DIAGNOSING SECOND LANGUAGE PRONUNCIATION

By

Daniel Richard Isbell

A DISSERTATION

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

Second Language Studies – Doctor of Philosophy

2019

ABSTRACT

DIAGNOSING SECOND LANGUAGE PRONUNCIATION

By

Daniel Richard Isbell

Pronunciation presents a significant, persistent challenge to second language (L2) learners (Derwing & Munro, 2015; Piske, McKay, & Flege, 2001). Fortunately, pronunciation instruction works (Lee, Jang, & Plonsky, 2015). However, pronunciation receives relatively little attention in language classrooms. Further complicating the matter are classrooms with learners from diverse linguistic backgrounds and/or differing levels of pronunciation ability, which may impact the effectiveness of one-size-fits-all whole-class pronunciation instruction (e.g., Isbell, Park, & Lee, 2019). Diagnostic language assessment (Alderson, 2005; Alderson, Brunfaut, & Harding, 2014), which prioritizes identifying the specific strengths and weaknesses of learners, is a potentially useful approach to addressing individuals' pronunciation needs.

Addressing these issues, I developed a new diagnostic tool for segmental L2 Korean pronunciation called the Korean Pronunciation Diagnostic (KPD). The KPD consists of two sections, production and perception, each with two tasks that tap into phonological knowledge and abilities. KPD feedback includes a list of a learner's most-difficult phonemes to prioritize in instruction and accuracy scores for production and perception of all phonemes. To evaluate the quality and usefulness of the test, I constructed a validity argument (Kane, 2013) for the interpretation and use of KPD scores, which included inferences on the operationalization of relevant theory, evaluation of observations, generalization of scores, explanation of scores with respect to underlying theory, extrapolation of scores to general language use, utilization of feedback by stakeholders, and the usefulness and impact of applying scores.

I sought support for these inferences from two main sources: field testing with 198 L2 Korean learners and interviews with 21 learners and one Korean language teacher. Field testing participants completed a background questionnaire, pronunciation self-assessment, independent speaking task, the KPD, and a standardized measure of oral proficiency. Interview participants completed an initial semi-structured interview where they received their KPD score report; 14 learners completed a follow-up interview approximately 3 months later where they discussed recent pronunciation learning activity and took the KPD again. I used several quantitative techniques, including measurement analyses (classical test theory and Rasch), correlations, cluster analysis, to analyze the field testing data. I analyzed interview data qualitatively.

Support for the operationalization, generalization, and explanation inferences was strong, supporting the interpretation of KPD scores as strengths and weaknesses in the production and perception of Korean phonemes. Support for the extrapolation inference was positive but limited. Correlations between KPD scores and learner self-assessments were positive but not large, as learners had limited awareness of their fine-grained pronunciation abilities. Similarly, the KPD's discrete and delimited measurements of phoneme accuracy had limited overlap with pronunciation in spontaneous, meaning-focused speech. The utilization inference was well-supported, though improvements to the KPD score report could further enhance stakeholder interpretation of results. Finally, positive but limited evidence for the usefulness and impact of the KPD was found: Findings suggest that learner application of KPD results has the potential to support pronunciation development, but this is conditional on learner effort. I determined that more evidence is needed to sufficiently support this inference. Overall, the interpretation and use of KPD scores is supported, but future development and research efforts should focus on the effective application of the KPD's diagnostic feedback.

For Kyujin.

ACKNOWLEDGEMENTS

This dissertation is the capstone of one incredible year. In this past year, I temporarily relocated to South Korea, became a father, started data collection, navigated the academic job market, finished data collection, completed analyses, finished writing this, and returned to the United States (in roughly that order). I wouldn't have been able to do this without lots of help.

First, this dissertation was made possible thanks to the financial support of the Fulbright U.S. Student program, an Educational Testing Service TOEFL Doctoral Dissertation Research Support Grant, and several sources at MSU: a Research Enhancement Award from the Graduate College, a Dissertation Completion Fellowship from the College of Arts and Letters and the Graduate College, funds from the College of Arts and Letters and the Second Language Studies program. Additionally, the Asian Studies Center at MSU and the U.S. Department of Education provided me with a Foreign Language and Area Studies Summer Fellowship to study Korean.

I have benefited immensely from Dr. Paula Winke's guidance throughout my doctoral studies. Paula was not just an extremely knowledgeable expert in the field of language testing perfectly suited to chair this dissertation. She was also the most positive, supportive advisor a graduate student could have. I learned so much about language assessment, the language testing industry, doing research, writing about research, and applying for funding from Paula, but my greatest takeaway from her mentorship is how to be a good mentor. Through her example, I learned about connecting people with opportunities, promoting the work and talents of others, and supporting someone both as a scholar and as a well-rounded human being. My go-to heuristic when mentoring students in the future will be the question "What would Paula do?"

I wish to express my gratitude to the other members of my committee. Complementing Paula's assessment knowledge, I was able to assemble a dream team of topical expertise perfectly aligned to this dissertation: Dr. Susan Gass (SLA, and so much more), Dr. Debra Hardison (L2 pronunciation and speech perception), Dr. Junkyu Lee (L2 pronunciation, language learning and research environment in Korea), and Dr. Shawn Loewen (instructed SLA, research methodology). Advice I received from my committee was indispensable. I am especially appreciative of Debra's perspective and feedback in the early stages of developing my ideas and Junkyu's support and guidance during my time in Korea. I am grateful to Sue and Shawn for the opportunity to collaborate on other projects during my time at MSU; I learned a great deal about doing research that helped me carry out this project.

Many others have helped me along the way. Thank you to my fellow SLS students for being great colleagues and friends, and especially Jin Soo Choi, Kathy Minhye Kim, Susie Kim, Jongbong Lee, Shinhye Lee, Jungmin Lim, and Myeongeun Son for their generous feedback on instruments help with trialing. Thanks to Dustin Crowther for always being available to talk L2 pronunciation and bounce ideas around, to Dr. Jai Ok Shim and Heidi Little at the Korean-American Educational Commission and Keehye Shin at Hankuk University of Foreign Studies for logistical support, and to HUFS TESOL graduate students Yerin An, Haewon Kim, Sohee Lee, Minchae Shin, and YounJu Yoo for coding and transcription assistance.

Last and certainly not least, I want to thank my family. I am thankful to my in-laws for their hospitality and support over the last year in Korea. To my son, Euan: Thank you for remembering me after I came back from trips, and thanks for being a good sleeper! It meant a lot to me. To my wife, Kyujin: Thank you for being Euan's mom, thank you for being my partner in life, and thank you for being my language expert. I couldn't have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ACTFL	American Council on Teaching Foreign Languages
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
CAH	Contrastive Analysis Hypothesis
CDA	Cognitive Diagnostic Assessment
CDM	Cognitive Diagnostic Model
CEFR	Common European Framework of Reference
CTT	Classical Test Theory
DLA	Diagnostic Language Assessment
EIT	Elicited Imitation Test
F_0	Fundamental Frequency (pitch)
FL	Functional Load (for Foreign Language, see L2)
HCA	Hierarchical Cluster Analysis
HVPT	High-Variability Phonetic Training
ICC	Intraclass Correlation Coefficient
IELTS	International English Language Testing System
IIF	Item Information Function
IPA	International Phonetic Alphabet
IRT	Item Response Theory
ISLA	Instructed Second Language Acquisition
KFL	Korean as a Foreign Language

KSL	Korean as a Second Language
KPD	Korean Pronunciation Diagnostic
L1	First Language
L2	Second Language (includes Foreign Language)
L3	Third Language (L3+ = third or later language)
LBQ	Language Background Questionnaire
NNS	Non-Native Speaker
NS	Native Speaker
OPIc	Oral Proficiency Interview – Computer
PAM	Perceptual Assimilation Model
PCA	Principal Components Analysis
PCM	(Rasch) Partial Credit Model
PTE	Pearson Test of English
SA	Self-Assessment
SAT	Skill Acquisition Theory
SD	Standard Deviation
SLA	Second Language Acquisition
SLM	Speech Learning Model
TA	Teaching Assistant
TIF	Test Information Function
TOEFL	Test of English as a Foreign Language
TOPIK	Test of Proficiency in Korean

INTRODUCTION

Second language (L2) pronunciation is a critical factor in the communicative success of L2 speakers. Without intelligible pronunciation, listeners experience greater difficulty (Lee, 2017a) and may fail to fully understand speakers (Kang, Thomson, & Moran, 2018), with communication breakdowns likely to occur (Jenkins, 2002; Loewen & Isbell, 2017; Matsumoto, 2011). Even when a speaker's pronunciation is largely intelligible, poor pronunciation can make listening a more difficult, effortful task (Crowther et al. 2015; Kang, Rubin, & Pickering, 2010; Saito, Trofimovich, & Isaacs, 2017). Further compounding the gravity of unintelligibility problems is the fact that pronunciation development presents a considerable challenge to language learners. For one, out of all aspects of second language competence, pronunciation appears to be affected most by age-related effects (Long, 2013). Simply put, it is extremely unlikely for learners with a post-puberty age of onset to acquire native-like pronunciation. And although a learner's other languages can be an asset in learning some aspects of a new second language, already known languages are a strong influence on L2 pronunciation and can be a source of confusion when learning new L2 speech sounds (Best & Tyler, 2007; Flege, 1995). Beyond an initial period of rapid familiarization with the phonological system of a new L2, some researchers have argued that subsequent pronunciation development is limited and/or unlikely to occur as a product of continued, naturalistic language use (Derwing & Munro, 2015).

Fortunately, L2 pronunciation is amenable to instruction (Lee, Jang, & Plonsky, 2015; Saito, 2012; Pennington, 1998; Thomson & Derwing, 2015), and native-like pronunciation is not necessary for an L2 speaker to be broadly intelligible and highly comprehensible (Derwing, Munro, & Wiebe, 1998; Jenkins, 2000; Munro & Derwing, 1995; Levis, 2005). Lee et al.'s (2015) meta-analysis of L2 pronunciation instruction studies found beneficial effects to be

comparable in magnitude to instructional treatments targeting other aspects of L2s, such as vocabulary and grammar. However, compared to vocabulary and grammar, pronunciation often receives little attention in L2 classrooms (Foote, Holtby, & Derwing, 2011) or language textbooks (Derwing, Diepenbroek, & Foote, 2012). Language teachers have reported low levels of confidence in teaching pronunciation (Derwing & Munro, 2015), owing to a lack of background knowledge in phonology and pronunciation teaching methods (Murphy, 2014). When it does occur in language classrooms, pronunciation instruction commonly takes a one-size-fits-all approach, where a group of learners are instructed on several features selected based on the intuitions of a teacher, researcher, or materials designer (e.g., Isbell, Park, & Lee, 2019).

While both testing (Lado, 1961) and L2 pronunciation experts (Derwing & Munro, 2015) have offered many helpful suggestions for assessing individuals' pronunciation difficulties, to my knowledge there are very few well-documented and researched assessment instruments or accounts of language teacher practices used to inform whole-class or individualized instruction. Some teacher-oriented books and classroom texts for L2 English pronunciation do present some helpful methods for assessing specific problems with *perceiving* phonological features, but their approach to assessing production involves reading aloud paragraph-length written text and free production (e.g., Celce-Murcia, Brinton, Goodwin, & Griner, 2010; Gilbert, 2005). The former approach is potentially problematic due to reading aloud being a specialized skill that differs from typical speech, requiring strong literacy and sound-symbol correspondence knowledge (Levis & Barriuso, 2012), and the latter approach is limited in the sense that there is no guarantee that features targeted for assessment will be used, or used enough times to obtain reliable information about. Two recent volumes on L2 pronunciation assessment (Isaacs & Trofimovich, 2017; Kang & Ginther, 2017) have done little to address this gap of identifying individuals'

pronunciation weaknesses, and virtually no attention is given to assessing learners' pronunciation in a way that informs instruction.

One potential avenue for helping teachers and learners make more informed and confident instructional decisions about pronunciation is Diagnostic Language Assessment (DLA) (Alderson, 2005; Alderson, Brunfaut, & Harding, 2014; Lee, 2015). Situated in a larger movement calling for assessment practices that directly support learning in language assessment (Turner & Purpura, 2015) and educational assessment more broadly (Pellegrino, DiBello, & Goldman, 2016), proponents of DLA emphasize providing detailed, instructionally-useful information on what a learner can and cannot do through well-constructed diagnostic instruments and carefully thought-out procedures. With detailed knowledge of what learners know and do not know, teachers or learners using DLA decide what to study and how to go about studying it.

This dissertation explores the potential of DLA to usefully inform intelligibility-focused L2 pronunciation learning. In the following chapters, I describe a project that spans the development, field testing, and validation of an instrument to diagnose L2 pronunciation called the Korean Pronunciation Diagnostic (KPD). In Chapter 1, I review literature on DLA and validity in language testing to establish guiding principles for diagnosing pronunciation and a framework for examining the validity of the diagnostic process. In Chapter 2, I make the case for a pronunciation diagnostic and summarize theory and research on L2 pronunciation that form the grounds for the design of the KPD. Chapter 2 culminates with a prospective validity argument for the KPD which guided the validation research I carried out. Chapter 3 features, in detail, the design of the KPD and chronicles its development through two rounds of pilot testing. Chapter 4 describes the methodology of this study, detailing instruments used other than the KPD and providing an overview of procedures for the validation research reported on in Chapters 5

through 8. In Chapter 5, I present the results of measurement analyses of the KPD based on a sample of 198 L2 Korean test-takers. In Chapter 6, I describe learner pronunciation profiles that emerged from a cluster analysis of KPD phoneme-level scores. In Chapter 7, I present the results of analyses that compare KPD scores to three external measures: a measure of overall Korean oral proficiency, learner segmental phonological errors in spontaneous speech, and learner self-assessments of Korean pronunciation abilities. In Chapter 8, I draw on interviews with 21 Korean learners and a teacher who taught two of those learners to explore how these key stakeholders interpret and apply KPD results. I also report on exploratory analyses for a subset of 14 learners who took the KPD again after 2-4 months of time in which they had an opportunity to engage in pronunciation learning activity. In Chapter 9, I review the results of the previous four chapters holistically through an explication and critical review of the KPD's validity argument. Finally, in Chapter 10, I close the dissertation with a discussion of implications of the KPD's development and validation followed by discussion of broader implications for diagnosing L2 pronunciation and diagnostic language assessment.

CHAPTER 1: DIAGNOSTIC LANGUAGE ASSESSMENT

In this chapter, I provide a broad overview of Diagnostic Language Assessment (DLA). I begin by defining DLA and situating it in relation to other types of assessments. Here I include examples of several DLA instruments. Next, I raise and discuss key concerns in DLA theory and practice that are of particular relevance to this dissertation. Finally, I discuss argument-based validity as means of (a) establishing support for using tests and (b) setting a validation research agenda for DLA.

What is Diagnostic Language Assessment?

Diagnostic language assessment (DLA) has the aim of uncovering a language learner's strengths and weaknesses for the purposes of informing instruction (Alderson et al., 2014). In this sense, DLA can be considered as a type of formative assessment, which is concerned with student progress toward achieving the goals or target outcomes of an educational curriculum. Indeed, for many decades now, teachers of languages and other subjects have been using formative assessments to inform the teaching of their courses and to help individual students, often through the provision of individualized feedback. One way that DLA can be distinguished from other types of formative assessment is its scope. Whereas many formative assessments are used to gauge student progress toward completing an in-progress task or achieving a near-term curricular outcome, DLA is concerned with a learner's overall level of ability and finding their weakest links in the use of that ability. Further, DLA also has an orientation to the future: DLA should yield information that is useful for subsequent instruction.

Before proceeding further, it is important to clarify the use of the term *diagnostic* in reference to DLA and other types of assessments. Unlike diagnostic tools used by psychologists, speech language pathologists, child development experts, and medical professionals, the

diagnosis yielded via DLA is not clinical in nature nor related to foundational cognitive or educational development. Having weaknesses in L2 skills and knowledge is, by and large, not pathological, and most adult L2 learners have successfully and fully acquired one language (and, often, literacy in that language) already. However, almost every L2 learner at some point experiences having some weaknesses or gaps in their L2 knowledge or skills that hamper the effective use of the L2, and the *treatments* that may be *prescribed* as a result of a DLA are simply commonly-used (but principled) teaching and learning activities. That being said, DLA theory does draw on other forms of diagnosis (Alderson et al., 2014; Alderson et al. 2015) and, I argue, may draw more directly on certain types of language-related diagnostic techniques and instruments used in other fields such as speech language pathology and educational psychology.

DLA has a clear emphasis on identifying the weaknesses of L2 learners (Alderson et al., 2014), as there is an obvious connection between these weaknesses and instructional planning. Still, determining learner strengths is not without some instructional utility: Instruction targeting mastered knowledge or proficient subskills can be confidently skipped over in favor of focusing on more pressing aspects of language competence. The reasons for using specialized assessment procedures to examine learner strengths and weaknesses are not new. Consider Lado (1961) on the challenges language teachers face in assessing their students' L2 pronunciation weaknesses:

> Informal contact with students, even the extended contact of the language classroom, is not very effective as a way to test a student's pronunciation. From this extended contact one can say that one student has better pronunciation than another in rough terms, but when asked to list the specific pronunciation problems of a particular student of ours we will remember only the very salient mispronunciations and will not as a rule be able to come anywhere near completeness. (Lado, 1961, p. 80)

The challenges Lado outlined over 50 years ago are still relevant in language classrooms and other instructional contexts today. As a means of addressing these challenges, Alderson, Brunfaut, and Harding (2014) argued for five guiding principles of DLA that, if followed, can

allow practitioners to bridge the gap between rough comparisons of ability and specific understanding of individual weaknesses (paraphrased):

1. A test user ultimately diagnoses, not the test.
2. Diagnostic instruments should be targeted and discrete and provide highly-detailed information about a learner's abilities.
3. Diagnostics should take account of multiple perspectives, including learner self-assessments.
4. DLA should involve four stages: observation, initial (informal) assessment, use of diagnostic instruments, and decision making.
5. DLA should be connected to future instruction.

Principle 1 highlights the role of the diagnostician, typically a teacher, and the role of their expertise in interpreting diagnostic information (Edelenbos & Kubarek-German, 2004) and agency in decision making. Lee (2015) added strong arguments for providing elaborate feedback (see Principle 2) and connecting results to future instruction (see Principle 5): These can be seen as essential and distinguishing components of DLA. Without these components, diagnostics are (a) unlikely to be very helpful and (b) essentially do nothing that other types of tests (achievement tests, proficiency tests) already do. In the field of L2 pronunciation, Trofimovich, Isaacs, Kennedy, Saito, and Crowther (2016) provided support for Principle 3. They found that learners frequently have poor self-assessments: Lower-ability learners overestimate their pronunciation quality, while higher-ability learners underestimate it. Trofimovich et al. suggested that using self-assessments alongside objective measures could help develop learner awareness and clarify goals for improvement. While these major principles provided important guidance for diagnosticians, previous work on DLA has elaborated in greater detail how

diagnostic tests might be designed. The following suggestions from Alderson (2005) further informed specifications for diagnostic tests:

- Diagnostic tests are based on a detailed theory of language development.
- Diagnostic tests are likely to be discrete and focused on specific elements rather than global language abilities.
- Diagnostic tests are likely to focus on lower-level (i.e., bottom-up) language skills rather than higher-order integrated skills.

These three suggestions from Alderson lay a type of foundational blueprint for designing new diagnostic instruments: a starting point, in a sense. The suggestions also provide a way to identify already-existing diagnostic tests, which may be important because not all diagnostic tests are labeled as such.

Operational Examples of DLA

While DLA has been theorized to a considerable degree, Alderson et al. (2015) noted that few specifically-tailored diagnostic language tests exist. More commonly, existing proficiency tests have been retrofitted for diagnostic purposes (e.g., Lee & Sawaki, 2009; Jang, 2009). Jang (2009) is arguably the quintessential example of this approach to diagnosis and worthy of additional consideration here due to its topical relevance and Jang's rigor of analysis and frankness in interpretation of her findings. In her paper, Jang (2009) describes application of a measurement technique called cognitive diagnostic assessment (CDA) to the reading section of *LanguEdge*, an early prototype of the TOEFL iBT. Through a rigorous, iterative analysis of *LanguEdge* items by judgments of a team of experts, Jang identified 9 subskills of reading comprehension that were tapped into by the various test items. Through the application of a sophisticated measurement model, Jang was able to estimate test-taker mastery of these 9

subskills and provide score reports with considerably more information on test-takers' reading abilities compared to just having a single reading ability score. Jang also collected data on test-taker self-assessments and conducted classroom case studies where she interviewed learners and teachers. While the CDA approach showed promise, Jang identified several obstacles to meaningfully diagnosing learners, such as some subskills being represented by too few items, very large (mostly > .8) correlations among subskills (questioning their separability), difficulty measuring very low and very high ability test-takers, issues with subskill labeling and divisibility of subskills across items, and questionable applicability of subskill feedback to instruction (though the awareness-raising capacity of the subskill feedback was noted positively by teachers). Jang connected most of these difficulties to the design of the test: *LanguEdge* was built as a proficiency test, not a diagnostic test.

