.02	0.32	audio script	0.13	0.32	no exposure	-0.15	0.46	-0.33	98	0.743	-0.07
4231	-0.02	0.32	audio script	0.37	0.19	baseline	-0.40	-0.37	-1.06	84	0.291	-0.23
4231	0.11	0.31	audio only	0.13	0.32	no exposure	-0.02	0.45	-0.04	103	0.970	-0.01
4231	0.11	0.31	audio only	0.37	0.19	baseline	-0.26	0.36	-0.73	98	0.467	-0.15
4231	0.13	0.32	no exposure	0.37	0.19	baseline	-0.24	0.37	-0.65	85	0.515	-0.14
4321	0.59	0.32	audio script	0.67	0.31	audio only	-0.08	0.45	-0.19	103	0.853	-0.04
4321	0.59	0.32	audio script	0.98	0.33	no exposure	-0.39	0.46	-0.84	98	0.403	-0.17

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
4321	0.59	0.32	audio script	0.92	0.20	baseline	-0.33	0.38	-0.87	88	0.386	-0.19
4321	0.67	0.31	audio only	0.98	0.33	no exposure	-0.31	0.45	-0.67	103	0.502	-0.13
4321	0.67	0.31	audio only	0.92	0.20	baseline	-0.25	0.37	-0.67	102	0.504	-0.13
4321	0.98	0.33	no exposure	0.92	0.20	baseline	0.06	0.39	0.16	87	0.877	0.03
4421	-0.56	0.33	audio script	-0.46	0.31	audio only	-0.09	0.46	-0.20	102	0.841	-0.04
4421	-0.56	0.33	audio script	-0.41	0.33	no exposure	-0.15	0.47	-0.31	98	0.758	-0.06
4421	-0.56	0.33	audio script	0.10	0.18	baseline	-0.65	-0.38	-1.71	81	0.090	-0.38
4421	-0.46	0.31	audio only	-0.41	0.33	no exposure	-0.05	0.46	-0.12	103	0.907	-0.02
4421	-0.46	0.31	audio only	0.10	0.18	baseline	-0.56	-0.36	-1.54	95	0.127	-0.32
4421	-0.41	0.33	no exposure	0.10	0.18	baseline	-0.51	-0.38	-1.33	82	0.186	-0.29
4521	0.08	0.32	audio script	0.11	0.31	audio only	-0.03	0.44	-0.07	102	0.945	-0.01
4521	0.08	0.32	audio script	0.44	0.32	no exposure	-0.36	0.46	-0.80	98	0.428	-0.16
4521	0.08	0.32	audio script	0.73	0.19	baseline	-0.65	-0.37	-1.74	87	0.085	-0.37
4521	0.11	0.31	audio only	0.44	0.32	no exposure	-0.33	0.45	-0.75	103	0.458	-0.15
4521	0.11	0.31	audio only	0.73	0.19	baseline	-0.62	-0.36	-1.72	101	0.089	-0.34
4521	0.44	0.32	no exposure	0.73	0.19	baseline	-0.29	0.38	-0.76	88	0.448	-0.16

Target item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target Contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
4633	-1.43	0.38	audio script	-1.46	0.36	audio only	0.03	0.52	0.06	102	0.949	0.01
4633	-1.43	0.38	audio script	-0.64	0.34	no exposure	-0.79	-0.51	-1.55	97	0.123	-0.31
4633	-1.43	0.38	audio script	-0.54	0.18	baseline	-0.88	-0.42	-2.10	73	0.040	-0.49
4633	-1.46	0.36	audio only	-0.64	0.34	no exposure	-0.83	-0.49	-1.67	104	0.098	-0.33
4633	-1.46	0.36	audio only	-0.54	0.18	baseline	-0.92	-0.40	-2.28	85	0.025	-0.49
4633	-0.64	0.34	no exposure	-0.54	0.18	baseline	-0.09	0.39	-0.24	81	0.810	-0.05

Appendix B.7 Pairwise Bias Report for Exposure Condition and Listening Proficiency at Item Level

Target proficiency	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	0.03	0.13	audio script	-0.13	0.13	audio only	0.16	0.18	0.88	712	0.380	0.07
High	0.03	0.13	audio script	-0.21	0.12	no exposure	0.24	0.17	1.39	740	0.166	0.10
High	0.03	0.13	audio script	0.16	0.08	baseline	-0.14	0.15	-0.92	686	0.359	-0.07
High	-0.13	0.13	audio only	-0.21	0.12	no exposure	0.08	0.18	0.47	725	0.638	0.03
High	-0.13	0.13	audio only	0.16	0.08	baseline	-0.30	0.15	-1.95	655	0.052	-0.15
High	-0.21	0.12	no exposure	0.16	0.08	baseline	-0.38	0.14	-2.62	764	0.009	-0.19
Medium	-0.14	0.11	audio script	-0.08	0.11	audio only	-0.06	0.16	-0.37	851	0.715	-0.03
Medium	-0.14	0.11	audio script	0.17	0.12	no exposure	-0.31	0.17	-1.84	778	0.066	-0.13
Medium	-0.14	0.11	audio script	0.01	0.06	baseline	-0.15	0.13	-1.11	705	0.267	-0.08
Medium	-0.08	0.11	audio only	0.17	0.12	no exposure	-0.25	0.16	-1.52	783	0.129	-0.11
Medium	-0.08	0.11	audio only	0.01	0.06	baseline	-0.09	0.13	-0.69	769	0.489	-0.05
Medium	0.17	0.12	no exposure	0.01	0.06	baseline	0.16	0.14	1.17	599	0.243	0.10
Low	0.16	0.13	audio script	0.22	0.12	audio only	-0.07	0.18	-0.37	732	0.712	-0.03
Low	0.16	0.13	audio script	0.05	0.13	no exposure	0.10	0.19	0.56	719	0.572	0.04
Low	0.16	0.13	audio script	-0.14	0.07	baseline	0.29	0.15	1.96	550	0.051	0.17

Target proficiency	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	0.22	0.12	audio only	0.05	0.13	no exposure	0.17	0.18	0.97	797	0.334	0.07
Low	0.22	0.12	audio only	-0.14	0.07	baseline	0.36	0.14	2.59	749	0.010	0.19
Low	0.05	0.13	no exposure	-0.14	0.07	baseline	0.19	0.15	1.28	630	0.200	0.10

Appendix B.8 Pairwise Bias Report for Exposure Condition, Test Item, and Listening Proficiency

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	1111	-2.19	0.51	audio script	-1.30	0.55	audio only	-0.90	0.75	-1.20	33	0.238	-0.42
High	1111	-2.19	0.51	audio script	-2.04	0.52	no exposure	-0.16	0.72	-0.21	38	0.831	-0.07
High	1111	-2.19	0.51	audio script	-1.51	0.33	baseline	-0.68	0.60	-1.13	37	0.268	-0.37
High	1111	-1.30	0.55	audio only	-2.04	0.52	no exposure	0.74	0.76	0.98	32	0.333	0.35
High	1111	-1.30	0.55	audio only	-1.51	0.33	baseline	0.22	0.64	0.34	26	0.735	0.13
High	1111	-2.04	0.52	no exposure	-1.51	0.33	baseline	-0.52	0.61	-0.85	34	0.400	-0.29
High	1221	1.38	0.78	audio script	0.27	0.61	audio only	1.11	1.00	1.12	34	0.272	0.38
High	1221	1.38	0.78	audio script	0.84	0.65	no exposure	0.54	1.02	0.53	38	0.600	0.17
High	1221	1.38	0.78	audio script	0.24	0.45	baseline	1.14	0.91	1.25	33	0.219	0.44
High	1221	0.27	0.61	audio only	0.84	0.65	no exposure	-0.57	0.89	-0.64	33	0.525	-0.22
High	1221	0.27	0.61	audio only	0.24	0.45	baseline	0.02	0.76	0.03	32	0.975	0.01
High	1221	0.84	0.65	no exposure	0.24	0.45	baseline	0.60	0.79	0.76	38	0.455	0.25
High	1321	-0.13	0.52	audio script	-1.30	0.55	audio only	1.16	0.75	1.55	33	0.132	0.54
High	1321	-0.13	0.52	audio script	-0.62	0.48	no exposure	0.49	0.71	0.70	38	0.490	0.23

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	1321	-0.13	0.52	audio script	-1.18	0.34	baseline	1.05	0.62	1.71	37	0.095	0.56
High	1321	-1.30	0.55	audio only	-0.62	0.48	no exposure	-0.67	0.73	-0.92	32	0.366	-0.33
High	1321	-1.30	0.55	audio only	-1.18	0.34	baseline	-0.11	0.64	-0.17	26	0.864	-0.07
High	1321	-0.62	0.48	no exposure	-1.18	0.34	baseline	0.56	0.59	0.95	37	0.347	0.31
High	1431	-2.77	0.58	audio script	-1.61	0.57	audio only	-1.17	0.81	-1.44	34	0.158	-0.49
High	1431	-2.77	0.58	audio script	-2.32	0.55	no exposure	-0.45	0.80	-0.57	38	0.575	-0.18
High	1431	-2.77	0.58	audio script	-1.41	0.33	baseline	-1.37	0.66	-2.05	33	0.048	-0.71
High	1431	-1.61	0.57	audio only	-2.32	0.55	no exposure	0.72	0.79	0.91	33	0.370	0.32
High	1431	-1.61	0.57	audio only	-1.41	0.33	baseline	-0.20	0.65	-0.30	25	0.763	-0.12
High	1431	-2.32	0.55	no exposure	-1.41	0.33	baseline	-0.92	0.64	-1.43	33	0.162	-0.50
High	1533	0.86	0.66	audio script	1.21	0.79	audio only	-0.35	1.03	-0.33	31	0.740	-0.12
High	1533	0.86	0.66	audio script	1.34	0.77	no exposure	-0.47	1.01	-0.46	37	0.645	-0.15
High	1533	0.86	0.66	audio script	1.05	0.61	baseline	-0.19	0.90	-0.21	50	0.834	-0.06
High	1533	1.21	0.79	audio only	1.34	0.77	no exposure	-0.13	1.10	-0.11	33	0.910	-0.04
High	1533	1.21	0.79	audio only	1.05	0.61	baseline	0.16	1.00	0.16	33	0.877	0.06

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	1533	1.34	0.77	no exposure	1.05	0.61	baseline	0.28	0.98	0.29	42	0.775	0.09
High	2111	0.86	0.66	audio script	2.03	1.06	audio only	-1.16	1.25	-0.93	26	0.361	-0.36
High	2111	0.86	0.66	audio script	0.15	0.54	no exposure	0.72	0.86	0.83	37	0.409	0.27
High	2111	0.86	0.66	audio script	1.05	0.61	baseline	-0.19	0.90	-0.21	50	0.834	-0.06
High	2111	2.03	1.06	audio only	0.15	0.54	no exposure	1.88	1.19	1.58	22	0.129	0.67
High	2111	2.03	1.06	audio only	1.05	0.61	baseline	0.97	1.22	0.80	25	0.434	0.32
High	2111	0.15	0.54	no exposure	1.05	0.61	baseline	-0.91	0.82	-1.11	56	0.272	-0.30
High	2221	-0.62	0.48	audio script	-1.00	0.54	audio only	0.38	0.72	0.52	32	0.606	0.18
High	2221	-0.62	0.48	audio script	-1.54	0.49	no exposure	0.92	0.68	1.35	38	0.186	0.44
High	2221	-0.62	0.48	audio script	-1.51	0.33	baseline	0.89	0.58	1.54	39	0.132	0.49
High	2221	-1.00	0.54	audio only	-1.54	0.49	no exposure	0.54	0.73	0.74	32	0.463	0.26
High	2221	-1.00	0.54	audio only	-1.51	0.33	baseline	0.52	0.63	0.81	26	0.423	0.32
High	2221	-1.54	0.49	no exposure	-1.51	0.33	baseline	-0.02	0.59	-0.04	36	0.967	-0.01
High	2321	-0.62	0.48	audio script	-0.70	0.54	audio only	0.08	0.72	0.11	32	0.910	0.04
High	2321	-0.62	0.48	audio script	-0.38	0.50	no exposure	-0.24	0.69	-0.34	38	0.733	-0.11

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	2321	-0.62	0.48	audio script	0.05	0.43	baseline	-0.67	0.64	-1.04	49	0.302	-0.30
High	2321	-0.70	0.54	audio only	-0.38	0.50	no exposure	-0.32	0.74	-0.43	32	0.667	-0.15
High	2321	-0.70	0.54	audio only	0.05	0.43	baseline	-0.75	0.69	-1.09	34	0.285	-0.37
High	2321	-0.38	0.50	no exposure	0.05	0.43	baseline	-0.43	0.66	-0.66	46	0.512	-0.19
High	2421	-0.85	0.47	audio script	-1.61	0.57	audio only	0.76	0.73	1.03	31	0.309	0.37
High	2421	-0.85	0.47	audio script	-0.38	0.50	no exposure	-0.46	0.68	-0.68	38	0.504	-0.22
High	2421	-0.85	0.47	audio script	-0.13	0.41	baseline	-0.72	0.62	-1.16	48	0.253	-0.33
High	2421	-1.61	0.57	audio only	-0.38	0.50	no exposure	-1.22	0.75	-1.62	32	0.114	-0.57
High	2421	-1.61	0.57	audio only	-0.13	0.41	baseline	-1.48	0.70	-2.12	31	0.042	-0.76
High	2421	-0.38	0.50	no exposure	-0.13	0.41	baseline	-0.26	0.64	-0.40	44	0.690	-0.12
High	2532	0.86	0.66	audio script	0.27	0.61	audio only	0.60	0.90	0.66	34	0.514	0.23
High	2532	0.86	0.66	audio script	0.47	0.58	no exposure	0.40	0.88	0.45	38	0.654	0.15
High	2532	0.86	0.66	audio script	0.24	0.45	baseline	0.62	0.80	0.77	39	0.445	0.25
High	2532	0.27	0.61	audio only	0.47	0.58	no exposure	-0.20	0.85	-0.23	33	0.818	-0.08
High	2532	0.27	0.61	audio only	0.24	0.45	baseline	0.02	0.76	0.03	32	0.975	0.01

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	2532	0.47	0.58	no exposure	0.24	0.45	baseline	0.22	0.74	0.30	42	0.767	0.09
High	2632	-0.38	0.49	audio script	-1.00	0.54	audio only	0.61	0.73	0.84	33	0.409	0.29
High	2632	-0.38	0.49	audio script	-1.31	0.48	no exposure	0.92	0.69	1.34	38	0.187	0.43
High	2632	-0.38	0.49	audio script	-1.07	0.34	baseline	0.69	0.60	1.14	39	0.260	0.37
High	2632	-1.00	0.54	audio only	-1.31	0.48	no exposure	0.31	0.72	0.43	32	0.672	0.15
High	2632	-1.00	0.54	audio only	-1.07	0.34	baseline	0.07	0.64	0.11	27	0.911	0.04
High	2632	-1.31	0.48	no exposure	-1.07	0.34	baseline	-0.24	0.59	-0.40	38	0.688	-0.13
High	3111	1.48	1.07	audio script	1.54	1.05	audio only	-0.06	1.50	-0.04	25	0.968	-0.02
High	3111	1.48	1.07	audio script	1.63	1.05	no exposure	-0.16	1.50	-0.11	24	0.917	-0.04
High	3111	1.48	1.07	audio script	0.81	0.45	baseline	0.67	1.16	0.57	15	0.575	0.29
High	3111	1.54	1.05	audio only	1.63	1.05	no exposure	-0.10	1.49	-0.07	28	0.948	-0.03
High	3111	1.54	1.05	audio only	0.81	0.45	baseline	0.73	1.14	0.64	20	0.532	0.29
High	3111	1.63	1.05	no exposure	0.81	0.45	baseline	0.82	1.14	0.72	19	0.481	0.33
High	3221	-0.39	0.65	audio script	-2.63	0.61	audio only	2.25	0.89	2.52	24	0.019	1.03
High	3221	-0.39	0.65	audio script	-2.86	0.69	no exposure	2.48	0.95	2.60	24	0.016	1.06

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	3221	-0.39	0.65	audio script	-0.89	0.30	baseline	0.51	0.72	0.71	15	0.492	0.37
High	3221	-2.63	0.61	audio only	-2.86	0.69	no exposure	0.23	0.92	0.25	28	0.806	0.09
High	3221	-2.63	0.61	audio only	-0.89	0.30	baseline	-1.74	0.68	-2.58	22	0.017	-1.10
High	3221	-2.86	0.69	no exposure	-0.89	0.30	baseline	-1.97	0.76	-2.61	19	0.017	-1.20
High	3331	0.07	0.71	audio script	0.74	0.78	audio only	-0.67	1.05	-0.64	25	0.529	-0.26
High	3331	0.07	0.71	audio script	-0.10	0.61	no exposure	0.17	0.94	0.18	23	0.855	0.08
High	3331	0.07	0.71	audio script	0.45	0.40	baseline	-0.38	0.81	-0.47	18	0.645	-0.22
High	3331	0.74	0.78	audio only	-0.10	0.61	no exposure	0.84	0.99	0.85	27	0.402	0.33
High	3331	0.74	0.78	audio only	0.45	0.40	baseline	0.29	0.87	0.33	23	0.743	0.14
High	3331	-0.10	0.61	no exposure	0.45	0.40	baseline	-0.55	0.73	-0.75	27	0.458	-0.29
High	3421	0.07	0.71	audio script	-1.40	0.53	audio only	1.47	0.88	1.67	21	0.111	0.73
High	3421	0.07	0.71	audio script	-1.09	0.56	no exposure	1.16	0.90	1.30	22	0.208	0.55
High	3421	0.07	0.71	audio script	-1.56	0.28	baseline	1.63	0.76	2.14	14	0.050	1.14
High	3421	-1.40	0.53	audio only	-1.09	0.56	no exposure	-0.31	0.77	-0.40	28	0.692	-0.15
High	3421	-1.40	0.53	audio only	-1.56	0.28	baseline	0.16	0.60	0.26	24	0.798	0.11