Aside from retrofitting proficiency tests to provide more detailed feedback, other tests labeled diagnostic have ended up measuring language abilities broadly (i.e., primarily function as proficiency tests) and/or been mostly used for course placement decisions (e.g., DIALANG: Alderson & Huhta, 2005; DELNA: Elder & von Randow, 2008; Knoch & Elder, 2016). With the KPD, and this dissertation, I aim to put contemporary DLA theory into practice, following the principles and suggestions offered by Alderson and others.

Similar efforts to put DLA principles into practice have recently been made by Kremmel (2017). Kremmel developed an instrument that diagnoses learner levels of (written) form-meaning vocabulary knowledge, information that is useful for understanding difficulties in L2 reading comprehension. Links between diagnosis and vocabulary instruction are readily available thanks to information provided by corpus-based word frequency (Kremmel, 2016) and analyses of lexical coverage of texts at various levels of sophistication. For example, learners can

be given tailored lists or spaced-repetition flashcard programs to study with independently or referred to reading materials at an appropriate lexical level to foster incidental form-meaning knowledge acquisition. Kremmel's work in this area built more formally on the longstanding use of vocabulary size tests (e.g., Nation & Beglar, 2007) to diagnose learner weaknesses in overall receptive vocabulary knowledge and to direct students to appropriate material in extensive reading programs to promote reading and vocabulary development (Nation, 2001).

Another outstanding example of DLA focuses on the same language and skill area as this dissertation: L2 Korean pronunciation (Kim, 2006). Kim developed an instrument used for diagnosing Korean learners' pronunciation difficulties and tracking their development over time. Kim's diagnostic included a broad range of pronunciation phenomena, going beyond individual sounds to include learner knowledge of phonological processes (e.g., nasalization, tensification, consonant cluster simplification) and suprasegmental aspects of pronunciation. Kim, in many ways ahead of the curve in DLA, also described a cyclical process of feedback, observation, and reevaluation that occurred after the administration of her diagnostic. This diagnostic was later included in a two-volume pronunciation textbook (Choi, Kim, Park, Jin, & Park, 2009a, 2009b) and the scoring and feedback form noted relevant textbook units for different categories of pronunciation features. Despite the many strengths and innovations of Kim's approach, all diagnostic test items consisted of word and sentence read-alouds and did not consider learner perception when diagnosing difficulties.

The *Criterion* software published by the Educational Testing Service (https://www.ets.org/criterion) is another example of a diagnostic test. *Criterion* is a program designed to help learners improve their writing for the Test of English as a Foreign Language (TOEFL). Learners write TOEFL-style essays, which are then given an estimated (computer-

generated) overall score, but more importantly, are also given detailed, computer-generated written corrective feedback that diagnoses their writing difficulties (Chapelle, Cotos, & Lee, 2015). Learners can see what their most common errors are and are given algorithm-based advice on how to address them. Teachers can also supplement the Criterion feedback received by students.

Finally, although not labeled as DLA, Dynamic Assessment (Poehner & Lantolf, 2013; Teo, 2012) shares many of the same aims and is worth considering from the perspective of DLA. In Dynamic Assessment, learner difficulties are probed via standard test tasks (e.g., reading comprehension multiple-choice questions, oral interviews). Where learners make mistakes, mediation is provided in the form of hints or other support which allow the learner to eventually arrive at a correct answer (in the case of a discrete-point test) or otherwise improve their understanding or performance. Compared to DLA, however, Dynamic Assessment is not oriented to subsequent instruction in quite the same way. There is some overlap: Dynamic Assessment collects information on what a learner can and cannot do independently, information that can be applied to curricular placement or instructional decisions. However, Dynamic Assessment also emphasizes the mediation that occurs in the assessment event as instruction, effectively melding assessment, teaching, and learning (Poehner & Lantolf, 2013).

The diagnosis of L2 abilities for instructional purposes has not been strictly confined to the field of language assessment. Quite expectedly, L2 researchers and practitioners concerned with language teaching and learning have developed tools and instructional programs to identify and address individual learner needs, and this work has often been carried out without reference (originally, at least) to the work and theory of Alderson, Y. Lee, or other key figures in DLA. Specific to L2 pronunciation, several computer programs have recently been created that, overtly

or covertly, diagnose learner difficulties with phoneme perception and/or production and then adjust the content of program learning activities. The *NetProfII* program (https://netprof.ll.mit.edu/netprof/), developed by MIT's Lincoln Labs for the U.S. Defense Language Institute, features vocabulary and pronunciation training that provides evaluation and feedback through automated speech recognition. This program also maintains detailed records of learner performance over extended use of the program, yielding detailed reports on phoneme accuracy ratings. Although no initial diagnostic test is available, over time learners' difficulties with phoneme production are profiled and made available to the learner through an interactive dashboard; in theory a learner could then select words containing difficult phonemes to focus on in subsequent practice sessions.

Focusing on perception instead of production, Qian, Chukharev-Hudilainen, and Levis (2018) developed a program for English-language learners that provided adaptive high-variability phonetic training (HVPT) for English segments. They developed this program in response to several previous calls for greater personalization and efficiency in computer-based HVPT programming (Levis, 2007; Munro, Derwing, & Thomson, 2015). The program required learners to accurately discriminate phoneme contrasts in minimal pairs. For each training session, a phoneme was targeted five times, and if a learner met or exceeded 80% accuracy (i.e., 4+ out of 5 correct), the learner would 'exit' further training on that phoneme and instead focus on more subjectively-difficult contrasts (i.e., those responded to correctly less than 80% of trials). From a DLA perspective, Qian et al.'s (2018) program smoothly integrated diagnosis and instructional planning.

Some Key Concerns in DLA

In this section, I review some key questions in DLA that are especially relevant to the present study. In some cases, these questions reflect a lack of research or development in practice, and in others the questions reflect areas of controversy.

Practicality

Practicality in assessment is an ever-present concern: Any usefulness an assessment has can be made irrelevant by untenable time, money, or expertise requirements. That said, the greater the assessment stakes, the greater the resources are that are deemed reasonable. In medicine, where the stakes are extremely high, it is completely reasonable to run many expensive laboratory tests (themselves originally researched and developed at great cost) in order to understand problems underlying painful symptoms and potentially resulting in a life being saved. For learning disabilities, screening all youngsters early and referring probable cases for more detailed, time-intensive diagnosis and treatment by a trained expert can have major positive impacts on a child's education and long-term quality of life. For DLA, however, the stakes are generally low. Not to say that that the benefits of facile language abilities are trivial, but gaps in adult L2 ability are (a) not usually a matter of life-and-death, (b) may represent relatively minor inconveniences able to be overcome through communication strategies and/or sympathetic interlocutors, or (c) may eventually be ameliorated without specific intervention, given enough time, L2 exposure, and/or conventional instruction.

Thus, the kind of extremely rigorous scientific analyses and/or technically-savvy tools and procedures available thanks to laboratory phonology/acoustic phonetics (e.g., spectrogram analysis, ultrasound), cognitive science (e.g., event-related potentials), psycholinguistics (e.g., eye-tracking, reaction time analyses), and computational linguistics (e.g., automated speech

recognition, natural language processing) are rarely practical for real-world DLA due to reluctance to expend money and expert labor on the development and scoring of low-stakes assessments. Similarly, tests of considerable length (e.g., the 3+ hours and two visits necessary to complete an IELTS exam) and rigorous scoring procedures (e.g., multiple, trained human raters and computer scoring engine on TOEFL productive tasks) are also likely to be out of acceptable practicality bounds. While Alderson et al. (2014) did not specifically say that DLA must be brief, at the very least it should be practical for teachers and learners to do and sensitive to the many time demands on language teaching and learning. Developers of diagnostic instruments and procedures should ask: What can be provided that is practical for learners and teachers? How can one maximize, or at least strike a reasonable balance, between technical quality and resource expenditures?

Grain Size and Score Reporting: How Detailed Should Diagnosis Be?

The level of detail in information provided by test scores is a key, if not defining, feature of DLA. Clearly, a single score describing ability in a language skill area is insufficient; such a score may only be appropriate for describing global levels of proficiency. However, there is no clear guidance on what level of granularity in scores is necessary to meet the needs of DLA, i.e., identifying specific strengths and weaknesses at a level useful for instruction. The provision of a handful of subscale scores associated may or may not be sufficient for DLA. Many language proficiency exams, for example, provide subscores for each of the traditional four macroskills of reading, writing, listening, and speaking (e.g., the TOEFL), but this level of detail is unlikely to uncover anything but broad-stroke areas of strengths and weaknesses. Even presenting a handful of subscores in the context of assessing a more delimited area of language ability, such as reading ability (e.g., see the 9 subcomponents of reading ability in Jang, 2009), may not be

sufficiently informative for understanding specific student weaknesses, nor for planning instruction.

This question about how fine-grained diagnostic language assessments should be cannot be addressed entirely by the quantity of information reported. The Pearson Test of English (PTE, https://pearsonpte.com/), a standardized test of English proficiency, provides highly-detailed score reports that feature an overall scale score, macroskill subscores, and subscores for six enabling skills (grammar, oral fluency, pronunciation, spelling, vocabulary, written discourse) that underlie performance in the skill areas (i.e., a total of 11 scores, Pearson, 2018). Pearson (2018) avoid the word diagnostic, yet describe the enabling scores as "information about particular strengths and weaknesses of a test taker's ability to communicate in speaking or writing" which "may be useful to determine the type of further English study" a learner should engage in to improve (p. 42). I would hazard to say that most experts would not describe the PTE as a particularly useful diagnostic instrument, a single piece of information about, say, a learner's grammar provides little specific guidance on how or what to study. However, it may be appropriate to say that PTE scores have some diagnostic qualities. Thus, in part, the question of grain size must consider quality. Scores/subscores provided by DLA tools will likely be large in number, but must also contain diagnostically-actionable information, based on a detailed description of language ability and understanding of language development.

Although large grain-size in DLA is clearly undesirable, there may be limits on information granularity due to practicality and utilization issues. Extremely high-granularity may require untenably long observation procedures or instrument designs, making use of such techniques impractical, and stability of diagnostic classifications is likely to be lower at finer grain-size (Lee & Sawaki, 2009). Making use of information from a high-granularity diagnostic

procedure may also prove challenging or otherwise overwhelming. If the grain-size of a diagnostic is keyed to the minutest details of learning theory and language, the resulting information may be too technical and/or too voluminous for learners and teachers to fruitfully apply. Imagine being a language learner (or a teacher) and being told that you (or your student) have voice onset times for word-initial stop consonants that are on average 34.11ms too long. Without substantial training in phonology and phonetics of the target language, it might be difficult to comprehend what that information means, much less apply it. Now imagine receiving parallel information for other acoustic qualities, such as intensity, for other syllable/word contexts, and other types of sounds. Background in phonology and phonetics aside, the sheer volume of such information would likely be overwhelming, perhaps debilitatingly so. Thus, grain-size is a Goldilocks issue for DLA practitioners: Not too large, not too small, not too vague, and not too technical: *just right* should be strived for.

In DLA, score reporting has been framed in terms of providing feedback (Alderson, 2005; Kunnan & Jang, 2010) rather than simply informing a stakeholder of a test result. This is one more way in which DLA emphasizes a connection to subsequent learning: Just like immediate corrective feedback in a classroom interaction (e.g., Saito & Lyster, 2012) or delayed feedback on written assignments (e.g., Ferris, 2010), I argue that the primary purpose of feedback to learners from a diagnostic test is to raise awareness of linguistic form in order for the learner to subsequently apply conscious attention to form in both deliberate learning activity and general language use. Theoretically, this view of diagnostic feedback is well-aligned with SLA hypotheses and theories that suggest that learners need, or at least can benefit from, conscious attention to forms (i.e., vocabulary and/or grammar) to develop and ultimately acquire or otherwise master those forms (e.g., DeKeyser, 2017; Robinson, 1995; Schmidt, 1990, 1993;

Schmidt & Frota, 1986), but perhaps not SLA theory that suggests that all learners need is implicit (i.e., unaware) learning of form in order to develop and acquire the forms (e.g., Krashen, 1982; Truscott, 1996; VanPatten & Rothman, 2015). In Schmidt's (Schmidt, 1990, 1993; Schmidt & Frota, 1986) influential Noticing Hypothesis, it is claimed that learners must be aware of linguistic forms at a conscious level (i.e., they must *notice* forms). Noticing is what allows learners to direct attention to a form, which in turn promotes storage in memory and learning. Complementing this hypothesis, Robinson (1995) detailed the process by which noticing and attention to form in short-term memory is a necessary condition for storage in long-term memory. Such a process also factors into Gass and Mackey's (2006) Interaction Hypothesis, where through input from interlocutors and interactional feedback learners' awareness of and attention to form is promoted, facilitating acquisition. Coming from a slightly different perspective, a Skill Acquisition Theory (SAT) approach to SLA (DeKeyser, 2017) suggests that a considerable amount of practice with attention given to linguistic forms is necessary to achieve fluent, automatized skill in using them. This practice can come in the form of pre-planned instruction (e.g., a classroom activity) or learners' own conscious monitoring of explicit knowledge (or *declarative* knowledge in DeKeyser's framework) during authentic language use (e.g., daily interactions during study abroad).

More specific to the present study, learner awareness and attention to form is known to be helpful to L2 speech learning (Guion & Pedersen, 2007; Kennedy & Trofimovich, 2010; Moyer, 2014; Saito, 2018; Thomson, 2012). It is widely accepted that explicit phonetic instruction (e.g., pronunciation instruction based on explicit description of articulation) is beneficial, with learners generally showing improvements on the phonological forms they are taught (Lee et al., 2014). However, all phonological forms cannot be taught or paid attention to all the time. This is where

learner autonomy and independent use of learning strategies (or metacognitive strategies) (Moyer, 2014) also comes into play when considering the utility of diagnostic feedback: An experienced, well-trained, strategic learner who is aware of their weaknesses may be able to deliberately pursue study activities or utilize techniques that address their specific needs. In Moyer's (2014) review of highly successful L2 phonology acquirers, she specifically points out the autonomous deployment of strategies such as "self-monitoring", "explicit attention to accent", and "conscious concern for accent" (p. 430), which all draw on learner awareness and attention to form. Along these lines, Kunnan and Jang (2010) suggested that for diagnostic feedback to be most useful, it should be presented in a way that encourages learners to "reset their own learning goals by breaking down goals into manageable tasks" (p. 617). In other words, by guiding learners to linguistic forms in most need of attention, diagnostic feedback potentially enhances the learner's efficacy in autonomous learning and strategy deployment.

It is also worthwhile to consider DLA score reporting from the perspective of teachers or tutors. Although a teacher's conscious attention to linguistic form is not a primary concern, a teacher's awareness of student weaknesses can be deployed to induce or reinforce learner awareness through well-matched pedagogical responses, such as *in situ* corrective feedback or deliberate provision of pronunciation learning opportunities, such as the creation (or modification) of classroom activities or the selection of learning materials, drawing on the teacher's training and knowledge of pronunciation teaching (Baker, 2014; Burri, Baker, & Chen, 2017). As Alderson et al (2015) pointed out, the whole DLA process is for naught if no party, learner or teacher, appropriately considers and then acts upon diagnostic feedback – a sentiment that would surely be agreed upon by scholars supportive (e.g., Ferris, 2010; Saito & Lyster, 2010) and critical (e.g., Truscott, 1996) of feedback in SLA and language teaching.

Some reports of stakeholder understanding of diagnostic results have been cause for considerable concern. Huhta (2010) found that language learners paid most attention to the overall proficiency level and ignored many other parts of DIALANG results. Similarly, Yang (2003) found that DIALANG test-takers compared their overall scores to their TOEFL or IELTS scores and did not substantially engage with the diagnostic information provided. Jang and Wagner (2014) emphasized that learners with different goals and motivations are likely to differ in their uptake and application of diagnostic feedback. A key question, then, is: How might DLA score reports be designed to effectively promote awareness of linguistic forms (or perhaps other relevant aspects of performance)?

There does not currently appear to be a simple answer to this question. For one, the feedback of different types of diagnostic assessments will often, and perhaps necessarily, take different forms: Diagnostic feedback on L2 writing may involve annotation of learner text (e.g., from a teacher or a computer program), while diagnostic feedback on a reading test could utilize item-level hints during the test, and item-level feedback after the test. Despite these skill/content area and method considerations, there are few, if any, specific guidelines for presenting diagnostic information. Some useful advice, though vague in terms of format, comes from Alderson and colleagues (2015), who suggested that diagnostic feedback could (and perhaps should) attempt to link together weaknesses, probable causes, and next steps for learning. They also offered the following key characteristics of diagnostic feedback (p. 169):

- it is much more detailed than, for example, a reading test score;
- it is not limited to the actual errors a learner makes;
- it is based on an understanding of what probably underlies those errors; and finally,

- it is not limited to errors but also addresses what the learner could do to improve the skill involved

When considering what the literature says about provision of diagnostic feedback, I wonder whether, for a test designed for diagnostic purposes from the ground-up, if any kind of total score is necessary. Although it is common to provide a total score (Alderson, 2005; Jang, 2009; Lee & Sawaki, 2009; Sawaki, Kim, & Gentile, 2009), I question whether the practice was simply born out of habit or just a byproduct of retrofitting proficiency tests for diagnostic purposes. This is not to say that providing an overall ability score is wholly inappropriate, but excluding any overall scores and presenting only detailed information on specific aspects of ability and suggestions for improvement could avoid the problem of learners finding a total score and ceasing further engagement with feedback. Indeed, effective feedback in classrooms is not contingent on a teacher telling a learner overall how good they are before getting into the specifics of an error or recurring difficulties.

Measurement Models and Techniques

A measurement model can be simply defined as the way scores are assigned to objects. In the case of language learning, the *objects* of measurement are typically L2 learners, and these learners are assigned scores on some attribute (a skill, a domain of linguistic knowledge) as the result of an assessment procedure (e.g., an interviewer's overall judgment of a learner's proficiency level, a conventional reading proficiency test). In several treatments of DLA, measurement is little discussed (e.g., Alderson, 2005; Alderson et al., 2015) while in others, measurement techniques are on center stage (Jang, 2005, 2009; Lee & Sawaki, 2005).

Perhaps the issue of measurement is sometimes avoided due to the potential thorniness of dimensionality in DLA. In measurement, dimensionality refers to the number of dimensions

along which examinees are meaningfully compared on the basis of an assessment procedure. Most commonly, and especially in high-stakes educational and language proficiency testing, measurement is unidimensional: Learners' are assigned a single score that refers to their ability along a single dimension. Unidimensional measurement is supported by well-tested and widely-used techniques and analysis software familiar to many assessment practitioners. However, as previously discussed, DLA requires more than a single score in order to be truly diagnostically useful. Ideally, DLA yields multiple scores that allow for meaningful inferences on the status of subcomponents and more narrowly defined knowledge bases that influence macro abilities. Rigorous and simultaneous measurement of multiple dimensions presents a marked increase in theoretical and technical complexity and is unfamiliar territory for many language testing and assessment specialists. DLA, which will typically report multiple scores targeting different aspects of an ability, may on the surface appear to be a prime example of a multidimensional measurement opportunity.

It is important to note that a measurement dimension is not the same as a construct or attribute. Rather, measurement dimensions are mathematical/statistical abstractions of assessment data; the relationship between a measurement dimension and a theoretical construct must be inferred and supported by additional evidence (Reckase, 2009), such as the test content or investigation of item response processes. Because human knowledge structures and mental abilities are complex, it is possible to conceive of theoretical constructs as abstractions of complicated, multicomponent mental processes. Language ability is no exception (e.g., Bachman & Palmer, 2010). For this reason, unidimensional measurement has occasionally been criticized as fundamentally flawed (Buck & Tatsuoka, 1998). However, a higher-level abstraction like *reading comprehension*, which obviously involves identifiable subcomponents such as

lexicogrammatical knowledge and grapheme decoding, can be justifiably measured along a single dimension, focusing on global performance rather than attempting to directly and separately measure each relevant knowledge base and processing routine. Nonetheless, there is still a need in DLA to get information on those knowledge bases and processes.