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	3421	-1.09	0.56	no exposure	-1.56	0.28	baseline	0.46	0.62	0.74	21	0.466	0.32
High	3533	2.24<	1.46	audio script	0.23	0.67	audio only	2.01	1.61	1.25	15	0.230	0.65
High	3533	2.24<	1.46	audio script	1.63	1.05	no exposure	0.61	1.80	0.34	20	0.740	0.15
High	3533	2.24<	1.46	audio script	1.62	0.61	baseline	0.62	1.58	0.39	15	0.699	0.20
High	3533	0.23	0.67	audio only	1.63	1.05	no exposure	-1.40	1.24	-1.13	23	0.271	-0.47
High	3533	0.23	0.67	audio only	1.62	0.61	baseline	-1.39	0.90	-1.54	42	0.132	-0.48
High	3533	1.63	1.05	no exposure	1.62	0.61	baseline	0.02	1.22	0.01	24	0.988	0.00
High	4111	-1.57	0.63	audio script	-0.51	0.57	audio only	-1.06	0.85	-1.25	24	0.223	-0.51
High	4111	-1.57	0.63	audio script	-0.78	0.56	no exposure	-0.79	0.84	-0.94	23	0.357	-0.39
High	4111	-1.57	0.63	audio script	0.30	0.38	baseline	-1.87	0.74	-2.55	20	0.019	-1.14
High	4111	-0.51	0.57	audio only	-0.78	0.56	no exposure	0.27	0.80	0.34	28	0.740	0.13
High	4111	-0.51	0.57	audio only	0.30	0.38	baseline	-0.81	0.69	-1.18	29	0.246	-0.44
High	4111	-0.78	0.56	no exposure	0.30	0.38	baseline	-1.08	0.68	-1.59	28	0.124	-0.60
High	4231	-0.39	0.65	audio script	0.23	0.67	audio only	-0.62	0.93	-0.66	25	0.515	-0.26
High	4231	-0.39	0.65	audio script	-0.46	0.58	no exposure	0.07	0.87	0.08	23	0.938	0.03

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	4231	-0.39	0.65	audio script	0.30	0.38	baseline	-0.69	0.76	-0.90	19	0.377	-0.41
High	4231	0.23	0.67	audio only	-0.46	0.58	no exposure	0.69	0.88	0.78	28	0.444	0.29
High	4231	0.23	0.67	audio only	0.30	0.38	baseline	-0.07	0.77	-0.09	25	0.930	-0.04
High	4231	-0.46	0.58	no exposure	0.30	0.38	baseline	-0.75	0.70	-1.08	27	0.288	-0.42
High	4321	-0.39	0.65	audio script	0.74	0.78	audio only	-1.13	1.02	-1.11	25	0.277	-0.44
High	4321	-0.39	0.65	audio script	1.63	1.05	no exposure	-2.02	1.24	-1.63	22	0.117	-0.70
High	4321	-0.39	0.65	audio script	1.29	0.54	baseline	-1.68	0.85	-1.98	28	0.057	-0.75
High	4321	0.74	0.78	audio only	1.63	1.05	no exposure	-0.89	1.31	-0.68	26	0.501	-0.27
High	4321	0.74	0.78	audio only	1.29	0.54	baseline	-0.55	0.94	-0.58	30	0.566	-0.21
High	4321	1.63	1.05	no exposure	1.29	0.54	baseline	0.34	1.18	0.29	21	0.774	0.13
High	4421	-0.39	0.65	audio script	-0.51	0.57	audio only	0.13	0.87	0.15	23	0.885	0.06
High	4421	-0.39	0.65	audio script	-1.40	0.56	no exposure	1.02	0.86	1.18	23	0.250	0.49
High	4421	-0.39	0.65	audio script	0.81	0.45	baseline	-1.20	0.79	-1.51	22	0.146	-0.64
High	4421	-0.51	0.57	audio only	-1.40	0.56	no exposure	0.89	0.80	1.12	28	0.274	0.42
High	4421	-0.51	0.57	audio only	0.81	0.45	baseline	-1.32	0.73	-1.82	35	0.077	-0.62

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	4421	-1.40	0.56	no exposure	0.81	0.45	baseline	-2.21	0.72	-3.09	34	0.004	-1.06
High	4521	-0.80	0.63	audio script	1.54	1.05	audio only	-2.33	1.22	-1.91	23	0.069	-0.80
High	4521	-0.80	0.63	audio script	1.63	1.05	no exposure	-2.43	1.23	-1.98	22	0.060	-0.84
High	4521	-0.80	0.63	audio script	2.06	0.73	baseline	-2.86	0.97	-2.95	45	0.005	-0.88
High	4521	1.54	1.05	audio only	1.63	1.05	no exposure	-0.10	1.49	-0.07	28	0.948	-0.03
High	4521	1.54	1.05	audio only	2.06	0.73	baseline	-0.52	1.28	-0.41	31	0.685	-0.15
High	4521	1.63	1.05	no exposure	2.06	0.73	baseline	-0.43	1.28	-0.33	29	0.742	-0.12
High	4633	-0.39	0.65	audio script	-1.12	0.54	audio only	0.73	0.85	0.86	23	0.397	0.36
High	4633	-0.39	0.65	audio script	-1.72	0.57	no exposure	1.33	0.87	1.54	23	0.138	0.64
High	4633	-0.39	0.65	audio script	-0.98	0.29	baseline	0.59	0.72	0.83	15	0.421	0.43
High	4633	-1.12	0.54	audio only	-1.72	0.57	no exposure	0.60	0.78	0.77	28	0.448	0.29
High	4633	-1.12	0.54	audio only	-0.98	0.29	baseline	-0.14	0.61	-0.22	24	0.825	-0.09
High	4633	-1.72	0.57	no exposure	-0.98	0.29	baseline	-0.74	0.64	-1.15	22	0.261	-0.49
Medium	1111	-1.79	0.79	audio script	-2.14	0.77	audio only	0.35	1.10	0.32	33	0.754	0.11
Medium	1111	-1.79	0.79	audio script	-3.39>	1.45	no exposure	1.60	1.65	0.97	24	0.342	0.40

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	1111	-1.79	0.79	audio script	-1.98	0.35	baseline	0.19	0.86	0.22	22	0.828	0.09
Medium	1111	-2.14	0.77	audio only	-3.39>	1.45	no exposure	1.25	1.64	0.76	24	0.454	0.31
Medium	1111	-2.14	0.77	audio only	-1.98	0.35	baseline	-0.16	0.85	-0.19	25	0.852	-0.08
Medium	1111	-3.39>	1.45	no exposure	-1.98	0.35	baseline	-1.41	1.49	-0.94	17	0.358	-0.46
Medium	1221	0.91	0.53	audio script	0.64	0.50	audio only	0.27	0.73	0.37	33	0.716	0.13
Medium	1221	0.91	0.53	audio script	1.97	0.63	no exposure	-1.06	0.82	-1.29	31	0.206	-0.46
Medium	1221	0.91	0.53	audio script	0.23	0.29	baseline	0.68	0.60	1.12	26	0.272	0.44
Medium	1221	0.64	0.50	audio only	1.97	0.63	no exposure	-1.33	0.80	-1.65	31	0.109	-0.59
Medium	1221	0.64	0.50	audio only	0.23	0.29	baseline	0.41	0.58	0.71	30	0.485	0.26
Medium	1221	1.97	0.63	no exposure	0.23	0.29	baseline	1.73	0.69	2.52	23	0.019	1.05
Medium	1321	-1.27	0.67	audio script	-0.92	0.55	audio only	-0.35	0.87	-0.40	32	0.693	-0.14
Medium	1321	-1.27	0.67	audio script	-1.84	0.78	no exposure	0.57	1.03	0.55	31	0.586	0.20
Medium	1321	-1.27	0.67	audio script	-1.76	0.33	baseline	0.48	0.75	0.65	24	0.523	0.27
Medium	1321	-0.92	0.55	audio only	-1.84	0.78	no exposure	0.92	0.96	0.95	29	0.348	0.35
Medium	1321	-0.92	0.55	audio only	-1.76	0.33	baseline	0.83	0.64	1.29	31	0.206	0.46

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	1321	-1.84	0.78	no exposure	-1.76	0.33	baseline	-0.08	0.85	-0.10	21	0.922	-0.04
Medium	1431	-1.27	0.67	audio script	-2.14	0.77	audio only	0.87	1.02	0.85	33	0.401	0.30
Medium	1431	-1.27	0.67	audio script	-1.32	0.67	no exposure	0.05	0.95	0.05	31	0.963	0.02
Medium	1431	-1.27	0.67	audio script	-1.87	0.34	baseline	0.59	0.75	0.79	24	0.436	0.32
Medium	1431	-2.14	0.77	audio only	-1.32	0.67	no exposure	-0.83	1.02	-0.81	33	0.426	-0.28
Medium	1431	-2.14	0.77	audio only	-1.87	0.34	baseline	-0.28	0.84	-0.33	25	0.746	-0.13
Medium	1431	-1.32	0.67	no exposure	-1.87	0.34	baseline	0.55	0.75	0.73	24	0.473	0.30
Medium	1533	0.91	0.53	audio script	0.39	0.50	audio only	0.52	0.73	0.71	33	0.482	0.25
Medium	1533	0.91	0.53	audio script	0.08	0.55	no exposure	0.83	0.76	1.09	31	0.283	0.39
Medium	1533	0.91	0.53	audio script	1.29	0.35	baseline	-0.39	0.63	-0.61	31	0.548	-0.22
Medium	1533	0.39	0.50	audio only	0.08	0.55	no exposure	0.31	0.74	0.42	33	0.674	0.15
Medium	1533	0.39	0.50	audio only	1.29	0.35	baseline	-0.90	0.61	-1.48	37	0.147	-0.49
Medium	1533	0.08	0.55	no exposure	1.29	0.35	baseline	-1.22	0.65	-1.87	30	0.071	-0.68
Medium	2111	0.36	0.52	audio script	-0.10	0.50	audio only	0.46	0.73	0.64	33	0.527	0.22
Medium	2111	0.36	0.52	audio script	0.66	0.54	no exposure	-0.31	0.75	-0.41	31	0.687	-0.15

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	2111	0.36	0.52	audio script	0.85	0.32	baseline	-0.50	0.61	-0.81	28	0.424	-0.31
Medium	2111	-0.10	0.50	audio only	0.66	0.54	no exposure	-0.77	0.74	-1.04	33	0.305	-0.36
Medium	2111	-0.10	0.50	audio only	0.85	0.32	baseline	-0.96	0.59	-1.62	33	0.115	-0.56
Medium	2111	0.66	0.54	no exposure	0.85	0.32	baseline	-0.19	0.63	-0.30	27	0.764	-0.12
Medium	2221	-1.27	0.67	audio script	-0.36	0.51	audio only	-0.91	0.85	-1.08	30	0.290	-0.39
Medium	2221	-1.27	0.67	audio script	-0.23	0.56	no exposure	-1.05	0.87	-1.20	30	0.241	-0.44
Medium	2221	-1.27	0.67	audio script	-1.36	0.30	baseline	0.09	0.74	0.12	22	0.903	0.05
Medium	2221	-0.36	0.51	audio only	-0.23	0.56	no exposure	-0.14	0.76	-0.18	33	0.858	-0.06
Medium	2221	-0.36	0.51	audio only	-1.36	0.30	baseline	1.00	0.59	1.68	31	0.103	0.60
Medium	2221	-0.23	0.56	no exposure	-1.36	0.30	baseline	1.14	0.63	1.79	26	0.085	0.70
Medium	2321	0.08	0.53	audio script	-0.10	0.50	audio only	0.19	0.73	0.25	33	0.801	0.09
Medium	2321	0.08	0.53	audio script	-0.90	0.62	no exposure	0.99	0.81	1.21	31	0.235	0.43
Medium	2321	0.08	0.53	audio script	0.07	0.28	baseline	0.01	0.60	0.01	25	0.989	0.00
Medium	2321	-0.10	0.50	audio only	-0.90	0.62	no exposure	0.80	0.79	1.01	31	0.322	0.36
Medium	2321	-0.10	0.50	audio only	0.07	0.28	baseline	-0.18	0.58	-0.31	30	0.760	-0.11

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	2321	-0.90	0.62	no exposure	0.07	0.28	baseline	-0.98	0.68	-1.44	23	0.162	-0.60
Medium	2421	0.36	0.52	audio script	0.14	0.50	audio only	0.21	0.72	0.30	33	0.769	0.10
Medium	2421	0.36	0.52	audio script	0.66	0.54	no exposure	-0.31	0.75	-0.41	31	0.687	-0.15
Medium	2421	0.36	0.52	audio script	-0.62	0.28	baseline	0.98	0.59	1.65	25	0.111	0.66
Medium	2421	0.14	0.50	audio only	0.66	0.54	no exposure	-0.52	0.74	-0.71	33	0.484	-0.25
Medium	2421	0.14	0.50	audio only	-0.62	0.28	baseline	0.76	0.57	1.34	30	0.190	0.49
Medium	2421	0.66	0.54	no exposure	-0.62	0.28	baseline	1.29	0.61	2.11	25	0.045	0.84
Medium	2532	0.08	0.53	audio script	0.14	0.50	audio only	-0.06	0.73	-0.09	33	0.931	-0.03
Medium	2532	0.08	0.53	audio script	0.96	0.55	no exposure	-0.88	0.76	-1.15	31	0.258	-0.41
Medium	2532	0.08	0.53	audio script	0.15	0.28	baseline	-0.07	0.60	-0.12	25	0.907	-0.05
Medium	2532	0.14	0.50	audio only	0.96	0.55	no exposure	-0.82	0.74	-1.10	33	0.278	-0.38
Medium	2532	0.14	0.50	audio only	0.15	0.28	baseline	-0.01	0.57	-0.01	30	0.990	0.00
Medium	2532	0.96	0.55	no exposure	0.15	0.28	baseline	0.81	0.62	1.31	25	0.202	0.52
Medium	2632	-1.27	0.67	audio script	-1.25	0.59	audio only	-0.02	0.90	-0.02	32	0.984	-0.01
Medium	2632	-1.27	0.67	audio script	-0.23	0.56	no exposure	-1.05	0.87	-1.20	30	0.241	-0.44

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	2632	-1.27	0.67	audio script	-0.78	0.28	baseline	-0.50	0.73	-0.68	21	0.504	-0.30
Medium	2632	-1.25	0.59	audio only	-0.23	0.56	no exposure	-1.03	0.81	-1.26	33	0.216	-0.44
Medium	2632	-1.25	0.59	audio only	-0.78	0.28	baseline	-0.48	0.66	-0.73	26	0.474	-0.29
Medium	2632	-0.23	0.56	no exposure	-0.78	0.28	baseline	0.55	0.62	0.88	24	0.387	0.36
Medium	3111	0.37	0.47	audio script	1.37	0.53	audio only	-1.00	0.70	-1.43	39	0.162	-0.46
Medium	3111	0.37	0.47	audio script	1.34	0.58	no exposure	-0.97	0.75	-1.30	32	0.201	-0.46
Medium	3111	0.37	0.47	audio script	0.81	0.32	baseline	-0.45	0.56	-0.79	39	0.432	-0.25
Medium	3111	1.37	0.53	audio only	1.34	0.58	no exposure	0.03	0.79	0.04	34	0.970	0.01
Medium	3111	1.37	0.53	audio only	0.81	0.32	baseline	0.56	0.62	0.90	35	0.373	0.30
Medium	3111	1.34	0.58	no exposure	0.81	0.32	baseline	0.53	0.66	0.79	26	0.435	0.31
Medium	3221	-0.50	0.47	audio script	0.01	0.45	audio only	-0.51	0.66	-0.78	39	0.442	-0.25
Medium	3221	-0.50	0.47	audio script	-0.43	0.55	no exposure	-0.07	0.72	-0.10	33	0.924	-0.03
Medium	3221	-0.50	0.47	audio script	-0.76	0.31	baseline	0.27	0.57	0.47	38	0.639	0.15
Medium	3221	0.01	0.45	audio only	-0.43	0.55	no exposure	0.44	0.71	0.62	33	0.541	0.22
Medium	3221	0.01	0.45	audio only	-0.76	0.31	baseline	0.78	0.55	1.41	39	0.166	0.45