There are at least three approaches to acquire such information in DLA: arithmetic subscore calculation, unidimensional Item Response Theory (IRT) or Rasch measurement with analysis of unexpected responses, and multidimensional measurement. These approaches differ substantially in their practicality. Simple subscore calculation has the smallest sample size requirements (essentially there is none) and the lowest technical expertise. One simply defines which items on a test or other assessment tool constitute meaningful subscales and computes sum scores. These subscores can then be added up to arrive at a total score representing an individual's overall ability. This approach is (implicitly) in line with Classical Test Theory (CTT), which posits that a person's true ability is represented by the sum of item/task scores, plus or minus measurement error. Due to its practicality, this method may be the most common approach to gleaning information on subcomponent knowledge and skills in language assessment (Jang, 2009). Although technically simple, the definition of meaningful subscales should nonetheless be principled, based on a thorough understanding of the underlying processes and the linguistic knowledge necessary to carry out higher-level tasks. It is also possible to apply weights to items and/or subscales, usually based on theory, but also possible based on technical quality, before adding them to produce an overall ability score (e.g., weighing the pronunciation scores for phonemes according to communicative importance, weighing subscores equally).

IRT and Rasch measurement techniques are common in language testing (McNamara, 1995; Knoch & McNamara, 2012). Comparatively, Rasch and the simplest form of an IRT

model require larger sample sizes (a minimum of 50-200, depending on desired precision of estimates and test design factors, DeAyala, 2009, Linacre, 1994) and greater technical expertise compared to CTT. Unlike CTT, IRT and Rasch consider the responses of individual examinees to individual items when determining the ability of people and the difficulty of items. In simpler Rasch/IRT analyses, when the data fit the model, raw total scores will correlate almost perfectly with a person's ability measure, allowing for straightforward (but more rigorously supported) interpretations of raw scores. Importantly, estimates of person ability and item difficulty are theoretically not sample dependent in IRT/Rasch (this is practically plausible when initially estimated with a sufficiently large and representative sample), which allows for detailed and generalizable consideration of item difficulty hierarchies that can be related to the theoretical understanding of the construct being assessed. In practice, extracting diagnostic subscore information using Rasch/IRT is very similar to computing subscores in a CTT model, but with greater confidence in the order of item/task difficulty and more precise probabilistic information on items that an examinee under- or overperforms on.

Using Rasch analysis to collect validity evidence for an aural vocabulary knowledge test, McLean, Kramer, and Beglar (2015) were able to show that item difficulty patterned reliably according to frequency, with items targeting more frequent vocabulary being easier than less frequent vocabulary. This aligns with exposure-based accounts of vocabulary acquisition and empirical findings of word frequency in natural language use, which in turn allows for developmental interpretations of the vocabulary test scores. For example, McLean et al.'s (2015) vocabulary test scores can be used to infer a learner's overall level of vocabulary knowledge, and highlight areas of weakness, such as an unexpected number of incorrect responses to items in a high-frequency (easier) band of vocabulary. Such a student could be referred to some remedial

vocabulary instruction. Rasch/IRT measurement can also accommodate polytomously-scored item/task responses, including the combining of several related items into one item parcel (also referred to as an *item bundle* or *superitem*). Justice, Bowles, and Skibbe (2006) used Rasch analysis on data from a developmental pre-literacy test that featured dichotomous and polytomous items; items were constructed to target specific knowledge facets of basic print concepts. A product of this analysis was an easy-to-use scoring sheet which visually incorporated relations between item difficulty and test-taker ability, allowing test-users to intuitively understand which item scores a child would be expected to receive given their overall ability level (based on their raw total score). Thus, broad instructional decisions could be made based on an overall score (e.g., referral to remedial pre-literacy instruction) and more specific instructional decisions could be made based on item performance (e.g., reviewing where the title of a book can be found).

Multidimensional measurement is the third and most demanding approach. There are a variety of multidimensional measurement techniques, most based on IRT (Reckase, 2009), which could be applied meaningfully in DLA. For the sake of brevity and relevance, this review will focus on Cognitive Diagnostic Models (CDM), a family of multidimensional IRT models that include variants such as the Rule Space Model (e.g., Buck & Tatsuoka, 1998) and the Fusion Model (e.g., Jang, 2009). Expanding on simpler IRT models, CDMs introduce additional dimensions based on cognitive attributes (skills, processes, knowledge) needed to successfully respond to items. Item attributes, usually coded by several experts with thorough understanding of knowledge bases and cognitive processes tapped by the larger construct being measured, can explain examinee responses to items in finer-grain detail: Examinee mastery of specified

cognitive attributes determine their odds of correct responses in accordance with the demands of each item.

CDM may represent an ideal measurement model for DLA. The technical potential of CDMs to provide detailed, precisely-measured information on a range of subordinate skills and knowledge aligns well with the goals of DLA. Criticism of CDM in DLA primarily stems from their post-hoc application to general proficiency tests, a phenomenon Alderson (2010, p. 99) described as "trying to retrofit a proficiency test into diagnostic uses." Much of Alderson's criticism of this approach stems from problems in accurately ascribing cognitive attributes to the kinds of questions found on typical reading or listening comprehension tests. For instance, even experts in L2 reading will not always agree whether a given reading question requires an inference to be made. Jang (2009), whose coders only achieved moderate agreement in assigning attributes to reading items, also recognized the limitations of the retrofitting CDM approaches, agreeing with Alderson's criticism that diagnostic tests need to be built from the ground up for diagnostic purposes in order to capitalize on the technical potential of CDMs.

Retrofitting is not the only weakness of CDMs. Being more technically sophisticated, potentially estimating numerous cognitive attributes, CDMs usually involve much larger samples than the other measurement models discussed so far. Jang's (2009) application of a CDM to the *LanguEdge* reading test (a TOEFL iBT precursor) involved 2,703 examinees; and Sawaki, Kim, and Gentile (2009) had over 3,000 examinees, while Buck and Tatsuoka (1998) used a more modest sample of 412. Even taking Buck and Tatsuoka (1998) as an acceptable sample size (for 15 cognitive attributes), it is clear that CDM analyses can be quite resource-demanding.

Self-Assessment

Self-assessment (SA) has been popular in L2 research and classroom practice and is a key step in effective DLA according to Alderson (2005; Alderson et al., 2015). Generally, the accuracy of SA (i.e, association between self-assessment and objective/expert assessment) for language learners has been found to be positive and moderate, yet widely variable: Ross' (1998) seminal meta-analysis found that the average correlation between learner SA and objective tests for overall proficiency was r = .63, with a range of .09 to .80. Especially relevant to the present dissertation, the average correlation between SA and an objective test for listening ability was r = .65 (range: .25 to .81) while the average correlation for speaking ability was slightly lower at r = .55 (range: .09 to .78). More recently, Ma and Winke (2019) found that L2 Chinese learners could fairly accurately self-assess their proficiency at broad levels but struggled to accurately assess their abilities at finer levels of distinction, especially if they were at Intermediate levels of proficiency (Novice and Advanced learners were better at self-assessing their oral skills than were Intermediate-level learners). For pronunciation self-assessment, findings pertaining to learner accuracy are mixed. Trofimovich et al. (2016) found weak to small correlations between SA and expert judgments of degree of foreign accentedness (r = .06) and comprehensibility (r = .18) for L2 English learners. Lappin-Fortin and Rye (2014) found that learners of French were reasonably accurate in self-assessing their global pronunciation and learned to more accurately assess specific features of French pronunciation that they had been taught explicitly, but nonetheless tended to overestimate their abilities. Dlaska and Krekeler's (2008) study, where learners of German assessed their segmental productions by comparing their recordings to native speaker models, found high overall learner agreement with expert judges, but the learners failed to identify roughly half of their mispronunciations – in other words, they tended to overestimate

the accuracy of their productions. Thus, it might appear that SA involving (a) productive and (b) more specific aspects of language proficiency may tend to be less accurate than other forms of SA. Interestingly, Trofimovich et al. (2016) also found that less-proficient learners (i.e., those with stronger foreign accents or lower comprehensibility) tended to overestimate their speech quality while more proficient speakers tended to underestimate. This finding reflects the well-known Dunning-Kruger effect (Kruger & Dunning, 1999), that is, the notion that those with less expertise tend to overestimate themselves while those with greater expertise tend to underestimate themselves, which throws another wrench into the machinery of self-assessment.

The apparent flaws of SA are not necessarily a problem for DLA. Rather, in DLA, they may be seen as a learning opportunity: For learners unaware of their weaknesses (or strengths, as it may be in the case of more experienced or proficient learners), reconciling SAs with expert/objective scores from diagnostic instruments can highlight gaps and create awareness in learners that will hopefully support subsequent learning. Indeed, if SA were so accurate that other steps of DLA showed learners nothing new, there would be little argument for doing anything beyond SA in the first place. In the *DIALANG* test for diagnosing foreign language ability (www.lancaster.ac.uk/researchenterprise/dialang/about.htm, Alderson, 2005), examinees complete a self-assessment prior to taking the DIALANG and then their DIALANG results are presented alongside their SA results after the test is complete. When there is a mismatch, the DIALANG system provides several possible explanations for why there is a mismatch, with the hope that learners will more carefully consider their abilities and take to heart the suggestions for future study provided with the test results. While this approach to utilizing self-assessment is somewhat simplistic, especially as the DIALANG focuses on language skills rather broadly, it

may nonetheless provide a useful wake-up call for someone presenting a strong Dunning-Kruger effect in the conceptualization of their language skills.

SA often takes the form of *can-do* statements, popularized by frameworks of language proficiency such as the Common European Framework of Reference (CEFR, Council of Europe, 2017) and the American Council on the Teaching of Foreign Languages' *Guidelines* (ACTFL, 2012). Can-do statements may be framed as yes/no questions (i.e., dichotomous responses) or involve longer rating scales (see Little, 2005; Tigchelaar, Bowles, Winke, & Gass, 2017; Ma & Winke, 2019). Otherwise, SA may employ other item types with rating scales anchored by short weak/low and strong/high descriptors of ability or performance, e.g., the accentedness and comprehensibility scales used by Trofimovich et al. (2016). In DLA, it would seem prudent for the grain-size of any SA to be roughly parallel to that of any diagnostic instrument used in the process. This is not to say that the inclusion of some broader, more general self-assessment items should be discouraged, but rather that it would seem easier and more useful for learners to compare self-assessment and diagnostic test results that are more directly relatable.

Validity in DLA

Alongside the previously discussed key questions in DLA, validity is a chief concern for any type of assessment. Validity also provides a framework for investigating and evaluating assessment instruments, procedures, and uses. In line with the larger field of educational assessment, validity in language assessment is widely conceived of in an argument-based framework (Bachman & Palmer, 2010; Kane, 2013; Chapelle, Enright, & Jamieson, 2008, 2010), and DLA is no exception (Chapelle, Cotos, & Lee, 2015). Whereas classical notions of validity have a narrow focus on whether a test measures what it claims to measure (or more precisely, whether variation in test scores reflects variation in the underlying construct(s) or trait(s))

(Borsboom, Mellenbergh, & van Heerden, 2006), the contemporary argument-based approach broadens the scope of validity to include the decisions made based on test scores and subsequent impacts on test stakeholders (Messick, 1989). Although argument-based validity theorists in educational assessment (e.g., Kane, 2013) and language assessment (Bachman & Palmer, 2010; Chapelle et al., 2008, 2010) differ somewhat in their specifications of validity arguments, the general structure involves a series of progressive inferences that lead from test-taker responses to the use of test results by a range of stakeholders (Figure 1). Each inference (indicated by curved arrows in Figure 1) requires some sort of backing (gray boxes with bullet-point examples).

The first inference in Figure 1 is evaluation. For scores to be meaningful, they must be appropriately assigned to responses elicited by well-designed test items or tasks in a way that reflects the targeted construct of language ability. Support for this inference comes from the theoretical background related to the construct and the connection between theory and operationalization in the form of test tasks and items, with well-reasoned scoring rules in place based on that connection. In a basic sense, this inference requires that the test content and tasks are a sensible snapshot of the way the targeted construct functions in real life. The next inference, generalization, reflects the assumption that scores from a given test observation are consistent with other possible observations. Support for this inference broadly involves estimation of the reliability of scores, which provides statistical backing for the notion that a test-taker is expected to receive similar scores, for example, if he took a slightly different form of the test, if he took the test two days earlier or later, or if a different teacher scored his responses.

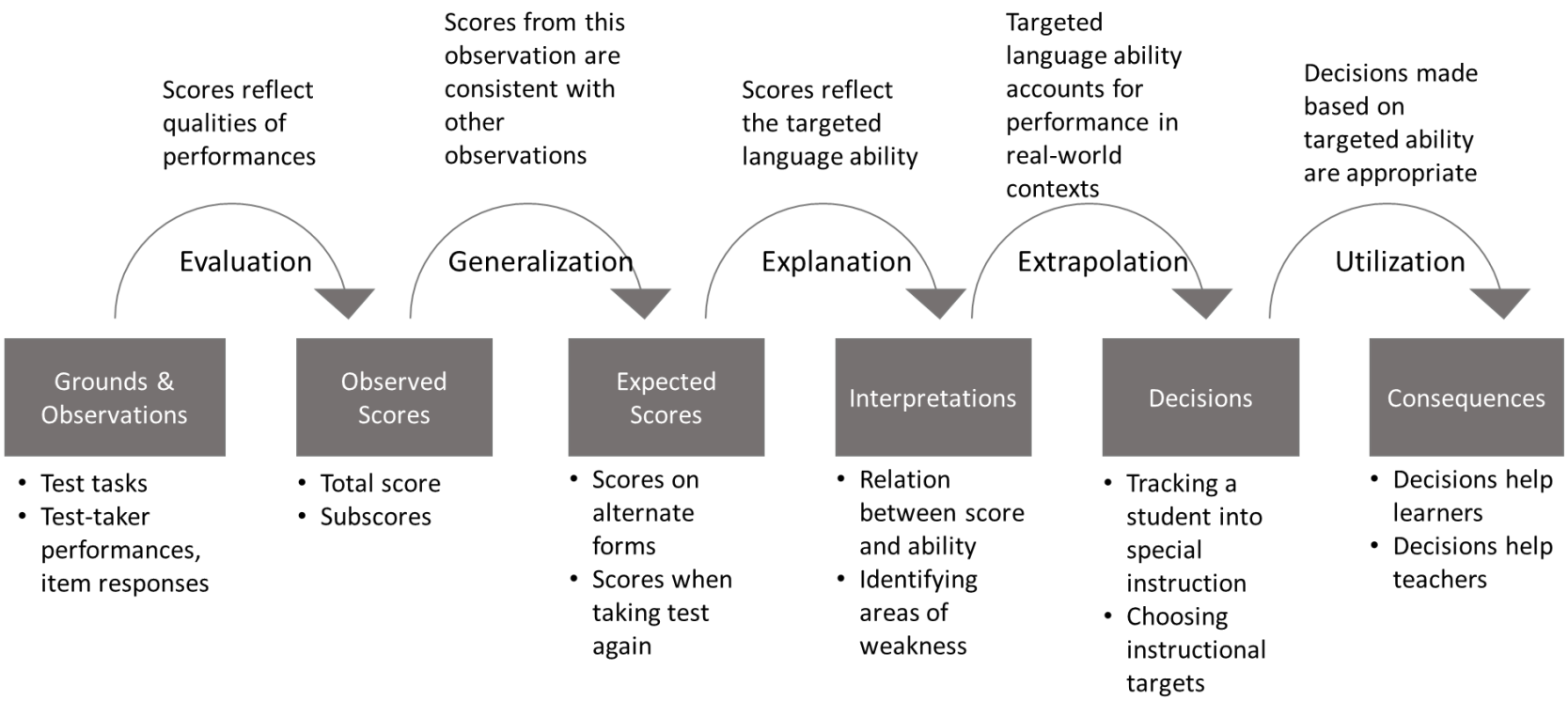


Figure 1.1. A series of inferences that typify validity arguments.

Next comes the explanation inference, which holds that differences in scores are explained by differences in the underlying language constructs. Conversely, scores are not attributable to irrelevant factors (e.g., a reading test should not depend on an examinee's mathematical ability). Support for this inference largely comes from measurement characteristics, such as the alignment between predicted and empirical item/task hierarchies and description of the internal structure of test items/tasks (e.g., dimensionality analysis, relationships among tasks). The extrapolation inference, which allows for test scores to be interpreted as reflective of performance in other, non-test situations, follows. Support for this inference often comes in the form of a relationship between test scores (or test-taker responses) and authentic (or semi-authentic) performance on a task with real-world relevance.

The last inference in the chain is utilization. Some validity theorists make a distinction between this inference and the previous inferences: Kane (2013), for example, distinguishes between 'interpretation' and 'use' of test scores; utilization falls into the latter category while the previous inferences relate more directly to the issue of interpreting the meaning of scores in terms of the targeted construct. The primary assumption in the utilization inference is that decisions made on the basis of test scores are useful, fair, and beneficial. Evidence is required that shows how decisions made help ensure or improve outcomes (learning, job performance, etc.), or otherwise beneficially serve a social function (e.g., allow a school to hire teachers with adequate language ability).

While these five inferences are at the core of most validity arguments, other inferences are possible, and often appended to either the beginning or end of this core chain. For example, the work of Carol Chapelle and colleagues (e.g., Chapelle et al., 2008, 2010; Chapelle et al., 2015) features validity arguments that begin with an additional inference (often referred to as

"domain definition," or sometimes "authenticity") related to the link between the design of test tasks and real-world language use or contexts of use. Chapelle and colleagues as well as Bachman and Palmer (2010) have also included additional inferences at the end of the chain related to the beneficial consequences of utilizing test results.

For DLA, in line with Alderson (2005), Alderson et al. (2014), and Harding et al.'s (2015) recommendations for diagnostic instruments to be constructed based on a detailed theory of learning and models of language processing, adding an inference at the beginning of the chain, connecting such theory to test-taker responses, would strengthen a DLA instrument's validity argument. Similarly, adding an inference at the end of the chain related to the beneficial impact of applying diagnostic results would be in-line with the emphasis on subsequent learning in DLA (e.g., Lee, 2015) and would enhance the persuasiveness of the validity argument. In following chapters, I will introduce a proposed validity argument for the KPD that I use to set the research agenda for this dissertation.

CHAPTER 2: DIAGNOSING SECOND LANGUAGE PRONUNCIATION

As Alderson (2005) suggested, a strong theory of language development ought to underpin any diagnostic language assessment. Theory also supports key inferences in validity arguments. In this chapter, I begin by making a case for developing a pronunciation diagnostic, highlighting a gap in instructionally-relevant pronunciation assessments. Then, I establish the theoretical grounding for the KPD by reviewing theories and research related to L2 pronunciation, both in general and specifically for Korean. Specifically, I review the linguistic, cognitive, and developmental bases that underpin the design of the KPD. I end the chapter by laying out the goals of the KPD development project and introducing a validity argument used to frame the validation research agenda in this dissertation.

Why Diagnose L2 Pronunciation?

In the Introduction of this dissertation, I pointed out that pronunciation can present persistent challenges to L2 learners, and that such problems can lead to intelligibility issues in real-world communication. I also pointed out that despite the well-documented effectiveness of pronunciation instruction, pronunciation is often neglected in language classrooms due to time/curricular restraints and in some cases lack of teacher confidence. At the same time, whole-class pronunciation instruction, when it is done, can be limited in its effectiveness, possibly due to instructional targets being sub-optimally matched to individual learner needs. Derwing and Munro (2014, p. 44) illustrated such a condition, where the resulting instructional decision was to mostly avoid teaching segmentals: “Little emphasis was placed on individual vowels and consonants, it turned out, because the students shared very few problems at the level of the segment.” Thus, *anything* that would aid teachers and learners in identifying critical targets for

pronunciation learning activity, whether in in-class or more individualized out-of-class formats, would appear beneficial.

But why would a diagnostic assessment, specifically, be beneficial? As reviewed in Chapter 1, DLA has several characteristics that make it particularly well-suited to supporting learning. Compared to other types of assessments, such as proficiency or achievement tests, diagnostic instruments are designed with learning theory in mind and provide highly-detailed feedback that can be used to inform instruction. By and large, the most common form of pronunciation assessment would appear to be as a component of large-scale, high-stakes speaking assessments (Isaacs, 2018; Isaacs & Harding, 2017). In these sorts of assessments, pronunciation is treated broadly as just one aspect of rubrics used to evaluate a learner's overall speaking abilities (e.g., IELTS, https://www.ielts.org/; OPIc, https://www.languagetesting.com/oral-proficiency-interview-by-computer-opic; TOEFL iBT https://www.ets.org/toefl), and results provide little to no guidance for subsequent learning activity. Isaacs, Trofimovich, and Foote (2018) developed a more detailed scale of global pronunciation quality that is theoretically well-grounded and could be used for upper-level instructional decisions, such as assigning international graduate students to pronunciation support classes. Similarly, for Korean, Lee (2017b) developed and examined pronunciation rating scales that can be used to augment speaking assessments. However, these scales ultimately fall short of providing individualized, instructionally-relevant information about learners' abilities.