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	3221	-0.43	0.55	no exposure	-0.76	0.31	baseline	0.34	0.63	0.54	27	0.597	0.21
Medium	3331	0.82	0.49	audio script	0.01	0.45	audio only	0.80	0.67	1.20	39	0.236	0.38
Medium	3331	0.82	0.49	audio script	0.43	0.53	no exposure	0.39	0.72	0.53	34	0.596	0.18
Medium	3331	0.82	0.49	audio script	-0.12	0.30	baseline	0.93	0.57	1.64	36	0.110	0.55
Medium	3331	0.01	0.45	audio only	0.43	0.53	no exposure	-0.42	0.70	-0.60	33	0.555	-0.21
Medium	3331	0.01	0.45	audio only	-0.12	0.30	baseline	0.13	0.54	0.24	38	0.808	0.08
Medium	3331	0.43	0.53	no exposure	-0.12	0.30	baseline	0.55	0.61	0.90	26	0.376	0.35
Medium	3421	-1.91	0.64	audio script	-0.87	0.50	audio only	-1.04	0.81	-1.28	37	0.207	-0.42
Medium	3421	-1.91	0.64	audio script	-1.49	0.67	no exposure	-0.42	0.93	-0.45	35	0.654	-0.15
Medium	3421	-1.91	0.64	audio script	-2.65	0.49	baseline	0.73	0.80	0.91	43	0.367	0.28
Medium	3421	-0.87	0.50	audio only	-1.49	0.67	no exposure	0.62	0.84	0.75	30	0.462	0.27
Medium	3421	-0.87	0.50	audio only	-2.65	0.49	baseline	1.78	0.70	2.55	56	0.013	0.68
Medium	3421	-1.49	0.67	no exposure	-2.65	0.49	baseline	1.15	0.83	1.39	34	0.173	0.48
Medium	3533	0.82	0.49	audio script	2.03	0.64	audio only	-1.22	0.80	-1.52	37	0.138	-0.50
Medium	3533	0.82	0.49	audio script	1.70	0.62	no exposure	-0.89	0.79	-1.12	31	0.271	-0.40

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	3533	0.82	0.49	audio script	1.49	0.36	baseline	-0.68	0.61	-1.12	43	0.270	-0.34
Medium	3533	2.03	0.64	audio only	1.70	0.62	no exposure	0.33	0.89	0.37	35	0.715	0.13
Medium	3533	2.03	0.64	audio only	1.49	0.36	baseline	0.54	0.73	0.74	33	0.467	0.26
Medium	3533	1.70	0.62	no exposure	1.49	0.36	baseline	0.21	0.72	0.29	27	0.773	0.11
Medium	4111	-0.73	0.49	audio script	-1.43	0.57	audio only	0.71	0.75	0.94	39	0.351	0.30
Medium	4111	-0.73	0.49	audio script	-1.49	0.67	no exposure	0.77	0.83	0.93	30	0.361	0.34
Medium	4111	-0.73	0.49	audio script	-0.12	0.30	baseline	-0.61	0.57	-1.06	36	0.295	-0.35
Medium	4111	-1.43	0.57	audio only	-1.49	0.67	no exposure	0.06	0.88	0.07	33	0.944	0.02
Medium	4111	-1.43	0.57	audio only	-0.12	0.30	baseline	-1.31	0.64	-2.04	31	0.050	-0.73
Medium	4111	-1.49	0.67	no exposure	-0.12	0.30	baseline	-1.38	0.74	-1.87	22	0.075	-0.80
Medium	4231	-0.50	0.47	audio script	-0.41	0.47	audio only	-0.09	0.66	-0.13	39	0.896	-0.04
Medium	4231	-0.50	0.47	audio script	-0.43	0.55	no exposure	-0.07	0.72	-0.10	33	0.924	-0.03
Medium	4231	-0.50	0.47	audio script	0.33	0.30	baseline	-0.83	0.56	-1.47	37	0.149	-0.48
Medium	4231	-0.41	0.47	audio only	-0.43	0.55	no exposure	0.02	0.72	0.03	33	0.980	0.01
Medium	4231	-0.41	0.47	audio only	0.33	0.30	baseline	-0.74	0.56	-1.33	37	0.191	-0.44

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	4231	-0.43	0.55	no exposure	0.33	0.30	baseline	-0.76	0.63	-1.21	26	0.237	-0.47
Medium	4321	1.33	0.53	audio script	0.87	0.48	audio only	0.46	0.72	0.64	39	0.523	0.20
Medium	4321	1.33	0.53	audio script	0.72	0.54	no exposure	0.61	0.76	0.81	35	0.425	0.27
Medium	4321	1.33	0.53	audio script	1.02	0.33	baseline	0.31	0.63	0.49	36	0.627	0.16
Medium	4321	0.87	0.48	audio only	0.72	0.54	no exposure	0.15	0.72	0.21	34	0.837	0.07
Medium	4321	0.87	0.48	audio only	1.02	0.33	baseline	-0.16	0.58	-0.27	39	0.791	-0.09
Medium	4321	0.72	0.54	no exposure	1.02	0.33	baseline	-0.31	0.63	-0.48	28	0.633	-0.18
Medium	4421	-1.24	0.53	audio script	-0.41	0.47	audio only	-0.83	0.71	-1.17	39	0.247	-0.37
Medium	4421	-1.24	0.53	audio script	-0.13	0.54	no exposure	-1.11	0.76	-1.47	35	0.152	-0.50
Medium	4421	-1.24	0.53	audio script	-0.30	0.30	baseline	-0.94	0.61	-1.54	33	0.134	-0.54
Medium	4421	-0.41	0.47	audio only	-0.13	0.54	no exposure	-0.28	0.71	-0.39	33	0.701	-0.14
Medium	4421	-0.41	0.47	audio only	-0.30	0.30	baseline	-0.11	0.56	-0.20	37	0.846	-0.07
Medium	4421	-0.13	0.54	no exposure	-0.30	0.30	baseline	0.17	0.61	0.27	26	0.789	0.11
Medium	4521	0.15	0.46	audio script	-0.41	0.47	audio only	0.56	0.66	0.85	39	0.399	0.27
Medium	4521	0.15	0.46	audio script	0.43	0.53	no exposure	-0.28	0.70	-0.40	33	0.694	-0.14

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	4521	0.15	0.46	audio script	1.02	0.33	baseline	-0.87	0.57	-1.54	41	0.132	-0.48
Medium	4521	-0.41	0.47	audio only	0.43	0.53	no exposure	-0.84	0.71	-1.19	34	0.244	-0.41
Medium	4521	-0.41	0.47	audio only	1.02	0.33	baseline	-1.43	0.57	-2.51	40	0.016	-0.79
Medium	4521	0.43	0.53	no exposure	1.02	0.33	baseline	-0.59	0.62	-0.95	29	0.351	-0.35
Medium	4633	-3.16	1.04	audio script	-2.27	0.75	audio only	-0.90	1.28	-0.70	36	0.489	-0.23
Medium	4633	-3.16	1.04	audio script	0.15	0.53	no exposure	-3.31	1.16	-2.85	29	0.008	-1.06
Medium	4633	-3.16	1.04	audio script	-0.30	0.30	baseline	-2.86	1.08	-2.66	23	0.014	-1.11
Medium	4633	-2.27	0.75	audio only	0.15	0.53	no exposure	-2.42	0.92	-2.62	34	0.013	-0.90
Medium	4633	-2.27	0.75	audio only	-0.30	0.30	baseline	-1.97	0.81	-2.43	26	0.023	-0.95
Medium	4633	0.15	0.53	no exposure	-0.30	0.30	baseline	0.45	0.61	0.74	27	0.467	0.28
Low	1111	-1.69	1.05	audio script	-1.04	0.75	audio only	-0.64	1.29	-0.50	25	0.623	-0.20
Low	1111	-1.69	1.05	audio script	-1.61	1.05	no exposure	-0.08	1.48	-0.05	27	0.959	-0.02
Low	1111	-1.69	1.05	audio script	-1.20	0.39	baseline	-0.48	1.12	-0.43	16	0.674	-0.22
Low	1111	-1.04	0.75	audio only	-1.61	1.05	no exposure	0.57	1.29	0.44	28	0.664	0.17
Low	1111	-1.04	0.75	audio only	-1.20	0.39	baseline	0.16	0.85	0.19	29	0.850	0.07

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	1111	-1.61	1.05	no exposure	-1.20	0.39	baseline	-0.40	1.12	-0.36	19	0.722	-0.17
Low	1221	1.65	0.58	audio script	1.04	0.46	audio only	0.61	0.74	0.83	27	0.414	0.32
Low	1221	1.65	0.58	audio script	0.09	0.60	no exposure	1.56	0.83	1.88	27	0.071	0.72
Low	1221	1.65	0.58	audio script	0.99	0.28	baseline	0.66	0.64	1.04	19	0.313	0.48
Low	1221	1.04	0.46	audio only	0.09	0.60	no exposure	0.95	0.76	1.25	29	0.220	0.46
Low	1221	1.04	0.46	audio only	0.99	0.28	baseline	0.05	0.54	0.10	34	0.923	0.03
Low	1221	0.09	0.60	no exposure	0.99	0.28	baseline	-0.90	0.66	-1.35	22	0.191	-0.58
Low	1321	-0.89	0.78	audio script	-0.20	0.57	audio only	-0.68	0.97	-0.71	25	0.487	-0.28
Low	1321	-0.89	0.78	audio script	-0.31	0.66	no exposure	-0.58	1.02	-0.56	26	0.577	-0.22
Low	1321	-0.89	0.78	audio script	-2.01	0.53	baseline	1.13	0.94	1.20	26	0.242	0.47
Low	1321	-0.20	0.57	audio only	-0.31	0.66	no exposure	0.10	0.87	0.12	31	0.907	0.04
Low	1321	-0.20	0.57	audio only	-2.01	0.53	baseline	1.81	0.78	2.33	52	0.024	0.65
Low	1321	-0.31	0.66	no exposure	-2.01	0.53	baseline	1.71	0.85	2.01	36	0.052	0.67
Low	1431	0.04	0.61	audio script	-0.57	0.64	audio only	0.60	0.88	0.68	31	0.499	0.24
Low	1431	0.04	0.61	audio script	-0.31	0.66	no exposure	0.35	0.90	0.38	27	0.704	0.15

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	1431	0.04	0.61	audio script	-0.68	0.34	baseline	0.72	0.70	1.03	21	0.317	0.45
Low	1431	-0.57	0.64	audio only	-0.31	0.66	no exposure	-0.26	0.92	-0.28	33	0.781	-0.10
Low	1431	-0.57	0.64	audio only	-0.68	0.34	baseline	0.11	0.72	0.16	30	0.878	0.06
Low	1431	-0.31	0.66	no exposure	-0.68	0.34	baseline	0.37	0.74	0.50	23	0.624	0.21
Low	1533	-0.37	0.67	audio script	0.10	0.53	audio only	-0.46	0.85	-0.54	27	0.591	-0.21
Low	1533	-0.37	0.67	audio script	-0.82	0.78	no exposure	0.45	1.02	0.44	27	0.666	0.17
Low	1533	-0.37	0.67	audio script	0.01	0.30	baseline	-0.38	0.73	-0.52	18	0.611	-0.25
Low	1533	0.10	0.53	audio only	-0.82	0.78	no exposure	0.91	0.94	0.97	27	0.340	0.37
Low	1533	0.10	0.53	audio only	0.01	0.30	baseline	0.08	0.60	0.14	31	0.890	0.05
Low	1533	-0.82	0.78	no exposure	0.01	0.30	baseline	-0.83	0.83	-1.00	19	0.332	-0.46
Low	2111	0.71	0.56	audio script	0.36	0.50	audio only	0.35	0.75	0.47	29	0.641	0.17
Low	2111	0.71	0.56	audio script	-0.31	0.66	no exposure	1.02	0.87	1.18	27	0.250	0.45
Low	2111	0.71	0.56	audio script	-0.68	0.34	baseline	1.39	0.65	2.12	23	0.045	0.88
Low	2111	0.36	0.50	audio only	-0.31	0.66	no exposure	0.67	0.83	0.80	29	0.429	0.30
Low	2111	0.36	0.50	audio only	-0.68	0.34	baseline	1.04	0.60	1.72	37	0.094	0.57

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	2111	-0.31	0.66	no exposure	-0.68	0.34	baseline	0.37	0.74	0.50	23	0.624	0.21
Low	2221	-0.89	0.78	audio script	-2.53>	1.44	audio only	1.64	1.64	1.00	28	0.324	0.38
Low	2221	-0.89	0.78	audio script	-1.61	1.05	no exposure	0.72	1.31	0.55	26	0.585	0.22
Low	2221	-0.89	0.78	audio script	-4.19>	1.42	baseline	3.31	1.62	2.04	66	0.046	0.50
Low	2221	-2.53>	1.44	audio only	-1.61	1.05	no exposure	-0.92	1.78	-0.52	32	0.609	-0.18
Low	2221	-2.53>	1.44	audio only	-4.19>	1.42	baseline	1.67	2.02	0.82	55	0.414	0.22
Low	2221	-1.61	1.05	no exposure	-4.19>	1.42	baseline	2.59	1.77	1.46	62	0.149	0.37
Low	2321	-0.37	0.67	audio script	-0.57	0.64	audio only	0.20	0.92	0.21	30	0.833	0.08
Low	2321	-0.37	0.67	audio script	1.01	0.53	no exposure	-1.38	0.85	-1.62	25	0.118	-0.65
Low	2321	-0.37	0.67	audio script	-0.68	0.34	baseline	0.31	0.75	0.41	20	0.685	0.18
Low	2321	-0.57	0.64	audio only	1.01	0.53	no exposure	-1.58	0.83	-1.91	33	0.065	-0.66
Low	2321	-0.57	0.64	audio only	-0.68	0.34	baseline	0.11	0.72	0.16	30	0.878	0.06
Low	2321	1.01	0.53	no exposure	-0.68	0.34	baseline	1.69	0.63	2.70	28	0.012	1.02
Low	2421	0.39	0.58	audio script	2.40	0.53	audio only	-2.01	0.78	-2.57	29	0.016	-0.95
Low	2421	0.39	0.58	audio script	1.29	0.52	no exposure	-0.90	0.78	-1.16	27	0.258	-0.45

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	2421	0.39	0.58	audio script	0.51	0.28	baseline	-0.12	0.64	-0.18	19	0.857	-0.08
Low	2421	2.40	0.53	audio only	1.29	0.52	no exposure	1.11	0.74	1.49	33	0.145	0.52
Low	2421	2.40	0.53	audio only	0.51	0.28	baseline	1.89	0.60	3.17	30	0.004	1.16
Low	2421	1.29	0.52	no exposure	0.51	0.28	baseline	0.78	0.60	1.31	24	0.201	0.53
Low	2532	-0.37	0.67	audio script	-0.20	0.57	audio only	-0.16	0.88	-0.19	28	0.854	-0.07
Low	2532	-0.37	0.67	audio script	0.09	0.60	no exposure	-0.46	0.90	-0.51	27	0.616	-0.20
Low	2532	-0.37	0.67	audio script	-0.68	0.34	baseline	0.31	0.75	0.41	20	0.685	0.18
Low	2532	-0.20	0.57	audio only	0.09	0.60	no exposure	-0.29	0.83	-0.35	33	0.726	-0.12
Low	2532	-0.20	0.57	audio only	-0.68	0.34	baseline	0.47	0.66	0.71	33	0.481	0.25
Low	2532	0.09	0.60	no exposure	-0.68	0.34	baseline	0.77	0.69	1.11	25	0.277	0.44
Low	2632	0.04	0.61	audio script	-0.20	0.57	audio only	0.24	0.84	0.29	29	0.773	0.11
Low	2632	0.04	0.61	audio script	-1.61	1.05	no exposure	1.65	1.21	1.36	23	0.188	0.57
Low	2632	0.04	0.61	audio script	-0.68	0.34	baseline	0.72	0.70	1.03	21	0.317	0.45
Low	2632	-0.20	0.57	audio only	-1.61	1.05	no exposure	1.40	1.19	1.18	23	0.251	0.49
Low	2632	-0.20	0.57	audio only	-0.68	0.34	baseline	0.47	0.66	0.71	33	0.481	0.25

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	2632	-1.61	1.05	no exposure	-0.68	0.34	baseline	-0.93	1.10	-0.85	18	0.409	-0.40
Low	3111	1.10	0.51	audio script	1.60	0.48	audio only	-0.50	0.70	-0.71	33	0.484	-0.25
Low	3111	1.10	0.51	audio script	0.26	0.52	no exposure	0.84	0.73	1.16	33	0.254	0.40
Low	3111	1.10	0.51	audio script	0.27	0.30	baseline	0.83	0.59	1.42	28	0.168	0.54
Low	3111	1.60	0.48	audio only	0.26	0.52	no exposure	1.34	0.71	1.88	35	0.068	0.64
Low	3111	1.60	0.48	audio only	0.27	0.30	baseline	1.33	0.57	2.33	32	0.026	0.82
Low	3111	0.26	0.52	no exposure	0.27	0.30	baseline	-0.01	0.60	-0.02	30	0.987	-0.01
Low	3221	-0.41	0.65	audio script	-0.14	0.57	audio only	-0.27	0.87	-0.31	32	0.759	-0.11
Low	3221	-0.41	0.65	audio script	-0.35	0.59	no exposure	-0.07	0.88	-0.08	33	0.940	-0.03
Low	3221	-0.41	0.65	audio script	-0.78	0.36	baseline	0.36	0.75	0.49	26	0.630	0.19
Low	3221	-0.14	0.57	audio only	-0.35	0.59	no exposure	0.20	0.82	0.25	35	0.807	0.08
Low	3221	-0.14	0.57	audio only	-0.78	0.36	baseline	0.63	0.68	0.93	33	0.359	0.32
Low	3221	-0.35	0.59	no exposure	-0.78	0.36	baseline	0.43	0.69	0.62	32	0.539	0.22
Low	3331	0.29	0.55	audio script	-0.14	0.57	audio only	0.43	0.79	0.55	33	0.589	0.19
Low	3331	0.29	0.55	audio script	0.77	0.49	no exposure	-0.48	0.74	-0.65	33	0.519	-0.23