Other pronunciation assessments have engaged more meaningfully with detailed, individualized results informative to learning. Lappin-Fortin and Rye's (2014) self-assessment approach is commendable for its detail, requiring students to think about their global

pronunciation quality as well as their ability to produce individual features of French phonology, such as vowel and consonant segments and features of connected speech. Dlaska and Krekeler (2008), working with learners of German, also took a self-assessment approach that relies on learner comparisons of self-recordings to native speaker audio models to raise learner awareness of segmental pronunciation difficulties. Tsurutani (2008) took advantage of automated speech recognition (ASR) to provide detailed feedback on learner Japanese pronunciation that is integrated with training activities. Kim (2006), discussed in the previous chapter, aimed to diagnose difficulties with individual Korean phonemes and suprasegmental features. Similarly, Celce-Murcia et al. (2010) provided a tool to diagnose L2 English speakers' production difficulties and included some tasks targeting perception as well.

However, each of these examples, while certainly of considerable utility, could be improved on. Many of them (e.g., Celce-Murcia et al., 2010, p. 481; Dlaska & Krekeler, 2008; Kim, 2006; Lappin-Fortin & Rye, 2014) evaluate pronunciation based entirely on read-aloud words or sentences (which can be prone to non-pronunciation influences, Levis & Barriuso, 2012; Munro, 2008), rely on a native-speaker standard (Dlaska & Krekeler, 2008; Tsurutani, 2008), or have limited observations of pronunciation targets (Dlaska & Krekeler, 2008; Kim, 2006). Aside from some suggestions for perception items from Celce-Murcia et al. (2010, but not included on their diagnostic) that mirror Lado (1961), none incorporate production and perception of pronunciation features, a design which has strong motivations in pronunciation learning theory (more details follow in later sections). Specific to Korean pronunciation, Lee (2017b) stated that Kim (2006) appears to be the only example of a detailed pronunciation assessment for L2 learners, and further noted that research on Korean pronunciation assessment is lacking in general. In her state-of-the-art review of pronunciation assessment, Isaacs (2018)

lamented that Lado's (1961) nearly 60-year-old text is still the most comprehensive treatment of pronunciation assessment, signaling that new advances are sorely needed. I agreed with Isaacs, especially in regards to lower-stakes, instructionally-relevant pronunciation assessments, and I saw an opportunity to fill these gaps in pronunciation assessment by developing a state-of-the-art yet practical assessment tool, in line with diagnostic principles elaborated by Alderson (2005) and colleagues (2014), that (a) diagnoses learner phoneme-level strengths and weaknesses in pronunciation, (b) integrates both production and perception, (c) explicitly promotes intelligibility-based evaluation of pronunciation, (d) does not rely exclusively on read-aloud tasks, (e) is relatively easy to administer and score, and (f) beneficially informs pronunciation learning, and evaluate it rigorously.

What is L2 Pronunciation?

Pronunciation refers to how humans produce speech using the vocal apparatus. Speech begins inside the mind, and through complex neural-motor activity, the lungs, vocal tract, and mouth move to produce sounds that represent language. The different ways in which humans use the vocal apparatus affect the resulting sounds produced in terms of both acoustic and temporal features. Pronunciation encompasses the qualities of segmental features that define words, i.e., phonemes, and suprasegmental (or prosodic) features that take shape over multiple segments, such as intonation, pitch accent, and stress. Features commonly associated with speech fluency, like speech rate and pauses, are also related to pronunciation. In naturalistic speech, these features can be difficult to tease apart, but nonetheless form a meaningful and practical basis for examining pronunciation.

L2 researchers have commonly examined pronunciation quality in terms of pronunciation's impact on a listener. Derwing and Munro (2015) offered a useful (and widely-

used) framework for considering listener-based dimensions of pronunciation. The degree to which a speaker's message is accurately received by a listener is referred to as *intelligibility*. A related but partially independent dimension is *comprehensibility*, which refers to the listener's ease of understanding a speaker. Seen another way, comprehensibility is analogous to the amount of effort a listener must put forth to comprehend speech. *Accentedness* is the difference between the speaker's pronunciation and the listener's own speech variety. When dealing with L2s, accentedness can also be understood as degree of foreign accent (rather than accents associated with L1 regional dialects). While all three dimensions are worth considering, Derwing and Munro declared that intelligibility is "the most fundamental characteristic of successful oral communication" (p. 1).

If the sounds produced by a speaker (in a L1 or L2) are not intelligible to listeners, communication will not be successful. The importance of speech intelligibility has long been recognized throughout the field of L2 pronunciation (e.g., Abercrombie, 1949), but has not always been emphasized in language teaching and assessment. Recently, the importance of intelligibility has been stressed in pedagogy by Levis (2005), who contrasted the previous emphasis in language teaching on achieving nativelike speech (the *Nativeness Principle*) with a more contemporary focus on learner intelligibility (*Intelligibility Principle*). Intelligible, rather than native-like, speech has concomitantly seen greater emphasis in descriptive frameworks of communicative second language ability, such as the Common European Framework of Reference (CEFR, Council of Europe, 2017) and the American Council on Teaching Foreign Languages' *ACTFL Guidelines* (2012).

In these proficiency frameworks, used in both pedagogical settings and assessment, intelligibility is generally depicted as something that lower-proficiency learners will struggle

with. Their interlocutors must put forth “some effort” and engage in “collaboration” with the speaker to establish meaning (Council of Europe, 2017, p. 134-135) or be “sympathetic” and/or “accustomed” (ACTFL, 2012, p. 9) to L2 speech in order for communication to be successful. At intermediate levels, learners are *generally* intelligible, but still mispronounce some sounds regularly. At higher levels of proficiency, learners are assumed to have sufficient control over the production of almost all L2 sounds (and indeed do have high accuracy in the production of the most critical sounds for distinguishing words in an L2, Kang & Moran, 2014), at which point suprasegmental and fluency-related aspects of L2 pronunciation may figure more prominently in communicative effect and ease of understanding from the listener’s perspective.

Empirical research on factors influencing speech intelligibility has suggested an integral role for segmental pronunciation. In monologic speech, research has shown that segmental accuracy has a clear effect on intelligibility (Kang, Thomson, & Moran, 2018a, 2018b). In Kang et al. (2018b), segmental features had the greatest influence on the intelligibility of individual sentences as well as on the comprehension of extended monologues. Along similar lines, Jenkins (2002) argued that most pronunciation-related breakdowns between L2 users (i.e., L2 pronunciation being processed by an L2 listener) are related to segmental features. Jenkins, focusing on English as an international language, went on to propose a pared-down, intelligibility-oriented pronunciation syllabus for L2 English learners, prioritizing consonant phonemes and deemphasizing many suprasegmental features. While context is often pointed to as a resource that interlocutors can use to help maintain intelligibility when mispronunciations occur, Jenkins found that this occurs less often when the interlocutor is a non-native speaker. Other L2 research has reinforced the importance of segmental pronunciation in interactive speech, including (but not limited to) research on English (Loewen & Isbell, 2017; Matsumoto,

2011), French (Kennedy, Guénette, Murphy, & Allard, 2015), and Spanish (Bowles, Toth, & Adams, 2014).

In sum, intelligibility is widely considered to be the most important aspect of L2 pronunciation. Intelligibility fails when listeners cannot associate a speaker's sounds with linguistic forms. Accordingly, segmental features are perhaps the most critical aspects of pronunciation to be mastered by L2 learners, as they form the basis of word forms, though it is not necessary to have native-like production of all segments. While duly noting the communicative functions of suprasegmental features and the effect they can have on listeners (e.g., Kang, Rubin, & Pickering, 2010) and the role of the listener in maintaining intelligibility through contextual cues and communicative strategies, I have focused the KPD and the remainder of this literature review on phonemes as criteria for identifying pronunciation weaknesses across a wide range of L2 proficiency levels.

The Linguistic Basis of Intelligible Pronunciation

As these frameworks of language proficiency suggest, segmental aspects of L2 phonology form the foundation of successful communication for L2 users. At a basic linguistic level, all phonemes are useful in distinguishing higher-level linguistic forms (i.e., words), allowing access to their associated meanings. With natural languages being composed of tens or hundreds of thousands of words, there are bound to be many that have highly similar forms, e.g., minimal pairs which differ by a single phoneme (*cap* and *cab*, in English). The concept of Functional Load (FL), first described by Brown (1988), explains the importance of segmental phonological contrasts by examining how much utility (a) phoneme contrasts (e.g., /n/-/m/) and (b) individual phonemes have when it comes to distinguishing the words that compose a language's lexicon (see also Oh, Coupé, Marsico, & Pellegrino, 2015). For example, the English

contrast of /n/-/t/ has a high FL due to the frequency at which those two phonemes distinguish similar words (e.g., *nap/tap, night/tight*). Because those phonemes are frequent in the lexicon and form crucial contrasts with other phonemes, /n/ and /t/ (as individual phonemes) are said to have high FLs. Oh et al.'s recent survey of FL in several typologically different languages suggested that consonants generally have higher FL than vowels, with vowels gaining some ground when considering inflectional derivations. FL information within a language can provide insights as to how likely a mispronunciation of a phoneme will lead to listener difficulty. Examining L2 production as understood by L1 listeners, Munro and Derwing (2006) found that (a) utterances with higher FL errors and (b) utterances with more high-FL errors created greater difficulty in listener understanding. This implies that some mispronunciations are more severe and present a greater threat to intelligible speech for L2 learners.

I now turn to the linguistic specifics of Korean pronunciation. According to Shin, Kiaer, and Cha (2012), the contemporary Korean spoken in South Korea (and particularly by younger people in the capitol region) has 28 phonemes. Among these phonemes are 7 vowels, 19 consonants, and 2 glides; the glides combine with vowels to form 10 diphthongs (Table 1). Cross-linguistically, Korean is somewhat rare in that it has a *tension* featural distinction for some consonants, resulting in a two- (tension) or three-way (tension X voicing) distinction among consonants with the same place and manner of articulation. Tensing requires pharyngeal articulation, somewhat longer stop/fricative/affricate duration, and tends to result in higher pitch (F_0) of the following vowel. Korean has a (C)(G)V(C) syllable structure. The allophonic distribution of Korean consonants is generally sensitive to syllable and word context. One notable idiosyncrasy involves the /s/ and /s*/: when these phonemes are followed by the vowel /i/ or glide /j/, the place of articulation changes to the alveopalatal area.

Some research on Korean Functional Load suggests that consonants are more critical than vowels for distinguishing words (Oh et al., 2015). Among Korean consonants, /n, k, l, s, t/ have the greatest FL, in that order. Additionally, the contrast between /l/ and /n/ has a notably higher FL than other contrasts; /n/ features in other top-ranking contrasts, too. Although vowels are somewhat less critical, the vowels /i, a, o/ are comparable in FL to the previously listed consonants, and several vowel contrasts carry greater FL than most consonant contrasts, e.g., /i-ɛ/, /o-i/, /i- ɑ/. Thus, it seems a fair assessment to say consonants and vowels are of similar, if not equal, importance in shaping and distinguishing words.

Table 2.1

Korean Phoneme Inventory

Consonants	Bilabial	Alveolar	Alveopalatal	Velar	Glottal
Stop					
Lax	p	t		k	
Tense	p*	t*		k*	
Aspirated	p^h	t^h		k^h	
Fricative					
Lax			s		h
Tense			s*		
Affricate					
Lax			tɕ		
Tense			tɕ*		
Aspirated			$tɕ^h$		
Nasal	m	n		ŋ	
Liquid		l			

	Front	Back	
Vowels	Unrounded	Unrounded	Rounded
High	i	ɯ	u
Mid	ɛ	ʌ	o
Low		ɑ	
Glide	j, w		

Note. Information compiled from Shin et al. (2013).

The Cognitive Basis of Pronunciation

From a psycholinguistic perspective, phonemes play a key role in models of spoken word recognition: Incoming soundwaves, after being decoded into phonemic units, can then inform the activation of potential lexical matches and suppression of competitors (McClelland & Elman, 1986; McQueen, Norris, & Cutler, 1994). Or, in other words, if a sound produced by a speaker is unrecognizable as the intended phoneme, the listener's word recognition is impeded, and intelligibility may suffer. Failure to identify a speaker's intended word causes immediate deterioration in intelligibility and can potentially cause ripple effects in subsequent word recognition: the previously (mis)identified word contributes to top-down activation and suppression processes, where the listener attempts to apply their understanding of the current discourse context and general world knowledge. Research has found that segmental pronunciation can substantially affect the intelligibility of L2 utterances (Isbell, 2017; Kang et al., 2018a, 2018b; Loewen & Isbell, 2017; Zoghbor, 2018).

Before proceeding further in discussing the role of phonemes in pronunciation, it is appropriate to highlight some important issues related to the mental representation of phonemes in speakers and listeners. As Field (2014) pointed out in his synthesis of historical and contemporary perspectives on phonemes, it is unlikely that language users (L1 or L2) possess a distinct, minimalist inventory of phonemes in their minds due to substantial variation in phonetic realization across speakers (e.g., pitch differences among men, women, and children) and contexts including local linguistic contexts (e.g., those leading to co-articulation phenomena) as well as social contexts (e.g., phonetic differences in phoneme realizations among national, social, and ethnic varieties of a named language). Instead, Field suggested that theories which account for this variation, such as multiple trace-based accounts that center on users' experience hearing

countless variations of individual sounds (Bybee, 2001), are more plausible. In other words, primarily (if not only) through substantial language experience can users of a language develop a sense of what acoustic patterns underlie the sounds used to encode meaning in language, i.e., the abstractions linguists refer to as phonemes. Thus, Field argued that while phonemes are still a valid unit of discussing learner pronunciation and intelligibility, more sophisticated input-based approaches are needed for building up learner knowledge of variation in phoneme realizations.

The foundational role of L2 phonemes in intelligible speech conveniently aligns with recommendations for DLA specifications offered by Harding, Alderson, and Brunfaut (2015). In their article, Harding et al. discussed potential avenues for implementing DLA that specifically targets L2 reading and listening skills. For L2 listening, Harding and colleagues cited Field's (2013) model, which is based on Cutler and Clifton's (1999) well-known model of L1 listening, as a strong and detailed model of language ability that could form the basis of a diagnostic test. Harding et al. (2015) emphasized the model's "obvious scope… for operationalizing elements of this model through discrete assessment tasks" (p. 329). A full summary of this model is beyond the scope of this dissertation, but I will highlight the lower-level processes most critical to DLA (Figure 2.1). In Field's model, auditory input is decoded into phonemes, syllable structures, and suprasegmental information which is subsequently used in lexical search. While the focus of this dissertation is on diagnosing pronunciation, an aspect of language production, lower-level listening processes (i.e., perception of phonemes in speech for word recognition) play an important role in L2 pronunciation development, an idea I will return to.

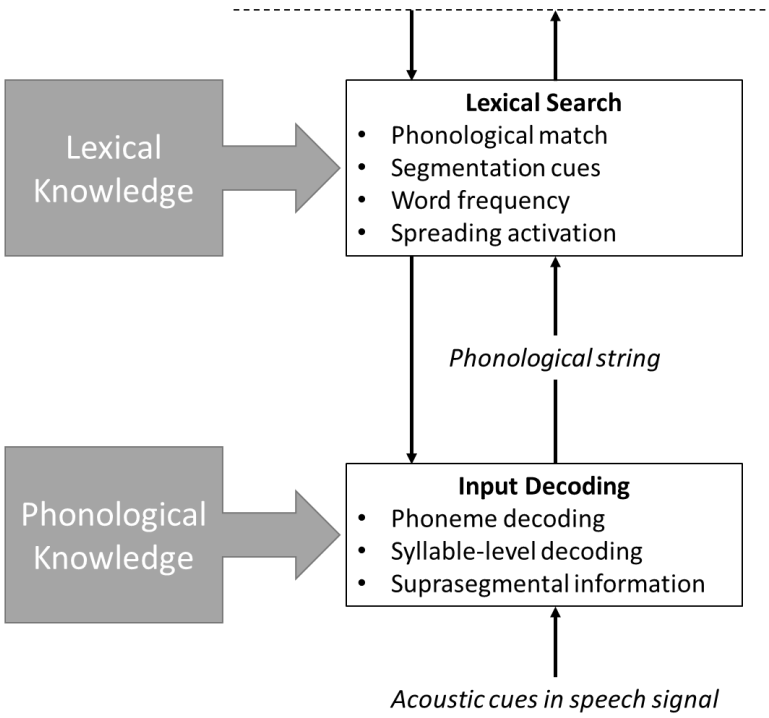


Figure 2.1. Lower-level listening processes, based on Field (2013, p. 97).

Although Harding et al. (2015) did not specifically address language production, their advice in selecting a detailed process model of language ability can easily be applied to productive skills. Once again, Field's work has proven valuable. In 2011, Field articulated a process model of L2 speaking (Figure 2.2), based on Levelt's (1993) seminal L1 speaking model. Field's speaking model and listening models are not simply mirror images, but their parallels in lower-level processes are obvious: The phonetic encoding and phonological encoding of speaking align with the input decoding and lexical search of listening. In the speaking model, messages that have been grammatically encoded are then converted to strings of phonemes. These phonemes direct phonetic articulatory settings that ultimately result in sounds being produced. Importantly, both speaking and listening rely on phonological knowledge in the lower-level processes. Field's (2011) speaking model also connects speaking to listening: After

articulating a chunk of speech, the speaker can self-monitor in order to repair mispronunciations or dysfluencies (or other errors).

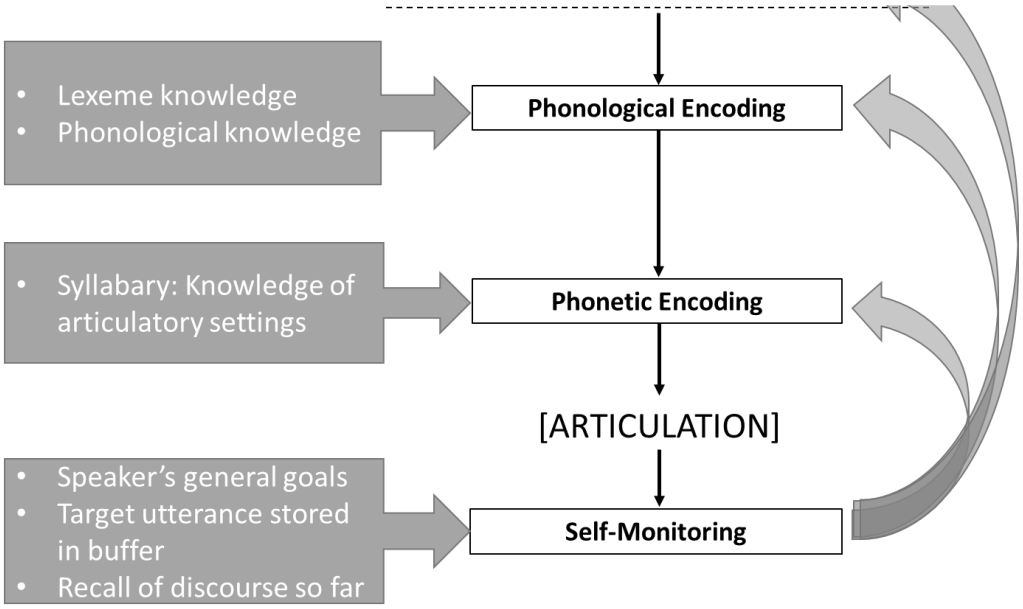


Figure 2.2. Lower-level speaking processes, based on Field (2011, p. 77).

An important feature of Field's (2011, 2013) models is a distinction between knowledge and processes. For example, in Figure 2.2, *phonological knowledge* and the *syllabary* are sources of knowledge that are drawn on in speech production. On the other hand, *phonological encoding* and *phonetic encoding* are processes; these may also be thought of as (sub)skills or abilities. Thus, it is possible for a speaker to possess the relevant knowledge to produce a sound (i.e., they might know what a segment sounds like, or how/where their speech articulators operate to produce it), but they may nonetheless fairly accurately articulate a sound at times due to a hiccup or failure in a process. Similarly, learners may have imperfect knowledge (e.g., a poorly-defined phonological category), but processes that are sufficiently tuned to produce intelligible (if not native-like) articulations more often than not (e.g., Sheldon & Strange, 1982). This distinction can be related to the competence-performance dichotomy in language assessment and can also be

related to different knowledge types in SLA theory. For example, in DeKeyser's (2017) application of skill acquisition theory to SLA, a distinction is made between declarative knowledge (knowledge *of*) and proceduralized (and ultimately automatized) knowledge (knowledge *how*, with an emphasis on expedient use). This distinction has been recently picked up by Saito and Plonsky (in press) in their measurement framework for L2 pronunciation, which contrasts controlled pronunciation tasks, which largely tap into declarative knowledge bases, and spontaneous production tasks, which measure accuracy and efficiency in processing.