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	3331	0.29	0.55	audio script	-0.31	0.33	baseline	0.60	0.64	0.93	28	0.359	0.35
Low	3331	-0.14	0.57	audio only	0.77	0.49	no exposure	-0.91	0.76	-1.21	35	0.234	-0.41
Low	3331	-0.14	0.57	audio only	-0.31	0.33	baseline	0.16	0.66	0.25	30	0.807	0.09
Low	3331	0.77	0.49	no exposure	-0.31	0.33	baseline	1.08	0.59	1.83	34	0.076	0.63
Low	3421	-0.91	0.76	audio script	-0.14	0.57	audio only	-0.76	0.96	-0.80	30	0.432	-0.29
Low	3421	-0.91	0.76	audio script	-2.76>	1.45	no exposure	1.86	1.64	1.13	27	0.267	0.43
Low	3421	-0.91	0.76	audio script	-1.88	0.53	baseline	0.97	0.93	1.05	32	0.303	0.37
Low	3421	-0.14	0.57	audio only	-2.76>	1.45	no exposure	2.62	1.56	1.68	23	0.106	0.70
Low	3421	-0.14	0.57	audio only	-1.88	0.53	baseline	1.73	0.78	2.22	48	0.031	0.64
Low	3421	-2.76>	1.45	no exposure	-1.88	0.53	baseline	-0.88	1.54	-0.57	22	0.572	-0.24
Low	3533	0.85	0.51	audio script	0.68	0.49	audio only	0.17	0.71	0.24	33	0.815	0.08
Low	3533	0.85	0.51	audio script	0.77	0.49	no exposure	0.08	0.71	0.11	33	0.915	0.04
Low	3533	0.85	0.51	audio script	1.62	0.32	baseline	-0.77	0.60	-1.28	29	0.212	-0.48
Low	3533	0.68	0.49	audio only	0.77	0.49	no exposure	-0.09	0.69	-0.13	35	0.897	-0.04
Low	3533	0.68	0.49	audio only	1.62	0.32	baseline	-0.94	0.58	-1.60	34	0.118	-0.55

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	3533	0.77	0.49	no exposure	1.62	0.32	baseline	-0.85	0.58	-1.45	34	0.157	-0.50
Low	4111	-1.68	1.04	audio script	-0.51	0.64	audio only	-1.17	1.22	-0.96	26	0.345	-0.38
Low	4111	-1.68	1.04	audio script	-2.02	1.05	no exposure	0.34	1.47	0.23	33	0.819	0.08
Low	4111	-1.68	1.04	audio script	-0.65	0.35	baseline	-1.03	1.10	-0.94	19	0.359	-0.43
Low	4111	-0.51	0.64	audio only	-2.02	1.05	no exposure	1.51	1.23	1.23	29	0.227	0.46
Low	4111	-0.51	0.64	audio only	-0.65	0.35	baseline	0.14	0.73	0.19	29	0.849	0.07
Low	4111	-2.02	1.05	no exposure	-0.65	0.35	baseline	-1.37	1.10	-1.24	22	0.227	-0.53
Low	4231	0.85	0.51	audio script	0.68	0.49	audio only	0.17	0.71	0.24	33	0.815	0.08
Low	4231	0.85	0.51	audio script	1.01	0.48	no exposure	-0.16	0.70	-0.23	33	0.821	-0.08
Low	4231	0.85	0.51	audio script	0.45	0.30	baseline	0.40	0.59	0.67	27	0.508	0.26
Low	4231	0.68	0.49	audio only	1.01	0.48	no exposure	-0.33	0.69	-0.48	35	0.637	-0.16
Low	4231	0.68	0.49	audio only	0.45	0.30	baseline	0.23	0.57	0.40	31	0.690	0.14
Low	4231	1.01	0.48	no exposure	0.45	0.30	baseline	0.56	0.57	0.98	32	0.333	0.35
Low	4321	0.29	0.55	audio script	0.43	0.51	audio only	-0.14	0.75	-0.19	33	0.851	-0.07
Low	4321	0.29	0.55	audio script	1.01	0.48	no exposure	-0.72	0.73	-0.98	32	0.335	-0.35

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	4321	0.29	0.55	audio script	0.71	0.29	baseline	-0.42	0.62	-0.67	25	0.507	-0.27
Low	4321	0.43	0.51	audio only	1.01	0.48	no exposure	-0.58	0.70	-0.82	35	0.417	-0.28
Low	4321	0.43	0.51	audio only	0.71	0.29	baseline	-0.28	0.59	-0.47	30	0.639	-0.17
Low	4321	1.01	0.48	no exposure	0.71	0.29	baseline	0.30	0.57	0.53	32	0.602	0.19
Low	4421	0.29	0.55	audio script	-0.51	0.64	audio only	0.80	0.84	0.95	33	0.351	0.33
Low	4421	0.29	0.55	audio script	0.26	0.52	no exposure	0.03	0.76	0.04	33	0.972	0.01
Low	4421	0.29	0.55	audio script	0.09	0.31	baseline	0.20	0.63	0.32	26	0.753	0.13
Low	4421	-0.51	0.64	audio only	0.26	0.52	no exposure	-0.77	0.82	-0.94	34	0.356	-0.32
Low	4421	-0.51	0.64	audio only	0.09	0.31	baseline	-0.60	0.71	-0.84	26	0.406	-0.33
Low	4421	0.26	0.52	no exposure	0.09	0.31	baseline	0.17	0.60	0.29	31	0.777	0.10
Low	4521	0.58	0.53	audio script	0.16	0.53	audio only	0.42	0.75	0.56	33	0.582	0.19
Low	4521	0.58	0.53	audio script	-0.02	0.55	no exposure	0.60	0.76	0.79	33	0.435	0.28
Low	4521	0.58	0.53	audio script	0.09	0.31	baseline	0.49	0.61	0.80	27	0.429	0.31
Low	4521	0.16	0.53	audio only	-0.02	0.55	no exposure	0.18	0.77	0.24	35	0.811	0.08
Low	4521	0.16	0.53	audio only	0.09	0.31	baseline	0.07	0.61	0.12	30	0.908	0.04

Target proficiency level	Item	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	4521	-0.02	0.55	no exposure	0.09	0.31	baseline	-0.11	0.63	-0.18	29	0.859	-0.07
Low	4633	-0.91	0.76	audio script	-0.99	0.76	audio only	0.08	1.08	0.08	33	0.940	0.03
Low	4633	-0.91	0.76	audio script	-0.35	0.59	no exposure	-0.56	0.97	-0.58	30	0.567	-0.21
Low	4633	-0.91	0.76	audio script	-0.31	0.33	baseline	-0.60	0.83	-0.72	22	0.479	-0.31
Low	4633	-0.99	0.76	audio only	-0.35	0.59	no exposure	-0.64	0.96	-0.67	34	0.508	-0.23
Low	4633	-0.99	0.76	audio only	-0.31	0.33	baseline	-0.68	0.82	-0.83	24	0.416	-0.34
Low	4633	-0.35	0.59	no exposure	-0.31	0.33	baseline	-0.04	0.67	-0.06	29	0.953	-0.02

Appendix B.9 Pairwise Bias Report for Exposure Condition and Item Type

Target item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Main idea	-0.12	0.17	audio script	0.18	0.16	audio only	-0.29	0.24	-1.24	421	0.215	-0.12
Main idea	-0.12	0.17	audio script	-0.28	0.17	no exposure	0.16	0.24	0.65	409	0.515	0.06
Main idea	-0.12	0.17	audio script	0.06	0.10	baseline	-0.18	0.20	-0.92	343	0.360	-0.10
Main idea	0.18	0.16	audio only	-0.28	0.17	no exposure	0.45	0.24	1.89	424	0.059	0.18
Main idea	0.18	0.16	audio only	0.06	0.10	baseline	0.11	0.19	0.60	385	0.548	0.06
Main idea	-0.28	0.17	no exposure	0.06	0.10	baseline	-0.34	0.20	-1.70	343	0.090	-0.18
Explicit detail	0.09	0.11	audio script	0.05	0.10	audio only	0.04	0.15	0.29	1057	0.770	0.02
Explicit detail	0.09	0.11	audio script	0.08	0.11	no exposure	0.01	0.15	0.05	1027	0.959	0.00
Explicit detail	0.09	0.11	audio script	-0.07	0.06	baseline	0.16	0.12	1.32	860	0.187	0.09
Explicit detail	0.05	0.10	audio only	0.08	0.11	no exposure	-0.04	0.15	-0.24	1067	0.811	-0.01
Explicit detail	0.05	0.10	audio only	-0.07	0.06	baseline	0.12	0.12	0.99	959	0.323	0.06
Explicit detail	0.08	0.11	no exposure	-0.07	0.06	baseline	0.15	0.12	1.25	871	0.212	0.08
Implicit detail	-0.06	0.12	audio script	-0.14	0.11	audio only	0.08	0.16	0.49	846	0.625	0.03
Implicit detail	-0.06	0.12	audio script	0.03	0.12	no exposure	-0.08	0.17	-0.50	821	0.618	-0.03