The Developmental Basis of L2 Pronunciation

Although L2 learners do tend to develop greater control over phonological features alongside their overall oral proficiency (Kang & Moran, 2014; Saito, Trofimovich, & Isaacs, 2016), it has long been understood that the development of L2 pronunciation is not a straightforward, predictable process. Abercrombie (1949, p. 118), in what would become an early landmark in pronunciation teaching, noted that:

> "People vary, to a surprising extent, in ability to learn the pronunciation of foreign language. Every phonetician must have had the experience, at some time or other, of meeting a person to whom the imitation of the most exotic sounds at first hearing presented no difficulty at all. At the other extreme are a more numerous minority who are hopelessly recalcitrant, and for whom any deviation from the native sound system is apparently impossible." (Abercrombie, 1949, p. 118).

Decades of subsequent observations from teaching practice, empirical research, and theory building would support Abercrombie's description of large variability in L2 pronunciation development outcomes. In this section, I review research on L2 pronunciation development, highlighting several salient factors found to influence this variability, including learner age,

cross-linguistic influence and bi/multilingualism, experience, the relationship between perception and production of L2 sounds, and instruction. Discussion of these factors is followed by a review of L1 and L2 Korean segmental development that is useful for establishing general expectations of pronunciation difficulties in the present study.

Age

One of the most consistent and robust findings in research on age-related constraints in SLA is related to phonological development: Learners who begin study of an L2 past the age of six (or more liberally, past puberty) are generally unlikely to acquire native-like perception and articulation of L2 sounds (Abrahamsson, 2012; Flege et al., 1999; Long, 2013; Piske, MacKay, & Flege, 2001). Thus, for older L2 learners, age of onset plays an important predictive role in defining the endpoint of pronunciation development. It is for this reason that Levis' Intelligibility Principle has taken a firm hold on the field of L2 pronunciation, as native-like pronunciation outcomes are simply not a realistic goal for many L2 learners (and in some cases may not be desired, e.g., by learners who strongly identify with their national/ethnic group). At the same time, even though much adolescent and adult L2 learning does not result in nativelike phonologies, L2 phonology does develop, typically in the direction of more intelligible and/or target-like perception and production of L2 sounds and sound patterns. Instruction has been found to improve various aspects of L2 phonology, including phoneme perception (e.g., Flege, 1991; Hardison, 2005; Thomson, 2012) and production (e.g., Thomson, 2011; Lee et al., 2015).

Cross-Linguistic Influence

A key feature of L2 pronunciation learning that distinguishes it from child L1 acquisition is the bi/multilingual phonemic inventory. The starting point for L2 learners is not a blank slate, and L2 learners do not simply turn off their L1 phonemic inventory or develop an entirely

separate phonological system when learning or using the L2. It is widely observed that learners (to varying degrees) substitute L1 phonemes, follow L1 syllable structure constraints, and apply L1 prosodic patterns to L2 speech. The earliest theoretical accounts of L1 transfer or influence, such as the Contrastive Analysis Hypothesis (CAH), relied completely on cross-linguistic differences in phonological systems to make strong predictions about difficulty and learning for various L1-L2 pairings (Lado, 1957; Stockwell & Bowen, 1965). In brief, the CAH predicted that phonological features not present in the L1 would be very difficult to acquire in the L2, features that are optional in the L1 yet obligatory in the L2 would be a moderate challenge, and features that were present and obligatory in both languages would be easily acquired (or transferred).

Ultimately, many specific predictions based on the CAH failed to pan out, and similarly the CAH failed to account for variation in learning and accuracy within L1 groups. Namely, L2 pronunciation research has shown that considerable variation in phoneme articulation exists within groups of speakers from the same L1 background (e.g., Abrahamsson, 2012), and that speakers from the same L1 group can vary greatly in the overall strength of foreign accent (e.g., Kang et al., 2010; Munro & Derwing, 1995). As Munro, Derwing, and Thomson (2015) pointed out, just because a contrastive analysis predicts a challenge based on L1-L2 pairing, in many cases the potential challenge is either (a) overcome quickly or (b) never actually presents substantial, long-lasting difficulty and thus, in either case, does not require much specific instruction. Further, while the L1 has an undeniable influence on L2 phonology and pronunciation, not all learners are influenced by their L1 in exactly the same way or to the same degree, and learners may make progress in different aspects of L2 pronunciation at different rates.

Additionally, L1 varieties and multilingualism create conditions that make L1-based predictions of pronunciation difficulties more difficult to carry out and less reliable for teachers. For example, (Mandarin) Chinese is one of the world's most widely spoken first languages, which might make some knowledge of Mandarin Chinese phonology useful for second/foreign language teachers in many contexts, but the varieties spoken throughout the Chinese-speaking world vary phonologically. While pronunciation textbooks have commonly provided information on the phonological systems of various learner L1s (e.g., Avery & Ehrlich, 1992; Kwon, 2017), they rarely provide information on non-dominant regional varieties. How much can a teacher be expected to know about their students' specific L1 varieties? At the same time, language classrooms are increasingly being populated by multilingual learners; many foreign language learners are technically L3+ learners. It remains unclear whether native languages or L2s primarily shape L3 phonology, but it is possible for L2 articulatory settings to be transferred to an L3, even when L1 settings would result in production closer to native-like targets (Llama, Cardoso, & Collins, 2010). Relevant specifically to Korean, Chen (2018) illustrated an interesting case where some Korean phonological difficulties experienced by Taiwanese Mandarin speakers, who also had some knowledge of Taiwanese and/or Hakka, differed in pronunciation difficulties from what the CAH would predict for Mandarin speakers from China.

Although the strongest accounts of L1 transfer such as the CAH have been abandoned, cross-linguistic influence remains prominent in theoretical accounts of L2 speech learning. L2 phonological development is adequately described by models involving perceptual assimilation (Best & Tyler, 2007; Flege, 1995). Empirical research has demonstrated that L1 phonemes remain active during L2 speech perception and influence word recognition (Imai, Walley, & Flege, 2005; Weber & Cutler, 2004), providing strong evidence for the influence of learners'

pre-existing phonemes on the L2 phonological system. Perceptual assimilation, for L2 learners, happens when a newly-encountered L2 sound is parsed as a similar, existing phoneme (typically, but not always, a L1 phoneme). For example, an English learner of Korean may assimilate the Korean /k/ and English /k/, as they share a number of acoustic and articulatory similarities. However, the same learner may also assimilate Korean /k^h/ to the L1-L2 /k/ phoneme category. In part, this is because aspiration is not phonemic in English, leading to the learner perceiving the two sounds as being more similar than they really are in Korean. With enough input, learners can separate these assimilated L1-L2 phonemes, but it is not guaranteed, and it can take quite a long time for L2 phonemes to become distinctly and robustly represented in the learner's inventory (recall Field's (2014) discussion of phoneme representation in the mind). Individual learners will also vary in the rate and potentially in the order of distinguishing new L2 phonemes; this variation is likely driven in part by differing amounts of L2 use/exposure and individual differences such as motivation, musical aptitude, and other cognitive/neurological differences (see Ingvalson, Ettlinger, & Wong, 2014, for a discussion of the latter).

Experience

The most dynamic period of L2 speech learning tends to occur within the first year or two of exposure to the language (Flege, 1988), at least in immersion contexts, which Derwing and Munro (2015) referred to as the *Window of Maximal Opportunity*. Within this window, development may not always be uniformly in the direction of target-like representations and articulation; sometimes learners experience ups and downs in accuracy due to the process of building new representations and reorganizing their phoneme inventories (Holliday, 2016). After the window passes, learners' L2 phonology may fossilize, whereby pronunciation of segments and suprasegments ceases developing toward more intelligible, comprehensible forms (Derwing

& Munro, 2013). For instructed L2 Korean learners in a low-input foreign language environment, I and my colleagues (Isbell, Park, & Lee, 2019) found support for this window as well. We found that students within their first year of exposure to Korean as a foreign language showed rapid improvements in pronunciation (greater comprehensibility as well as lower error rates) regardless of treatment, while second-year students without pronunciation instruction showed no improvements. The state of interlanguage phonology and corresponding quality of L2 pronunciation that exists after this period is perhaps of greater interest: While giving beginners a good start to L2 speech sounds and pronunciation is important, the greater challenge lies in the gradual disentanglement of assimilated phonemes and the development of more intelligible and comprehensible pronunciation, generally in the direction of target language norms. From the perspective of diagnosis and targeted instruction, weaknesses discovered in learners who have most likely cleared the Window of Maximal Opportunity are likely to be more stable and less likely to improve without instruction in a shorter time period (Derwing & Munro, 2014).

Instruction

As previously mentioned, L2 pronunciation instruction is known to be effective (Lee et al., 2014) and durable (Couper, 2006). Moreover, instruction is capable of aiding learner development even when long-term fossilization of L2 phonology has occurred (Derwing, Munro, Foote, Waugh, & Fleming, 2014). The L2 pronunciation instruction literature, both empirical and practice-oriented, is rich with techniques that promote pronunciation learning, such as *shadowing* (speaking alongside an audio model, Foote & McDonough, 2017), *read aloud* (reading text aloud, with feedback if possible, Horgues & Scheuer, 2014; McCrocklin, 2019), *choral repetition* (teacher led repetition of words/sentences, Baker, 2014), *explicit instruction* of acoustic and articulatory features (explaining how to produce sounds and what they should sound

like, e.g., Derwing et al., 1998), *communicative tasks* (such as conversation or information-gap tasks, Loewen & Isbell, 2017; Saito & Lyster, 2012), and *listening to self-recordings* (recording oneself and listening for aspects to improve, often comparing to a model) and *using visual aids* (looking at acoustic visuals of self- or other-productions, Hardison, 2004), among many others. While a complete treatment of the various types of pronunciation instruction and their associated benefits and limitations is beyond the scope of this chapter (though see Celce-Murcia et al., 2010; Derwing & Munro, 2015; Thomson & Derwing, 2015; Lee et al., 2015), I will revisit some specific instructional techniques later in the dissertation when relevant. Here, I focus my review of the literature on more general aspects of pronunciation instruction most relevant to diagnostic assessment.

One finding from the L2 phonological development literature that has important implications for instruction is the perception-production link (Flege, 1991; Derwing & Munro, 2015). Recent research in cognitive science has shown that areas of the brain responsible for articulation can also become active during and contribute to speech perception (Möttönen & Watkins, 2009). These same areas of the brain can also contribute to the learning of novel phonological forms (Nora, Renvall, Kim, Service, & Salmelin, 2015). In some ways, this relationship is quite intuitive: When a language user has a strong, consistent ability to perceive a specific sound, it suggests that they have a strong underlying mental representation of the sound and its distinguishing features, which in turn would lead to consistent, accurate articulation of the sound. Strong interpretations of the perception-production link include accurate perception (a) preceding and (b) predicting accurate production of L2 sounds. For example, if a Japanese learner of English cannot perceive the difference between /l/ and /r/ (instead assimilating both sounds to their Japanese /ɾ/ phoneme), it is unlikely that they will be able to produce the

distinction. In some cases, being trained to perceive a L2 phoneme results in improvements to production (e.g., Lee & Lyster, 2017; Thomson, 2011; see also Sakai & Moorman's 2018 meta-analysis supporting such findings across 18 different studies). At the same time, some learners will be able to quite reliably decode a given phoneme from speech but struggle to articulate it in their own production: English learners of Spanish frequently struggle in producing the trill /r/, but are usually quite able to distinguish it from the flap /ɾ/ in listening.

Another key finding of research on L2 speech perception and production is that focus on form, i.e., promoting awareness and directing attention to linguistic (in this case, articulatory/acoustic) form, is beneficial to learning (Derwing & Munro, 2015; Guion & Pedersen, 2007; Kennedy & Trofimovich, 2010; Moyer, 2014; Saito, 2018, Venkatagiri & Levis, 2007). Thomson (2012) discussed the role of attention on phonological learning, whereby learner attention to phonological form leads to improvement of perception and in turn production. Focus on form is often operationalized as corrective feedback in speech perception and pronunciation studies, where learners are alerted to their errors and given information to support more target-like performance in the future (e.g., Lee & Lyster, 2016, 2017). Explicit focus on form instruction is also useful with a primary focus on production: Learners receive explicit phonetic instruction prior to carrying out practice and/or communicative activities, where learners receive feedback on their production involving the provision of model input from a teacher or peer, and then go on to gradually produce more intelligible articulations with continued practice (e.g., Derwing et al., 1998; Isbell et al., 2019; Gooch, Saito, & Lyster, 2016; Saito & Lyster, 2012). This progression from explicit articulatory and acoustic knowledge to consistent, intelligible production aligns well with skill acquisition approaches to SLA (DeKeyser, 2017), where learners, particularly in instructed settings, first acquire declarative knowledge of L2 speech

sounds and eventually develop efficiency in producing them through attention-focused practice (Saito and Plonsky, in press).

When attempting to diagnose learner pronunciation issues for the purpose of setting instructional targets, Lado (1961) emphasized the assessment of both perception and production, highlighting that testing only one or the other results in an incomplete picture:

> If a student pronounces a sound contrast in a foreign language he will also hear it. … At the same time, students learn to hear sound contrasts usually before they are able to pronounce them, and so in testing production we would not discover everything the student has learned to hear. And what is more to the point in this chapter, by testing recognition of the sound segments we will not have tested what the student has learned to pronounce. Finally, the distance between recognition and pronunciation is not the same for every student. Some students who learn to hear reasonably well still have very poor pronunciation, whereas others learn to pronounce almost as well as they can hear. (Lado, 1961, p. 78)

Furthermore, as seen in the excerpt, Lado highlighted the variability in student speech learning and suggested the potential of identifying different sorts of profiles that characterize learners' pronunciation. Thus, instruction can benefit from pinpointing the source of individual learners' difficulties: An English instructor might begin with perception training for the Japanese learner who cannot perceive or produce /l/, or at least tackle both modes simultaneously. On the other hand, the same instructor may have another student work exclusively on production if the student can reliably hear the difference between /r/ and /l/. While the stronger claims of the perception-production link are up for debate, there is nonetheless a straightforward pedagogical argument for establishing perception first: Aural feedback on learner pronunciation has to be interpretable, and if a learner cannot tell the difference between what they produce and the model provided by a program, textbook audio CD, or teacher, adjustments to articulation seem less likely to occur. Exemplifying this, classroom-based pronunciation instruction research by Saito and Lyster (2012) showed that corrective feedback on pronunciation in the form of recasts, requiring

learners to hear the difference between their own productions and the model provided in feedback, could induce changes in L1 Japanese learners' phoneme representation and articulation of English /r/, a feature that was considered to be fossilized for many of the learners.

Research on L1 and L2 Korean Phonological Development

Research on the acquisition of Korean phonemes, receptively and productively, has yielded some useful insights. In child L1 Korean acquisition, Kim, Kim, and Stoel-Gammon (2017) report that the earliest acquired consonants tend to be /p, p*, p^h, t*, k, m, n, h/ while the latest acquired consonants are /tɕ, $tɕ^h$, s, s*, l/ (see also McLeod & Crowe, 2018, which synthesizes the findings from several L1 Korean consonantal acquisition studies). Consonants tend to be acquired earlier in syllable-initial contexts and later in clusters or word final positions. From a featural perspective, there are reasonably clear orders of acquisitions for place and manner of articulation. For place, the order is roughly bilabial → alveolar → velar → alveopalatal → liquid. For tension and voicing, children follow a tense → aspirated → lax sequence; for fricatives lax precedes tensed. These patterns can serve as a baseline for difficulty expectations and acquisition orders where L2 acquisition data are absent or insufficient.

While the data from L1 Korean children are potentially useful for understanding L2 development, research on L2/heritage learners of Korean (with English as an L1/dominant language) has shown a notable contrast: L2 learners tend to struggle with tensed consonant articulation (e.g., /k*/) even at advanced proficiency levels (Lee et al., 2009; Oh, Jun, Knightly, & Au, 2002) or considerable exposure (Holliday, 2015). Holliday (2015), a longitudinal study of Mandarin speakers' acquisition of Korean's lax/tense/aspirated stop consonant distinction, found that learner development trajectories varied considerably and that learners struggled to reliability produce the distinction even after one year of residence in South Korea. Yu (2016) even found

that young Korean-English bilinguals' tensed consonants are less tensed than their monolingual peers'. Some research suggests that among adult L1 English speakers, aspirated consonants are mastered first (e.g., Tark, 2016). Recall that the tensed feature is learned very early on by L1 children. Furthermore, even when adult L2 Korean learners have acquired a phoneme, their articulation may still differ from native speakers (NSs), such as through the use of different areas of the tongue when making alveolar stops (Ko, 2013). Nonetheless, several similarities do exist between child L1 and adult L2 learners. L2 learners also have been shown to struggle with /l/, particularly with respect to its allophone distribution (Kim & Park, 1995; Kim, 2007; Lee, 2012). While this aligns with the pattern children exhibit, it is at the same time somewhat surprising that even L1 English learners struggle with accurate production: Their L1 contains /l/ and /r/ (and a flap [ɾ], an allophone of /l/ in Korean, as an allophone of /t, d/), so the articulation of Korean phonemes is mostly within their existing oral-motor skillset. Kim (2015) suggested that syllable context plays an important role in the accuracy of production for /l/ and other consonants. Following Lee (2012), the apparent hierarchy of ease for /l/ allophones by position is onset > coda ≈ geminate (where an /l/ in a coda position is followed by an /l/ in the onset of the following syllable). The pattern of coda articulations being more difficult mirrors child L1 acquisition order (Lee, 2012). Empirical findings have also suggested that some back vowels, particularly /ʌ, ɯ/, can present a challenge to L1 English speakers (Kim & Silva, 2003). These findings on L2 Korean phoneme acquisition, complemented by the L1 research, provide a suitable basis for examining the overall hierarchy of difficulty found among targets in the KPD. However, it is important to note that most of the published research on L2 Korean phonological acquisition is based on L1 English speakers; developmental patterns for learners of other L1 backgrounds might be expected to differ.

Although relatively few in number, studies on pronunciation instruction for L2 Korean suggest that L2 Korean learners can improve their pronunciation, just like learners of any other language. Tark (2016) demonstrated that form-focused instruction with corrective feedback helped learners improve their mastery of stops, fricatives, and affricates. Focusing on some of the same targets, Shin (2007) highlighted how perception training for three-way consonant distinctions (i.e., lax, tensed, and aspirated stops) led to improvements in both learners' perception and production. Thus, even for features commonly observed as difficult for learners, good instruction can make a difference. Instructional treatments with a broader scope have also benefitted learners. In my prior work with colleagues, Isbell et al. (2019), students in their fourth semester of Korean study in the United States improved their speech comprehensibility after an 8-hour instructional treatment that targeted a set of segmental and suprasegmental features. Their fourth-semester peers who did not receive the treatment showed no development, indicating a benefit to this broader-scope Korean pronunciation instruction treatment. However, the evidence of improvement was not extremely robust, raising the question of whether better-targeted instruction, suited to learners' individual needs, could have made a bigger impact.

The Goal: Diagnosing L2 Korean Pronunciation

A diagnostic instrument that can help teachers and learners identify pronunciation weaknesses and in turn motivate well-targeted instruction is both desirable and plausible. For this dissertation, I developed and validated a new diagnostic language assessment for L2 Korean pronunciation, the KPD. All the inferences outlined previously in Chapter 1 (see Figure 1.1) are relevant to the KPD. However, some inferences for DLA are perhaps more important than others. In particular, I consider the utilization inference (or impact in Bachman & Palmer's (2010) terms) to be key in DLA, aligning with Lee's (2015) emphasis on how DLA results are

used for instruction. For instruction, I adopt Housen and Pierrard's (2005) definition of second language instruction: "any systematic attempt to enable or facilitate language learning by manipulating the mechanisms of learning and/or the conditions under which these occur" (p. 2). This broad view of instruction encompasses not only what a teacher does with students in traditional classrooms, but also the aims of learning materials (e.g., textbooks, software) and the deliberate learning activities undertaken by individual learners (Loewen, 2015); *learning activity* is a term I will use interchangeably with instruction throughout the dissertation to describe the relatively informal and ad-hoc yet still deliberate language-focused learning efforts used by learners. Thus, validity arguments for a DLA procedure should include impact on teachers in classrooms/tutorial sessions, on learner awareness and autonomous learning activity, and the selection of learning materials by either teachers or learners.

Figure 2.3 is my proposed validity argument for using the KPD to inform the learning and instruction of KFL/KSL learners. On the right side of the figure are the sources of information that provide backing (gray boxes) for key inferences (arrows pointing upward) in the argument for using KPD results. When comparing the KPD validity argument to the more generic structure in Figure 1.1, readers may notice that I have included two additional inferences: (1) operationalization and (2) usefulness and impact. The operationalization inference draws on the work by Chapelle et al. (2008, 2010), who highlighted in greater detail how the description of target abilities and situations of use should inform test design and item/task construction. Often, theoretical and descriptive grounds are grouped with test observations in validity arguments, but in the case of the KPD I feel that a finer distinction here is useful given Alderson et al.'s (2015) strong emphasis on detailed theory informing the construction of discrete, well-targeted items/tasks. I consider the final inference near the top of the figure, usefulness and impact, based

on how effectively learners are able to make improvements to their pronunciation after applying KPD results. Finally, the entire validity argument for using the KPD can be evaluated by synthesizing the strength of supporting evidence across inferences and considering outstanding shortcomings or gaps in support.

The boxes on the left detail how backing supports the inferences. Some backing for the inferences in the proposed argument already exists in the form of theoretical backing, test design, and initial piloting efforts. The former has been elaborated on already, and I detail the latter two shortly. Other backing is needed; these gaps are presented in the form of research questions (RQs). In total I have identified nine RQs, several with sub-questions (Note: RQs are reproduced in the Methods chapter for easier reading). More concisely, the work undertaken as part of this dissertation can be summarized in the following four aims:

- Aim 1: Create, pilot, and revise test items to result in a final form of a diagnostic instrument that a) functions well and takes minimal time to administer, b) provides detailed, meaningful information about a learner's mastery of Korean phonemes, and c) can be used productively by Korean language teacher and learners.
- Aim 2: Field test the final form with a suitable number of Korean language learners in order to collect normative data that facilitates the interpretation of results and consideration of diverse learner profiles.
- Aim 3: Study the relationship between diagnostic test scores and spontaneous speech, and plot phoneme acquisition patterns across proficiency levels.
- Aim 4: Study how Korean language teachers and learners interpret and act on test results.

The remainder of this dissertation documents the results of these efforts and reflects on the evidence they provide pertaining to the valid diagnosis of L2 Korean segmental pronunciation.

The next chapter describes the culmination of Aim 1, Chapters 5 and 6 cover Aim 2, Chapters 6 and 7 cover Aim 3, and Chapter 8 represents initial an initial exploration of Aim 4.

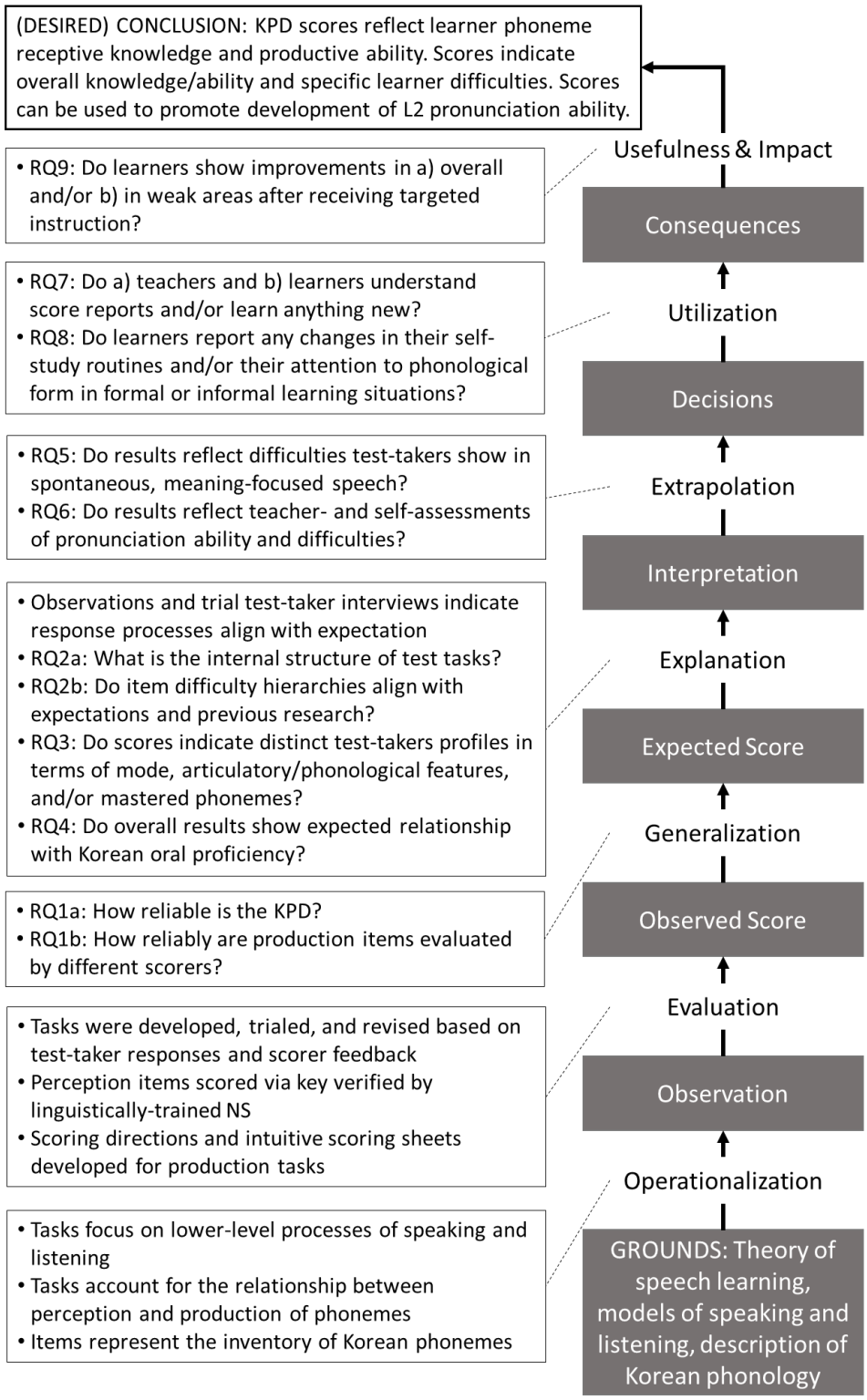


Figure 2.3. A proposed validity argument for using the KPD to inform learning and instruction.

CHAPTER 3: DESIGN AND DEVELOPMENT OF THE KOREAN PRONUNCIATION DIAGNOSTIC

In this chapter, I present the design and development of the KPD in detail. The design section features a detailed description of the operational version of the KPD used in the remainder of this dissertation. The development section chronicles the pre-operational changes to test design and items as a result of two phases of piloting.

Design

This section lays out the test purpose, appropriate uses, test specifications, item specifications, item creation process, scoring, and score reports.

Test Purpose

The purpose of the Korean Phonology Diagnostic (henceforth KPD) is to pinpoint strengths and weaknesses in L2 Korean learners' receptive and productive phonemic inventories, with the goal of then positively influencing learning through learner awareness and instructional remediation. Results are intended to be informative and instructionally relevant to individual learners and their teachers or tutors. Based on Field's models of listening (2013) and speaking (2011) in Chapter 2, the test focuses on phonological knowledge as it relates to lower-level listening and speaking processes. While the KPD would be suitable for a wide range of proficiencies, it is likely not suitable for true beginners or extremely novice learners. The KPD also requires a basic level of familiarity with *hangeul* (한글, the Korean alphabet). Familiarity with high-frequency Korean vocabulary is helpful, though does not need to be comprehensive.

Appropriate Uses

The KPD is intended to be used by students, teachers, and potentially language programs to increase awareness of pronunciation difficulties and guide instructional decisions relevant to

classroom instruction and autonomous learning activity. These are inherently low-stakes decisions with minimal potential for negative consequences. As examples, the following uses of the KPD would be appropriate:

- A Korean learner, now an undergraduate business student, feels that she struggles with intelligible pronunciation. She asks her old Korean teacher for guidance; the teacher administers and scores the KPD and provides feedback to the learner. The learner then selects material from a Korean pronunciation textbook to practice on her own time. The learner also focuses more on her perception and production of difficult sounds when using Korean in her coursework.
- A learner asks his teacher for help with his pronunciation. The teacher administers and scores the KPD and provides feedback to the student. The teacher meets with the student after class once per week for brief tutorial sessions and assigns some practice materials. The student pays more attention to difficult sounds during class time.
- A Korean language program offers a range of short-term supplemental courses, such as academic presentations, TOPIK preparation, and pronunciation fundamentals. To help ensure that students who sign up for the segmentally-focused pronunciation fundamentals course stand to benefit from it, the KPD is administered to ensure that students with generally strong control of Korean sounds are referred to other courses and only students with segmental difficulties take the pronunciation course. The KPD results are passed on to the teacher of the course to inform more detailed instructional decisions.
- A teacher of a pronunciation class wants to identify common difficulties and assign individualized homework to students. The teacher administers the KPD to each student in her class of 10 students and uses the score reports to select common targets for group

instruction. Targets not covered in whole-class sessions are assigned to learners according to their needs.

The KPD should not be used to interpret learners' overall Korean proficiency, speaking ability, or even overall pronunciation quality. The KPD should not be used to make decisions about immigration eligibility, visa status, university entrance, or employment. The KPD is inappropriate for high-stakes decisions, especially those meant to be based on more generalized communicative ability in Korean. The following examples illustrate some inappropriate uses of the KPD:

- A university in Korea has been using the Test of Proficiency in Korean (한국어능력시험, TOPIK, http://www.topik.go.kr) in making admission decisions for international students. The TOPIK does not have a speaking component, and the university is looking for a freely-available, quick, and easy-to-score speaking test. The university decides to use the KPD and require at least a 70% average across all phonemes in production for admission.
- A university in the United States has received complaints about the accents of their non-native teaching assistants (TAs) in the Korean program. The program director decides that all TAs should be able to demonstrate mastery of Korean segmental pronunciation in order to provide a good model for students. TAs who show significant difficulty in producing Korean sounds are excluded from teaching duties.
- A Korean coffee shop owner has received complaints about her international student baristas being difficult to understand. When interviewing new baristas, she administers the KPD to non-Koreans and does not make offers to people with "too many" troublesome sounds.

Structure and Item Specifications

The KPD involves two parts, with a total of four tasks (summarized in Table 3.1; a full table of specifications is available in Appendix A). My approach to task selection was informed by Harding et al.'s (2015) recommendations for using a detailed model of language production to focus on lower-level processes. The perception-production link, with its implications for development and pedagogical practice, was also a major influence. My design of the tasks themselves was inspired by recommendations in Lado (1961) and Derwing and Munro (2015) and heeded the latter's recommendation that "materials suitable for classroom testing are similar to many of those used in pronunciation research" (p. 115). I also took note of Munro's (2008) recommendation to avoid relying on a single task type for evaluating a learner's intelligibility, as most speech elicitation techniques have at least one drawback which should be counterbalanced. Item specifications (following Davidson & Lynch, 2002) for all tasks are in Appendix B.

Table 3.1

KPD Design Summary

Section	Task	Brief Item Specification	Number of Items
Production	Picture Naming*	Item: picture of a concrete noun Response: speaking the matching word	154 (in 35 words)
	Nonword Read-Aloud	Item: 1-2 syllable nonword Response: reading aloud the nonword	63
Perception	Pronunciation Judgment*	Item: picture of a concrete noun + audio recording of the word Response: forced choice whether audio recording was (in)correct	72 (plus 40 filler items)
	Nonword Identification*	Item: audio recording of a 1-2 syllable nonword Response: forced choice between two written 1-2 syllable nonwords	63

Note. *Task was part of initial pilot test design.

Within each section and task, the Korean phonemic inventory (Shin, Kiaer & Cha, 2012) served as a basis for selecting the number of items. As mentioned previously, there are 7 vowels, 19 consonants, and 2 glides that combine with vowels to form 10 diphthongs. A minimum of four items per phoneme were included in the Production section and at least three items per phoneme in the Perception section. Given the secondary utility of perception scores in relation to the test purpose and use, I felt it acceptable to collect somewhat less information in order to keep the test length more practical. An important consideration for consonant phoneme targets was the inclusion of targets in different syllable contexts to better capture the phoneme's allophonic distribution. While four items per phoneme was set as a general minimum, several phonemes were featured more due to their prevalence in real words (e.g., the vowels /ɑ/, /i/, /o/, /ʌ/). Some consonants, such as /l/, /s/, and /s*/ have additional items to account for markedly different and previously known to be difficult allophonic realizations. For example, when /s/ and /s*/ (both alveolar fricatives, the latter being tensed) are followed by the vowel /i/ or glide /j/, /s/ is realized as [ɕ] and /s*/ is realized as [ɕ*] (alveopalatal fricatives, the latter being tensed).

Production Tasks. The first part focuses on production and includes a Picture Naming task and a Nonword Reading task.

Picture Naming. For the Picture Naming task, learners are required to say the word that corresponds to a picture they are shown. This type of task is commonly used in assessments of children's L1 speech development (e.g., Kim, Pae & Lee, 2005; Seok et al., 2002). In terms of Field's (2011) process model of speaking, this task requires test-takers to activate lexeme and phonological knowledge to phonologically encode the target word, and then taps knowledge of articulatory settings to complete phonetic encoding immediately preceding articulation of the word. The quality of articulations provides information on phonetic encoding ability and

articulation knowledge, but at the same time may reflect malformed lexeme knowledge (i.e., having an erroneous phonological representation of the target word stored in the lexicon). To mitigate this latter possibility, all words are imageable nouns (thereby avoiding lexemes for potentially malformed verbal inflections) and most fall within the first 1,500 most common words in Korean; some exceptions were included because the words were known to be introduced relatively early in instructional settings (e.g., body parts, animals, foods) or due to a lack of other imageable nouns featuring a target phoneme.

Nonword Reading. The Nonword Reading task requires learners to read aloud a one or two syllable nonword; each nonword has only one target phoneme that is scored. Vowels are assessed in isolation, and consonants and glides are assessed in simple, legal syllable structures: (G/C)V, VC, or VCV. Through this task, written letters are used to tap into phonological knowledge, leading to phonological and then phonetic encoding and articulation similar to the Picture Naming task. To minimize potential interference from issues related to learners' orthographic knowledge, I constructed items that avoid sound-symbol mismatches (i.e., no phonological processes leading to a discrepancy between the written grapheme and the spoken phoneme). Variation in syllable context for consonants only affected allophonic realization of phonemes (e.g., [k̚o], [u.k̚u], [ok̚]). While the Picture Naming task avoids issues learners may have with orthography, the Nonword Reading task helps to ensure that consonant targets are represented in a variety of syllable contexts, but always in ways that are orthographically transparent (i.e., do not involve instances of grapheme-phoneme mismatches, such as nasalization or consonant relinking). For example, I was unable to find a suitable word for the Picture Description task that placed /p^h/ in an intervocalic (CVC) context but covering this was easy to do in the Nonword Reading task.

Perception Tasks. The second part of the KPD focuses on perception and includes a Pronunciation Judgment task and an Identification task.

Pronunciation Judgment. The pronunciation judgment task presents pictures of common Korean vocabulary and shortly after and while the picture is still visible plays an audio recording. The audio recording is either the correct phonological form of the word, or it contains a single phoneme deviation (typically the substitution of another phoneme with mostly similar features); the learner must judge whether the sound they heard was accurate for the picture they saw. This task type has recently been used in experimental psycholinguistics (e.g., Amengual, 2016). Only the items which contain mispronunciations contribute to scores for individual phonemes and features. In terms of Field's (2013) lower-level listening processes, the picture provides learners the target phonological string associated with the lexeme, and then test-takers decode the speech signal they hear and compare the phonemes they have decoded to the target string. If a test-taker can detect the phoneme in the stimulus that does not match the correct form of the word, it is inferred that their mental representation of that phoneme is robust enough to be distinct from the non-target (but somewhat similar) phoneme in the stimulus. This process is admittedly indirect but avoids some of the pitfalls of a task based on, for example, listening and then choosing between two words in a minimal pair. Minimal pair tasks can be difficult to construct due to a lack of minimal pairs, or minimal pairs that are likely to be known by learners. For example, most learners could be expected to know 강 (*river*), but may not know 간 (*liver*), which constitutes a minimal pair based on the /ŋ-n/ contrast in the word-final position. Finding a sufficient number of minimal pair sets where both words were imageable was another concern.

Nonword Identification. The Identification task presents nonword audio, and learners must choose between two written options that differ by only one phoneme. Here, test-takers must

tap into their phonological knowledge to decode the speech signal of each item into short strings of phonemes. Once the string of phonemes has been identified, test-takers select the written representation that best matches what they heard. Like the Nonword Reading task, nonword options consisted of 1-2 syllables; V, (G/C)V, VC, VCV. I created written keys and distractors that avoided any sound-symbol mismatches.

Item Writing

I was the primary item writer. Because my own Korean proficiency has limitations, I relied on a NS informant with a background in applied linguistics and Korean language teaching—a content expert—to verify keys, proofread, and spot any major problems at early stages of item creation. This type of test development arrangement is reportedly common for less-commonly taught languages and was explored in depth by Ryan and Brunfaut (2016), whose case study of a testing company found that testing experts and language informants working together produced higher-quality items. In the specific case of KPD item writing, the language assessment knowledge and classroom experience of the informant was extremely valuable and insightful. Importantly, items were revised after two stages of piloting. This process is documented later in the chapter.

Several key resources supported my item writing. Shin et al. (2013) was the primary linguistic resource consulted to verify information on Korean phonetics and phonology. For the Picture Naming and Pronunciation judgment tasks, I consulted Lee, Jang, and Seo's (2017) *Frequency Dictionary of Korean* for the selection of target words in the Picture Naming and Pronunciation Judgment Tasks. Openly-available picture collections for psycholinguistic experiments were drawn on for images used in these two tasks, including MultiPic (Duñabeitia et al., 2017) and BOSS (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010). I also utilized

images with Creative Commons licenses from www.pixabay.com. In a few cases, I produced original hand-drawn images (e.g., an image for *ramyeon,* Korean ramen noodles). Where necessary, I manually altered, combined, or otherwise edited images from these resources in the Paint.net image editing software.

Audio stimuli for both Perception tasks were recorded by the aforementioned expert informant, a female native speaker of Korean originally from the Gangwon province who attended university in Seoul and is a fluent speaker of Seoul Korean. Stimuli were recorded using a Snowball Blue microphone connected to a desktop computer and the audio recording and editing software Audacity. I applied Audacity's noise reduction and normalization filters to recordings and individual audio files were saved for each item.

Scoring

The KPD utilizes both human scoring (for production tasks) and objective scoring (for perception tasks), as shown in Table 3.2. All items on the KPD are scored dichotomously. The objectively-scored perception items are scored based on an answer key that I verified through native speaker consultation and involvement in stimulus recording and through piloting with several native speakers (more details on piloting follow later in the chapter).

Table 3.2

KPD Scoring Overview

Task	Scoring Method	Scoring Target	Scores
Picture Naming	Human	each phoneme in a word	clear (1) or unclear (0)
Nonword Reading	Human	target phoneme in a nonword	clear (1) or unclear (0)
Pronunciation Judgment	Objective	pronunciation error in stimulus word	correct (1) or incorrect (0)
Nonword Recognition	Objective	target phoneme in a nonword	correct (1) or incorrect (0)

For the production items, scoring can be carried out by any proficient speaker of Korean, ideally with some linguistic training and familiarity with the test items. The ideal scorer (and administrator) of the test would be a learner's Korean teacher or tutor; Sundqvist et al. (2018) argue that teachers are well-positioned to evaluate learning-oriented, low-stakes tests due to their subject knowledge and ability to apply results to instructional activities. A simple scoring sheet is available (Appendix C), which can be used to cross out unclear phonemes while listening to test-takers' responses. In consultation with the Korean instructor who scored responses in the pilot and operational testing, I also developed a set of scoring criteria to guide scoring decisions (Appendix D). These criteria emphasize intelligibility of test-taker productions, i.e., making scoring based on unambiguity of phonemes while not demanding productions to sound native-like.

Score Reports

Score reports for the KPD have so far undergone three phases of development. The earlier versions are discussed in the latter half of this chapter. A sample KPD score report from the operational version of the test is in Figure 3.1, annotated with translations of major features of the report. The goal of the score report is to provide detailed, instructionally-relatable information on specific pronunciation weaknesses, such as particular phonemes or articulatory features that are not well mastered by a learner. The current version of the score report was guided by two principles: (1) guide score users' attention to an individual's most severe pronunciation targets, and (2) present detailed information that can be used as a springboard for subsequent learning.