Target item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Implicit detail	-0.06	0.12	audio script	0.06	0.07	baseline	-0.12	0.14	-0.86	692	0.392	-0.07
Implicit detail	-0.14	0.11	audio only	0.03	0.12	no exposure	-0.16	0.17	-0.99	854	0.322	-0.07
Implicit detail	-0.14	0.11	audio only	0.06	0.07	baseline	-0.20	0.13	-1.48	767	0.140	-0.11
Implicit detail	0.03	0.12	no exposure	0.06	0.07	baseline	-0.03	0.14	-0.24	701	0.810	-0.02

Appendix B.10 Pairwise Bias Report for Exposure Condition, Item Type, and Listening Proficiency

Target proficiency	Item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	main idea	-0.19	0.29	audio script only	0.56	0.34	audio only	-0.75	0.45	-1.69	125	0.094	-0.30
High	main idea	-0.19	0.29	audio script	-0.11	0.29	no exposure	-0.09	0.41	-0.21	133	0.836	-0.04
High	main idea	-0.19	0.29	audio script	0.47	0.20	baseline	-0.66	0.36	-1.84	131	0.067	-0.32
High	main idea	0.56	0.34	audio only	-0.11	0.29	no exposure	0.67	0.44	1.51	127	0.133	0.27
High	main idea	0.56	0.34	audio only	0.47	0.20	baseline	0.10	0.39	0.24	112	0.808	0.05
High	main idea	-0.11	0.29	no exposure	0.47	0.20	baseline	-0.57	0.35	-1.63	142	0.106	-0.27
High	explicit detail	0.11	0.19	audio script	-0.44	0.18	audio only	0.54	0.26	2.08	322	0.038	0.23
High	explicit detail	0.11	0.19	audio script	-0.20	0.18	no exposure	0.30	0.26	1.18	335	0.240	0.13
High	explicit detail	0.11	0.19	audio script	0.14	0.12	baseline	-0.03	0.22	-0.13	302	0.897	-0.01
High	explicit detail	-0.44	0.18	audio only	-0.20	0.18	no exposure	-0.24	0.25	-0.96	330	0.340	-0.11
High	explicit detail	-0.44	0.18	audio only	0.14	0.12	baseline	-0.57	0.22	-2.64	300	0.009	-0.30
High	explicit detail	-0.20	0.18	no exposure	0.14	0.12	baseline	-0.33	0.21	-1.57	340	0.118	-0.17
High	implicit detail	0.04	0.21	audio script	-0.03	0.22	audio only	0.06	0.30	0.21	257	0.833	0.03
High	implicit detail	0.04	0.21	audio script	-0.29	0.20	no exposure	0.32	0.29	1.12	267	0.263	0.14

Target proficiency	Item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
High	implicit detail	0.04	0.21	audio script	0.06	0.14	baseline	-0.02	0.25	-0.09	247	0.930	-0.01
High	implicit detail	-0.03	0.22	audio only	-0.29	0.20	no exposure	0.26	0.29	0.88	260	0.377	0.11
High	implicit detail	-0.03	0.22	audio only	0.06	0.14	baseline	-0.09	0.26	-0.33	232	0.738	-0.04
High	implicit detail	-0.29	0.20	no exposure	0.06	0.14	baseline	-0.35	0.24	-1.44	278	0.150	-0.17
Medium	main idea	-0.19	0.27	audio script	-0.24	0.26	audio only	0.05	0.38	0.13	153	0.898	0.02
Medium	main idea	-0.19	0.27	audio script	-0.16	0.30	no exposure	-0.03	0.40	-0.07	139	0.943	-0.01
Medium	main idea	-0.19	0.27	audio script	0.21	0.15	baseline	-0.40	0.31	-1.28	127	0.203	-0.23
Medium	main idea	-0.24	0.26	audio only	-0.16	0.30	no exposure	-0.08	0.39	-0.20	140	0.845	-0.03
Medium	main idea	-0.24	0.26	audio only	0.21	0.15	baseline	-0.45	0.30	-1.47	137	0.143	-0.25
Medium	main idea	-0.16	0.30	no exposure	0.21	0.15	baseline	-0.37	0.33	-1.11	106	0.270	-0.22
Medium	explicit detail	-0.03	0.17	audio script	0.14	0.16	audio only	-0.17	0.23	-0.72	385	0.471	-0.07
Medium	explicit detail	-0.03	0.17	audio script	0.21	0.18	no exposure	-0.24	0.25	-0.97	352	0.331	-0.10
Medium	explicit detail	-0.03	0.17	audio script	-0.17	0.10	baseline	0.13	0.19	0.69	321	0.489	0.08
Medium	explicit detail	0.14	0.16	audio only	0.21	0.18	no exposure	-0.07	0.24	-0.30	354	0.765	-0.03
Medium	explicit detail	0.14	0.16	audio only	-0.17	0.10	baseline	0.30	0.19	1.61	352	0.108	0.17

Target proficiency	Item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Medium	explicit detail	0.21	0.18	no exposure	-0.17	0.10	baseline	0.37	0.20	1.83	272	0.068	0.22
Medium	implicit detail	-0.24	0.19	audio script	-0.27	0.18	audio only	0.03	0.26	0.11	308	0.909	0.01
Medium	implicit detail	-0.24	0.19	audio script	0.27	0.20	no exposure	-0.52	0.27	-1.89	282	0.060	-0.23
Medium	implicit detail	-0.24	0.19	audio script	0.13	0.11	baseline	-0.37	0.22	-1.71	255	0.089	-0.21
Medium	implicit detail	-0.27	0.18	audio only	0.27	0.20	no exposure	-0.55	0.27	-2.03	284	0.043	-0.24
Medium	implicit detail	-0.27	0.18	audio only	0.13	0.11	baseline	-0.40	0.21	-1.90	277	0.058	-0.23
Medium	implicit detail	0.27	0.20	no exposure	0.13	0.11	baseline	0.15	0.23	0.65	218	0.514	0.09
Low	main idea	0.07	0.32	audio script	0.37	0.28	audio only	-0.30	0.42	-0.72	129	0.472	-0.13
Low	main idea	0.07	0.32	audio script	-0.67	0.36	no exposure	0.74	0.48	1.55	129	0.124	0.27
Low	main idea	0.07	0.32	audio script	-0.43	0.18	baseline	0.50	0.36	1.37	101	0.172	0.27
Low	main idea	0.37	0.28	audio only	-0.67	0.36	no exposure	1.04	0.45	2.31	133	0.022	0.40
Low	main idea	0.37	0.28	audio only	-0.43	0.18	baseline	0.80	0.33	2.45	143	0.016	0.41
Low	main idea	-0.67	0.36	no exposure	-0.43	0.18	baseline	-0.24	0.40	-0.60	104	0.548	-0.12
Low	explicit detail	0.24	0.20	audio script	0.38	0.17	audio only	-0.14	0.26	-0.54	329	0.587	-0.06
Low	explicit detail	0.24	0.20	audio script	0.26	0.19	no exposure	-0.02	0.27	-0.07	324	0.944	-0.01

Target proficiency	Item type	Target measure +	S.E.	Exposure condition	Target measure +	S.E.	Exposure condition	Target contrast +	Joint S.E.	<i>t</i>	<i>d.f.</i>	<i>p</i>	Cohen's <i>d</i>
Low	explicit detail	0.24	0.20	audio script	-0.13	0.11	baseline	0.37	0.22	1.67	250	0.097	0.21
Low	explicit detail	0.38	0.17	audio only	0.26	0.19	no exposure	0.12	0.26	0.48	361	0.630	0.05
Low	explicit detail	0.38	0.17	audio only	-0.13	0.11	baseline	0.52	0.20	2.53	347	0.012	0.27
Low	explicit detail	0.26	0.19	no exposure	-0.13	0.11	baseline	0.39	0.22	1.82	292	0.070	0.21
Low	implicit detail	0.09	0.22	audio script	-0.06	0.20	audio only	0.16	0.30	0.52	269	0.601	0.06
Low	implicit detail	0.09	0.22	audio script	0.10	0.21	no exposure	0.00	0.30	0.00	259	0.998	0.00
Low	implicit detail	0.09	0.22	audio script	-0.02	0.11	baseline	0.11	0.24	0.46	195	0.648	0.07
Low	implicit detail	-0.06	0.20	audio only	0.10	0.21	no exposure	-0.16	0.29	-0.54	292	0.592	-0.06
Low	implicit detail	-0.06	0.20	audio only	-0.02	0.11	baseline	-0.04	0.23	-0.19	256	0.851	-0.02
Low	implicit detail	0.10	0.21	no exposure	-0.02	0.11	baseline	0.11	0.24	0.47	227	0.636	0.06