The first page of the score report focuses on the first principle. The text at the top of the page provides a brief explanation of the score report and directs users to "focus on these difficult

sounds and features in class or when studying on your own." No numeric scores are given on the first page; only lists of difficult to pronounce phonemes, articulatory features, and word contexts alongside explanations. Wherever possible, non-technical vocabulary is used in the report. For instance, 소리 (*sound*) is used in lieu of 음소 (*phoneme*). Korean characters are used to exemplify features, making the concepts more accessible to learners who may not know the precise linguistic terms; this was a particular concern for Korean-language versions of a score report. Focusing on the second principle, the second page of the score report features detailed scores for every phoneme (cf. Kim, Pae, & Lee, 2005; Seok, Park, Shin, & Park, 2002) in production and perception, along with examples of the sound in real words, drawn from items the examinees did not respond to clearly/correctly on the KPD (Kunnan & Jang, 2009).

Development

The remainder of this chapter reviews this development history, highlighting changes to the test structure, tasks, and items as a result of piloting. This provides a record of development and early validation efforts related to the grounds and operationalization inferences in the KPD's validity argument. To date, the KPD has existed in three versions: an initial 'Alpha' version of the test, a heavily-revised 'Beta' version, and an 'Operational' version based on limited revisions of the Beta. Both the Alpha and Beta versions were piloted with Korean learners and native speakers; pilot test data and my own observations and content analyses of items informed the development of subsequent versions of the KPD.

Alpha Version

The initial version of the KPD, henceforth the KPD Alpha, was developed in the spring of 2017. It consisted of five tasks organized in three sections (Table 3.3). A major difference between this version and later versions is the presence of a Repetition task and a Sound and

Articulation Knowledge task (the latter comprising an additional section for Explicit Knowledge) and the lack of the Nonword Reading task. The other tasks (Picture Naming, Pronunciation Judgment, and Nonword Recognition) survived to subsequent versions of the test with only minor modifications and item revisions. All five tasks were delivered using the PsychoPy experimental software (Peirce, 2009); this mode of delivery was also changed in subsequent versions of the test.

The Repetition task involved listening to and repeating a short 1-2 syllable nonword and was scored in a manner virtually identical to the Nonword Reading task. Much like the Nonword Reading task which replaced it, I had designed the Repetition task to elicit phonemes in particular syllable positions without being subjected to potentially malformed or absent phonological representations of words in the lexicon.

The Sound and Articulation Knowledge task involved three-option multiple choice questions in English about the acoustic and articulatory qualities of Korean phonemes. My rationale for including this section and task was based on empirical findings supporting explicit phonetic teaching (e.g., Lord, 2005, 2008, 2010) and widespread pedagogical recommendations to teach articulations explicitly (e.g., by using diagrams of the articulatory apparatus, by teaching students to manually check physical sensations such as vibrating vocal chords and release of air, see Celce-Murcia et al., 2010). If a learner lacked explicit knowledge necessary to produce a sound, she could be taught it; if the learner had the explicit knowledge but could not accurately produce a sound, the instructional emphasis would likely be on perception and/or production practice without belaboring basic explanations of phonetic qualities and articulatory settings.

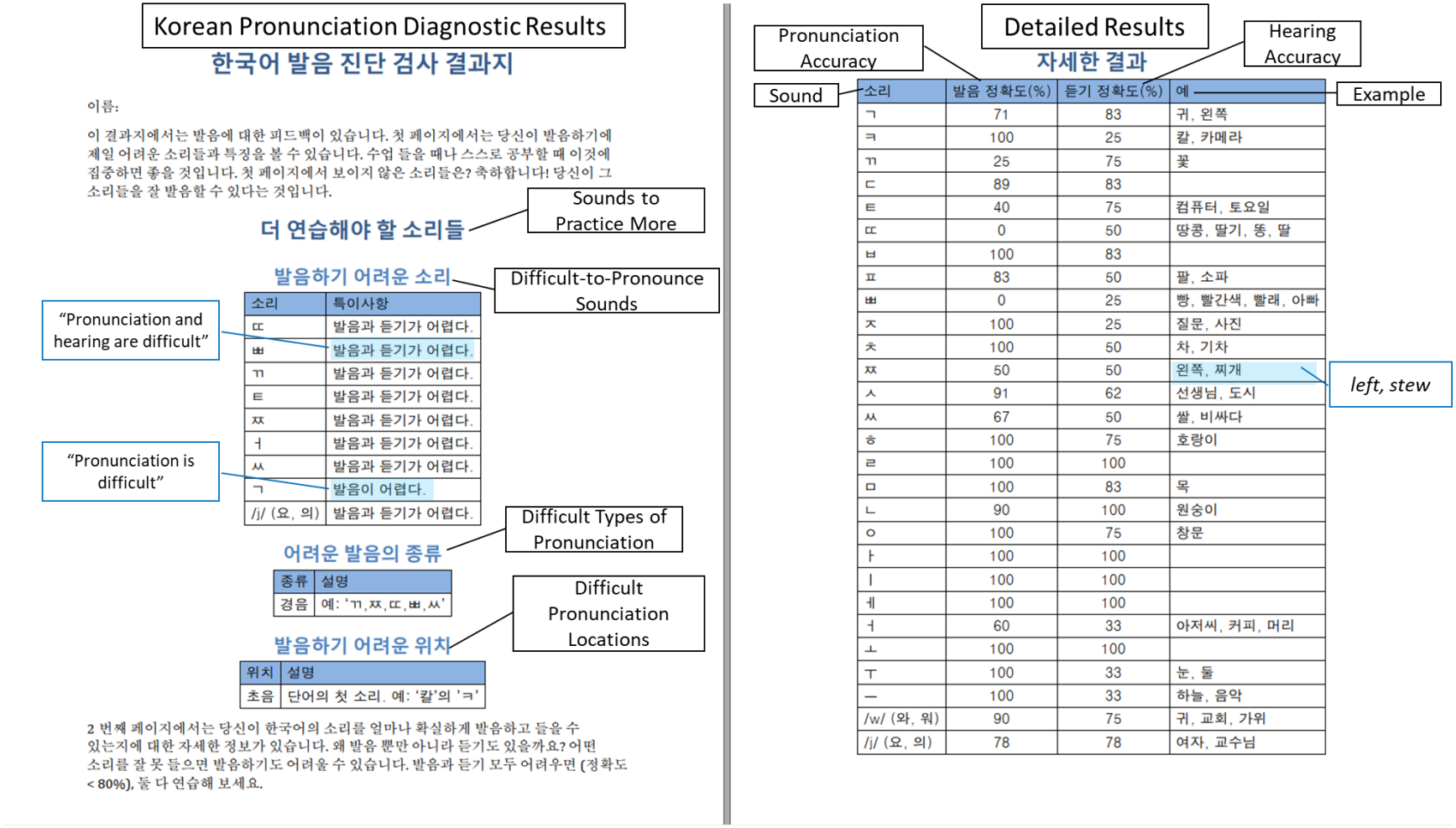


Figure 3.1. Diagram of a KPD score report. The first page of the score report is shown on the left, and the second page on the right.

Table 3.3

Initial KPD Design

Section	Tasks	Number of Items
Production	Picture Naming	35 words; 140 items
	Repetition	63
Perception	Pronunciation Judgment	72
	Nonword Identification	63
Explicit Knowledge	Sound and Articulation Knowledge	39

Piloting. This subsection details initial piloting of the KPD Alpha. The primary goal of this piloting was to investigate the alignment of test-taker processes and responses with what I had intended in task design, and to root out any undesirable task issues or item-level problems before proceeding to larger-scale piloting efforts. Essentially, I wanted to see if things generally worked (and what did not) and receive some guidance on initial item revisions. This kind of initial, small scale piloting is sometimes referred to as *prototyping* (see Nissen & Shedl, 2012), as the major aim is quickly finding major flaws before committing additional resources to test development and administration to larger numbers of (pilot) test-takers. The major consequence of this piloting was that two of the KPD Alpha tasks were eliminated and/or replaced after this pilot: The Repetition task, and the Sound and Articulation Knowledge task.

In the summer of 2017, I carried out this small-scale piloting of five initial tasks with four participants, including two L1 English learners of Korean and two Korean native speakers (Table 3.4). Each participant completed all five of the tasks, and after each task they completed a semi-structured interview with me. Figure 3.2 outlines the general procedures I followed for piloting and includes the list semi-structured interview questions.

Table 3.4

Alpha Pilot Participants

ID	L1	Sex	Notes
Alpha01	Korean	F	Graduate student in Second Language Studies. Expertise in psycholinguistics. Taught Korean as a foreign language for 1 year at an American university.
Alpha02	Korean	F	Graduate student in Second Language Studies. Expertise in language assessment. Taught Korean as a foreign language for 3 years at an American university.
Alpha03	English	F	Graduate student in Second Language Studies. Expertise in language assessment. Advanced speaker of French, novice in Korean. Basic Korean learned while teaching English in Korea for one year.
Alpha04	English	M	Undergraduate student in Information Technology, minoring in Korean. Previously learned Korean in the U.S. military. Has Korean spouse. Intermediate Korean proficiency.

1. KPD
   - 1.1 - Task 1: Picture Naming
   - 1.2 - Task 2: Repetition
   - 1.3 - Task 3: Pronunciation Judgment
   - 1.4 - Task 4: Identification
   - 1.5 - Task 5: Sound and Articulation Knowledge

After Each Task: Semi-Structured Interview
   - 1. Overall, what are your impressions of the task?
   - 2. What do you believe the task is about?
   - 3. Were the directions clear? Please explain any difficulties or confusion you encountered.
   - 4. Do you recall any particular questions that you felt were problematic? [*Participants shown items to stimulate comments*]
   - 5a. For Native Speakers: Did you generally feel confident in your answers?
     5b. For Learners: What were the easiest parts of the task? The hardest?
   - 6. Do you have any suggestions for improving the task?

2. Post-Test Interview
   - 2.1. Do you have any comments or questions about the test?

3. Background Questionnaire

Figure 3.2. KPD Alpha piloting procedures.

Findings and Revisions. I analyzed participants' performance on the KPD and their comments to guide my revisions to task design, individual items, and test administration. In terms of task design, I learned several important lessons from the alpha piloting. For tasks that

utilized picture stimuli (i.e., Picture Naming and Pronunciation Judgment), the selection of target words and selection/construction of the images are paramount. If a test-taker could not recognize the target word from the picture, I was unable to gain any information about their phonological abilities. Some pictures were simply not clear or obvious enough. Other pictures elicited multiple appropriate responses. For example, for an item intended to elicit *right* (as in *right hand,* 오른쪽 in Korean), I used a picture with an arrow pointing to the right. One NNS (Alpha 03) could not recall the word, one NNS immediately responded correctly (Alpha 04), and the two NSs instead responded with the Korean equivalent of *arrow sign/symbol* (화살표). Similarly, if a test taker did not focus on the targeted aspect or element of an image, they offered non-target responses. For example, one NNS, Alpha 04, offered 하트 (*heart*, a loanword) for an image intended to elicit 사랑 (*love*), focusing on the literal shape rather than what it commonly symbolizes. Thus, it became clear that images must be referents for commonly used and/or studied words, and the images should contain ample cues for word identification (e.g., shape, color, and additional symbols such as circles or 'X'). I also realized that the tester could provide some support, such as asking for another word, pointing to specific parts of an image, or giving clues, without compromising the intended response process (i.e., a test-taker recalls the phonological form of a target word and produces it).

For the repetition task, the post-task interviews with both native and non-native speakers revealed that perception ability played a major role in responses, even though I had originally intended the task to primarily measure production ability. I had hoped that an audio stimulus for short nonwords, rather than printed letters, would avoid an interference from shortcomings/mismatches related to learner orthographic and sound-symbol relationship knowledge. However, all respondents noted that items were easier to respond to if they felt they

had confidently identified what they heard in the stimulus. Thus, I eliminated the Repetition task and replaced it with a nonword reading task using the same stimuli in subsequent versions of the test.

Finally, I found that the Explicit Knowledge task had some insurmountable problems. On the positive side, my observations of the test-takers and their interview data showed that they were carefully thinking about and/or mouthing the articulations for target and distractor sounds, and so the task did appear to tap into explicit knowledge of articulatory and acoustic features. However, as one might expect, this section was difficult. Surprisingly, it was difficult for native speakers, too. My suspicion here is that because many of the Explicit Knowledge items relied on analogy to English sounds, the Korean NSs (who at the same time were English NNSs) were in a sense hampered. This is a considerable problem given that the target population of the KPD varies in L1; it seemed it would be inappropriate to give this task to those who do not speak (Standard American) English as an L1. Furthermore, one feature of items targeting articulations required test-takers to correctly choose descriptions of oral articulations. This worked well enough for articulations such as bilabial stops or lip-rounding for vowels, but items that required test-takers to identify what they were doing with their tongue and/or where the tongue was in the mouth were opaque. These items were also challenging for me to write while avoiding technical jargon (e.g., not using terms like *palate* or *alveolar ridge*). Thus, despite some evidence for the task and items functioning as intended and the potential usefulness of such information instructionally, I decided to remove this task from future versions of the KPD.

In terms of individual items, rather than salient task features or features affecting several items, the participants alerted me to several items with idiosyncratic problems. Primarily, these were vocabulary items in the Picture Naming and Pronunciation Judgment tasks. Some words

were simply unfamiliar to the NNSs, and so I have revised such items to target words that are either (a) higher frequency or (b) introduced earlier on in instructional materials (e.g., featured in the first level of the popular *Sogang Korean* or *Integrated Korean* textbooks). Other feedback pertained to audio quality and response options for some nonword items (Repetition and Identification tasks). When a native speaker felt an item had a stimulus and/or options that led to any lack of confidence in the response, I took that as a sign that something needed to be fixed. Through this process, I was able to identify several Identification items that would be re-recorded to ensure the clarity of keys.

Piloting the KPD Alpha also provided me an opportunity to begin developing the KPD score report. Figure 3.3 is an example of this early attempt. Compared to later versions, this initial score report was somewhat text-heavy, but contained many of the core elements that would reappear in subsequent revisions. For these initial score reports, I customized the text for each of the two learners who participated in the pilot. One major difference between this version of the score report and later versions is the provision of summary scores (total and task scores) at the top of the report.

Beta Version

The second iteration of the KPD, dubbed KPD Beta, was pared down to four tasks (Table 3.5). The administration of the tasks also underwent changes: The production tasks were administered via paper flipbooks (i.e., each item printed on its own 5.5 by 8.5 inch cardstock, with all items ring-bound in the top-left corner), and the receptive tasks were administered in OpenSesame (Mathôt, Schreij & Theeuwes, 2012), an experimental software similar to PsychoPy but with better (or at least more intuitive, to me) display settings. The Picture Naming and Pronunciation Judgment task stimuli were massively revamped. I opted for full-color

Alpha04

OVERALL SCORE: 87% (305 out of 351 attempted items)

By Skill and Tasks

PRONUNCIATION: 96% (170/177)

- Picture Naming: 98% (You attempted 114 out of 140 sounds; you got 112/114 correct)
- Repetition: 92% (58/63)

PERCEPTION: 80% (108/135)

- Pronunciation Judgment: 68% (49/72)
- Identification: 94% (59/63)

EXPLICIT KNOWLEDGE: 69% (27/39)

Overall feedback: You pronounce Korean sounds quite well- native speakers should generally have little trouble understanding you (based on sounds alone- vocab and grammar are another beast!). Your ability to hear and accurately distinguish Korean sounds is fairly good. The most difficult task for you was the pronunciation judgment- this may mean you might want to focus on getting more spoken vocabulary exposure to reinforce your knowledge of how words sound. Your explicit knowledge of Korean sounds is OK, but since you do quite well with pronunciation, this isn't a concern.

Additional Diagnostic Comments

By analyzing the mistakes you made, here are some specific areas for you to work on in the future:

Difficult Sounds:

- In general, consonants are more difficult for you than vowels. This is true for perception (hearing) and production.
- You have some trouble hearing ㅗ and ㅛ, and a little trouble producing them.
- You have some trouble producing and hearing ㄱ, ㄲ, and ㅋ.
- You have some trouble hearing ㅈ, ㅉ, and ㅊ.
- You have some trouble producing and hearing ㅅ and ㅆ.

Difficult Contexts:

- You have some difficulty producing and hearing consonants at the beginning of words, and some difficulty hearing consonants at the end of words.

Difficult Features:

- You have noticeable difficulty with tense sounds (ㄲ, ㄸ, ㅉ, ㅃ, ㅆ)
- Distinguishing back vowels is somewhat difficult for you (ㅏ, ㅓ, ㅗ, ㅜ, ㅣ).

Figure 3.3. Early draft of KPD score report.

pictures, drawing on open-source images previously used in psychological studies (MultiPic and BOSS), free full-color images (www.pixabay.com), and created or edited images manually as needed. In my image editing, I used various techniques to make target words more obvious, including arrows and circles to highlight key aspects and limited amounts of text to highlight a semantically-related word. For example, the new picture for 아저씨 (*ajeossi*, middle-aged man)

included a picture of a middle-aged woman labelled "아줌마" (*ajumma*, middle-aged woman) and a blank under the picture of the man.

Table 3.5

KPD Beta Design Summary

Section	Task	Brief Item Specification	Number of Items
Production	Picture Naming*	Item: picture of a concrete noun Response: speaking the matching word	154 (in 35 words)
	Nonword Read-Aloud	Item: 1-2 syllable nonword Response: reading aloud the nonword	63
Perception	Pronunciation Judgment*	Item: picture of a concrete noun + audio recording of the word Response: forced choice whether audio recording was (in)correct	72 (plus 40 filler items)
	Nonword Identification*	Item: audio recording of a 1-2 syllable nonword Response: forced choice between two written 1-2 syllable nonwords	63

Note. *Task was part of initial pilot test design.

Piloting. Compared to the KPD Alpha pilot, the focus of piloting the KPD Beta was more quantitative. I set out to collect data from a sample of learners that was just large enough for estimates of reliability and item statistics to be meaningful. I also collected data from a handful of Korean NSs. Having established tasks that generally worked (and after I had removed/addressed major flaws), I was more interested in fine-tuning at the item level. At the same time, this second piloting offered an opportunity to pilot other instruments that would be used in the main study, including a background questionnaire, pronunciation self-assessment, and an independent speaking task (see Chapter 4 for details). The piloting procedures for the KPD Beta are outlined in Figure 3.4.

The KPD Beta was piloted with 27 learners and 7 NSs of Korean recruited at Michigan State University. Of the learners, 25 were female and 2 were male. Learner breakdown by class level was as follows: 8 first year (KOR 102), 13 second year (KOR 202), 3 third year (KOR 302), 3 fourth year (KOR 402). For L1, 18 reported being English speakers, 5 were Chinese speakers, 2 spoke Malay, and there was 1 speaker each of Japanese and Thai.

PART 1 – Online

1. Send participant link to Qualtrics survey
2. Participant completes all parts of Qualtrics survey
    a. Informed Consent
    b. Part 1: Background
    c. Part 2: Korean Pronunciation Self-Assessment
        i. Global
        ii. Phoneme inventory (production and perception)

PART 2 – In-Person

1. Independent Speaking Task (2-3 minutes)
2. KPD (15-20 minutes)
    a. Production Tasks – Audio Recorded
        i. Picture Naming
        ii. Nonword Reading
    b. Perception Tasks – OpenSesame
        i. Pronunciation Judgment
        ii. Nonword Identification
3. Korean EIT (10 minutes) – Audio Recorded

Figure 3.4. KPD Beta piloting procedures.

Of the NSs, 6 were female and 1 was male. All NSs grew up primarily in South Korea and their dominant language was Korean, but all spoke English at a high level.

Findings. Rather than the learners' specific results and pronunciation strengths and weaknesses, I focus here on the technical qualities of the KPD itself, highlighting the specifics of test-taker responses where relevant. As a point of reference for the more detailed findings which follow, Table 3.6 presents summary statistics for the 27 learners who completed the KPD Beta.

Table 3.6

KPD Beta Learner Summary Statistics

Section	n	mean	SD	min	max	skew	kurt.
Production	217	193.74	14.17	153	154	-1.09	0.85
Task 1 – Picture Naming	154	142.7	7.50	124	153	-0.77	1.44
Task 2 – Nonword Reading	63	51.04	7.41	29	61	-1.28	1.34
Perception	135	99.56	14.00	71	124	-0.31	-0.94
Task 3 – Pronunciation Judgment*	72	41.33	11.63	18	61	-0.23	-0.84
Task 4 – Nonword Identification	63	58.22	3.29	50	63	-0.72	0.03

Developer Observations and Scorer Feedback. During piloting, I observed participants responding to items and took notes when I saw issues. I also took notes on any (unsolicited) verbal feedback participants gave on the tasks and items. At the task level, I noticed that the time of 1.0 seconds between initial stimulus presentation and audio in the receptive tasks seemed to be excessive. For the Nonword Reading task, I noticed an important error: I mistakenly included two additional items targeting the glide /j/ and failed to include any items targeting ㄲ /k*/. Other notes on individual items were as follows:

- T1_32-6 (ㅣ in 초콜릿): Appears to be substantial speaker variation; some NSs and learners use ㅖ instead of ㅣ. Excluding from analyses.
- T3_06 (팔): Picture would be clearer if it showed more of the upper body (to distinguish it from looking like a leg).
- T3_68 (예쁘다): The stimulus, “에쁘다”, is perhaps a slang/stylistic variation. Need to look into this.
- T3_16 (미국): The stimulus, “미국” (articulated with a /k^h/ in the coda), is not highly distinct and is also not a phonemic contrast. Consider changing to “미굿” or “미궁”.

- T3_50 (시장): The target phoneme is /i/, but the stimulus "쉬장" is not articulated to be distinct enough. "셔장" or "섀장"

Another source of feedback came from the Korean instructor who scored the production section of the KPD Beta. She found the noise from shuffling through the paper flipbooks present in the audio recordings to be a minor distraction. At the same time, I did notice that the flipbooks could occasionally be cumbersome.

Native Speaker Results. Due to the small number of examinees and the extremely high proportion of correct responses to most items, most conventional reliability and item analyses are not appropriate. Instead, the analyses of NS item responses focus solely on proportion of correct responses: A NS should generally be able to answer every item correctly, barring an occasional slip of the tongue or mishearing, and NS productions should otherwise be judged as acceptable.

The first key finding is that relatively few items—13 out of 366—had any incorrect responses from NSs. This provided general support for the notion the KPD Beta task designs, item specifications, individual items, and scoring procedures were working as intended: Speakers known to have robust Korean phonological systems (i.e., virtually all NSs) could successfully produce and perceive Korean segments according to KPD results. The items in Table 3.7, however, warranted extra scrutiny, because this desired success was not (totally) present. For 9 of the 13 potentially flawed items, there was only one NS incorrect response each. These incorrect responses may conceivably be attributable to accidental mis-presses on the keyboard (perception tasks) or slips of the tongue. In the case of the Picture Naming item, it may be an idiosyncratic scoring error rather than a speaker error. Nonetheless, I carefully reviewed these marginally problematic items when revising the KPD, focusing on stimulus clarity and distractor choices (as relevant). More pressing were items T3_01 (3 incorrect responses), T3_33 (7

incorrect), T3_44 (5 incorrect), and T3_50 (7 incorrect). These items required substantial revision and/or replacement.

Table 3.7

KPD Beta Items with Incorrect Responses from Korean NSs

Task	Item Code	Incorrect	Note
Task 1 – Picture Naming	T1_33-6	1/7	This is the ㄴ/n/ in 빨간색, it was substituted with an ㅇ/ŋ/ by one NS.
Task 2 – Nonword Reading	N/A	N/A	All items responded to correctly.
Task 3 – Pronunciation Judgment	OK_30	1/7	과일 /kwɑ.il/ *fruit*; filler item (pronounced correctly)
	T3_01	3/7	비 /pi/ *rain* pronounced as "피" /pʰi/
	T3_33	7/7	싸움 /s*a.um/ *fight* pronounced as "사움" /sa.um/
	T3_39	1/7	
	T3_41	1/7	하나 /ha.na/ *one* pronounced as "하마" /ha.ma/
	T3_44	5/7	창문 / tɕʰɑŋ.mun/ *window* pronounced as "찬문" /tɕʰɑn.mun/
	T3_50	7/7	시장 [ɕi.tɕɑŋ] *market* pronounced as "쉬장" [ɕwi.jɑɑŋ]
	T3_59	1/7	눈 /nun/ *eye* pronounced as "는" /n n/
	T3_68	1/7	예쁘다 /ye.p*ɯ.ta/ *pretty* pronounced as /e.p*ɯ.tɑ/
	T3_71	1/7	원 /wʌn/ *won (Korean currency)* pronounced as /wɑn/
Task 4 – Nonword Identification	T4_40	1/7	니 /ni/; distractor 미 /mi/
	T4_44	1/7	웅 /uŋ/; distractor 움 /um/

Task 1 – Picture Naming: Analysis of Non-Target Elicited Words. For the KPD Beta, Task 1 procedures were revised to allow for the tester to prompt test-takers when they provided a non-target word, up to and including modeling the word for the test-taker if it was completely unknown. While I deemed this accommodation necessary if the KPD were to be administered to

learners and L2 users across a reasonably wide range of general proficiency, I also had concerns about items that might consistently require extensive prompting and/or modeling: The flow of the task would be interrupted, and the overall time demand of the test would increase.

To investigate this new aspect of Task 1 procedures, I re-listened to all Task 1 audio recordings and logged each instance where a test-taker's first response to an item was off target. I logged what alternative(s) they provided and whether they ultimately required the tester (i.e., me) to model the word for them. Table 3.8 shows a summary of this analysis.

Table 3.8

Summary of KPD Beta Task 1 – Picture Naming Non-Target Responses

Group	N	Number of Non-Target Initial Responses (proportion*)	Number of Tester Models Supplied (proportion)
All	34	309 (26%)	204 (17%)
NSs	7	18 (7%)	0 (0%)
Learners	27	291 (31%)	204 (22%)

Note. *Proportion computed based on the total number of items administered to each (sub)group (35 items × N test-takers).

Focusing on specific items, there were only 8 words (out of 35) that elicited non-target responses from NSs (Table 3.9). The most frequently unclear items were 빵 (*bread*), 포도 (*grape*), and 돈 (*money*). The non-target alternatives provided for *bread* and *money* were more specific terms, while the alternatives provided for *grape* indicated some lack of clarity in the picture; it did not appear that most of the NSs could distinguish the picture as grapes and not some other similar-looking fruit.

Table 3.9

KPD Beta Task 1 Items which Elicited Non-Target NS Responses

Item	Translation	Freq.	Alternatives Provided
T1_1 빵	bread	5	바게트 (baguette)
T1_16 포도	grape	5	열매 (berry), 가지 (eggplant), 블루베리 (blueberry)
T1_17 돈	money	3	지폐 (bill), 화폐 (bill), 현금 (cash)
T1_11 택시	taxi	1	자동차 (car)
T1_24 그림	picture	1	액자 (picture frame)
T1_30 왼쪽	left (side/direction)	1	[mumbling]
T1_4 나비	butterfly	1	나바 (cf. 나방, moth)
T1_9 집	house	1	주택 (house/dwelling)

Table 3.10 lists which items most frequently elicited non-target responses from learners and those which most frequently required modeling by the tester (me). In total, 31 out of 35 items initially elicited a non-target word or no response by at least one Korean learner. I took these data with a grain of salt, given that much of the pilot learner sample was on the lower end of Korean proficiency due to having relatively minimal exposure to the language (e.g., second year students had only had roughly 150 hours of classroom instruction when they took the KPD). Like the NSs, the images for *grape* and *money* were somewhat ambiguous to the learners. Looking at the non-target words supplied, compared to NSs the learners often substituted more general terms or hypernyms. For example, the Korean word for "fruit" was given for the pictures of grapes and lemon, and the Korean word for "man" was given for a picture of a middle-aged man (n.b., the Korean word for middle-aged man is extremely commonly used). Learners also attempted to supply loanwords or words from other languages, such as the Japanese *tori* for Korean 새 (bird). In other instances, phonological word forms were inaccurately recalled.

Table 3.10

KPD Beta Task 1 Items with Frequent Non-Target Learner Responses

Item	Eng.	Freq.	Model Freq.	Prompting Success	Alternatives Provided
T1_14 땅콩	peanut	24	22	2/24	당근 (carrot), 돈 (money), 상추 (lettuce)
T1_25 용	dragon	23	18	5/23	룡 (similar to Chinese), 량 (amount), 공룡 (dinosaur), “dragon” (English)
T1_3 원숭이	monkey	21	21	0/21	마리 (animal counter), 동물 (animal)
T1_26 침대	bed	18	17	1/18	잠대 (sleep + second half of target word), 베드 (English “bed” in Korean pronunciation), “bed” (English)
T1_4 나비	butterfly	18	17	1/18	빠삐용 (Korean approximation of French for “butterfly”), 냄비 (cooking pot), 비자 (visa)
T1_16 포도	grape	17	13	4/17	과일 (fruit), 폼 (?), 블루… (blue…), “grapes” (English)
T1_11 택시	taxi	13	1	12/13	자동차/차 (car), 기겐샤 (Korean approximation of a Japanese word?)
T1_17 돈	money	13	5	8/13	원 (won, the Korean currency unit), 현금 (cash), 현킨 (malformed 현금/cash), 천원 (1,000 Korean won)
T1_28 왕	king	13	13	0/13	왕자 (prince), “king” (English)
T1_24 그림	picture	12	7	5/12	사진 (photograph), 꽃 (flower), 종이 (paper), “art” “painting” (English)
T1_10 새	bird	11	9	2/11	아가 (baby), 샘 (?), 토리 (Japanese), 파란색 (blue), 가새 (? + bird)
T1_23 의자	chair	10	7	3/10	자리 (seat), 자기 (oneself), 의사 (doctor)
T1_27 쓰레기	trash	9	6	3/9	휴지통 (wastebasket), 휴게통 (malformed 휴지통), 레서핑 (?), 나비스탄 (?)

Table 3.10 (cont'd)

T1_8 아저씨	middle-aged man	9	2	7/9	남자 (man), 아버..(beginning of "father"), 할아버지 (grandfather)
T1_18 레몬	lemon	8	0	8/8	과일 (fruit)
T1_19 시계	clock	8	8	0/8	시간 (hour), 시름 (?), "clock" (English)
T1_5 토끼	rabbit	8	8	0/8	또자 (?), 토자(?), "rabbit" (English)
T1_7 돼지	pig	8	7	1/8	뒤기 (malformed 되지)
T1_30 왼쪽	left	7	0	7/7	오른쪽 (right), 오른… (beginning of "right"), 왼쪽에 (left + to/on)
T1_1 빵	bread	5	1	4/5	밤 (chestnut; possible mispronunciation of target), 음식 (food), "bread" (English)
T1_13 귀	ear	5	5	0/5	이 (tooth), 얇.. (part of idiom "귀가 얇다", meaning gullible)
T1_31 불	fire	5	4	1/5	화 (Sino-Korean root meaning "fire")
T1_22 맥주	beer	4	1	4/5	물 (water), 술 (alcohol), 소주 (Korean traditional alcohol), 비어 (Korean pronunciation of loanword "beer"), "beer" (English)
T1_33 빨간색	red	4	4	0/4	none
T1_34 꽃	flower	4	4	0/4	"flower" (English)

Note. Items responded to with non-target words fewer than 4 times excluded from table.

For these items where non-target words were initially elicited, I was also interested in seeing where I was able to prompt learners to eventually provide the target word. This varied greatly. For words like *monkey*, which was not initially provided by 21 out of 28 learners, it seemed that they all were just unfamiliar with the word in Korean, and I had to provide a model to each of them. However, for words like *left*, I was able to successfully prompt all 7 learners who initially supplied something else (most commonly *right*). In general, I took away from this analysis that several pictures would need revising in order to minimize non-target responses and

modeling, yet at the same time I accepted that to some degree prompting and modeling may be necessary, particularly when administering the KPD to learners with limited Korean experience.

Reliability. I examined reliability of the KPD for the 27 learners by computing Cronbach's alpha. Two types of scoring models were explored: individual items and item parcels. For the individual items approach, I entered each item separately into reliability and item analyses. I carried these analyses out at the Task level (i.e., separately for Task 1, Task 2, etc.) and at the Mode level (i.e., Task1 & Task 2, Task 3 & Task 4). For the item parcels approach, I computed total scores across each phoneme in each mode, collapsing the several items corresponding to a phoneme into a single polytomous item (e.g., a sum score for all items targeting ㄱ /k/ in Task 1 and Task 2). Results of these reliability analyses are in Table 3.11. Generally, reliability results were within desirable ranges, and item parceling led to minimal degradation of reliability despite collapsing 100+ items into just 28. In sum, the test items (or item parcels) appeared to be strongly interrelated.

Table 3.11

Reliability of the KPD Beta

Section	n	Cronbach's alpha (individual items)	Cronbach's alpha (item parcels, n = 28)
Production	217	.92	.87
Task 1 – Picture Naming	154	.85	
Task 2 – Nonword Reading	63	.86	
Perception	135	.92	.91
Task 3 – Pronunciation Judgment*	72	.91	
Task 4 – Nonword Identification	63	.65	

*Excluding filler items.

Item Statistics. As another means of investigating the performance of individual items, I ran classical test theory (CTT) item analyses on the set of learner test data, separately for production and perception items, which yielded discrimination (D) and facility (P) statistics for

each individual KPD item. The diagnostic decisions made on KPD data are technically based on a cut score—actually, cut scores for parcels of items—which made criterion-referenced item statistics (i.e., the B index and facility differences between masters and non-masters) more appropriate. However, due to still being in early stages of developing an appropriate measurement model and framework for interpreting scores, I opted to go with the CTT analyses, which still gave a reasonably informative indicator of how well participants with generally more accurate pronunciation did on the items and how easy the items were overall. Additionally, I expected that items would have very high facility values. For example, items targeting ㅏ /ɑ/, a phoneme cross-linguistically common to many learner L1s, were expected to be rather easy. Thus, typical interpretations of CTT item analyses for norm-referenced tests (e.g., desirable values are between .25 and .75) were ignored. More weight was given to discrimination. In typical norm-referenced test contexts, discrimination values above .3 are desired (Carr, 2011), but I took a more liberal approach in line with my expectations that some items would be very easy (i.e., have high facility and thus poorly differentiate learners with stronger and weaker pronunciation or perception): I flagged items with negative discrimination (Table 3.12). Negative discrimination indicated that learners with generally more accurate production (or perception) tended to do poorly on the item. At the same time, given the small sample, a small number of people at the higher end of the total score range with similar pronunciation difficulties (e.g., great difficulties with phonemes predicted to be difficult, such as ㄹ /l/) could skew discrimination indices. Thus, I looked for larger negative discrimination values alongside facility values, and I considered the content of items.

Table 3.12

KPD Beta Items Flagged for Potential Revision

Item	D	P	Notes
Task 1 – Picture Naming			
T1_1-1	-.14	.70	ㅃ /p*/ in 빵
T1_12-1	-.17	.96	ㅌ /tʰ/ in 택시
T1_18-2	-.22	.96	ㅘ /wɑ/ in 화장실
T1_30-4	-.22	.96	ㅉ /tɕ*/ in 왼쪽
T1_32-5	-.05	.67	ㄹ /l/ (geminate) in 초콜릿
Task 2 – Nonword Reading			
T6_05	-.16	.78	ㅃ /p*/
T6_30	-.18	.74	ㅆ /s*/
Task 3 – Pronunciation Judgment			
T3_18	-.12	.15	(ㄱ)ㄲ /k*/ in 꿀
T3_34	-.07	.11	(ㅅ)ㅆ /s*/ in 접시
T3_45	-.19	.74	(ㄴ)ㄹ /l/ in 라디오
T3_68	-.21	.93	(ㅔ)ㅖ /je/ in 예쁘다
Task 4 – Nonword Identification			
T4_24	-.24	.85	쪼 /tɕ*/ (조)
T4_51	-.28	.96	으 /ɯ/ (우)

Note. For Tasks 3 and 4, distractors are indicated in parenthesis.

Many of the items with larger discrimination and/or lower facility targeted tensed phonemes, which was not unexpected given their cross-linguistic rarity, high degree of similarity with other Korean sounds (i.e., articulation differs with a lax phoneme only in tenseness), and previous empirical findings (e.g., Moon et al., 2009). Similarly, the phoneme /l/ (ㄹ) was flagged in one item. Other items involved English-origin loanwords. This may have been due to learners mixing the Korean phonological form with the one present in their native language.

Score Reporting. Each of the 27 learners in the second (Beta) pilot study received an individual score report (Figure 3.5). The reports were composed in English and consisted of two pages. The first page summarized their KPD results, highlighting phonemes that were deemed difficult to produce based on an arbitrary cutoff of 80% accuracy in production. The first page also included information on features (e.g., tenseness) and word contexts (e.g., word-initial) that

presented difficulty for learners, using the same 80% cutoff. Notably, the first page has no numeric scores. My intention was to require score report users to *read* the feedback instead of zero in on any total or part scores (see Alderson et al., 2015, pp. 188-192, for discussion of learners preferring traditional total scores and paying less attention to diagnostic feedback). The second page provided detailed information on learners' accuracy for each of the 28 phonemes in production and perception. It also included stimuli from items on which they made mistakes. My intentions here were to make the results more memorable ("ah, 왼쪽, I always mispronounce the ㅉ") and to provide some initial material for instruction. A learner could try recording the missed production items and ask his teacher to give feedback, or a teacher could provide dictation exercises based on the missed perception (and production) items.

Revisions Leading to Operational Version

Broad, task level revisions for the KPD Operational Version were few in number and relatively minor. The production tasks were converted to PowerPoint presentations that could be smoothly clicked through on a computer (although using flipbooks would still have been acceptable). For the Pronunciation Judgment task, the time between initial presentation of the image and start of the audio was reduced from 1.0 seconds to 0.5 seconds. Based on the previously presented Beta pilot findings and careful review of item content, I made the following changes to items:

Task 1 – Picture Naming

- T1_4 나비 (butterfly): The coloring of the image was manually altered to those of the iconic Monarch butterfly to avoid the non-target *moth*
- T1_16 포도 (grape): The original image only showed a single grape. I produced an image with a cluster of grapes.

Your Korean Pronunciation

Name: [redacted]

This report provides feedback on your pronunciation and hearing of Korean sounds. On the first page, your biggest pronunciation challenges are highlighted. These are things you should focus on when you are in class, studying by yourself, meeting with a conversation partner/tutor, or at your teacher's office hours. Anything that's not on this first page? Consider that good news! You seem to be able to intelligibly pronounce the other sounds most of the time.

On the second page, you will find complete information on how accurately you pronounce and hear Korean sounds. Why pronunciation *and* hearing? Well, there's generally a connection between how well you can hear a sound and how accurately and consistently you can pronounce it. So, if you have trouble pronouncing AND hearing a sound (< 80%), you'll want to work on both! If your pronunciation is accurate (> 80%), don't worry too much about the hearing. Note that hearing scores will tend to be lower, as the hearing tasks were more difficult.

Challenging Sounds: Standard Korean has 28 sounds, including 19 consonants, 7 vowels, and 2 glides (*glides* are something between vowels and consonants; in Korean these are the first sounds in 와 and 요).

- **ㄷ:** You have noticeable pronunciation difficulty.
- **ㅂ:** You have noticeable pronunciation difficulty and some hearing difficulty.
- **ㅉ:** You have some pronunciation difficulty and noticeable hearing difficulty.
- **ㅆ:** You have some pronunciation difficulty and some hearing difficulty.
- **ㄹ:** You have noticeable pronunciation difficulty.
- **ㅇ:** You have some pronunciation difficulty.
- **/w/ (e.g., 외):** You have some pronunciation difficulty.

Challenging Features: *Features* are ways of making sounds with language, and most sounds combine multiple features. Several sounds can share a feature. For example, the *tense* feature in Korean is shared by ㄲ, ㅉ, ㄸ, ㅃ, and ㅆ.

- **Sonorants:** This feature is found in ㅁ, ㄴ, ㅇ, and ㄹ. Sonorants are always voiced, but voicing is weaker at the beginning of words, making these sounds tricky.

Challenging Contexts: In language, a single sound can vary depending on the surrounding sounds, and some situations might make a sound harder to pronounce or hear.

- **Coda (end of word):** Some consonants change sound slightly at the end of words. It can be hard to hear the difference between some sounds in the coda position.

Korean Sound Profile

This chart tells you how accurately you pronounce and hear Korean sounds. For pronunciation, anything above 80% is considered sufficiently accurate. The 'Example Words/Sounds' columns show questions that you missed on the test. You can use these words/sounds to study!

		Pronunciation		Hearing	
	Sound	Accuracy (%)	Example Words/Sounds	Accuracy (%)	Example Words/Sounds
Consonants	ㄱ	86	왼쪽, 옥	83	미국
	ㅋ	83	키	75	칼
	ㄲ	100		75	꿀
	ㄷ	56	초콜릿, 꽃, 도, 알	100	
	ㅌ	100		100	
	ㄸ	100		50	똥, 딸
	ㅂ	57	불, 보, 압	67	비, 바
	ㅍ	100		100	
	ㅃ	100		75	빨래
	ㅈ	88	자	100	
	ㅊ	100		75	차
	ㅉ	75	맥주	50	찌개, 쪼
	ㅅ	100		75	소, 선생님
	ㅆ	71	택시, 우쑤	67	싸움, 접시
	ㅎ	100		75	호랑이
	ㄹ	42	화장실, 딸기, 그림, 쓰레기, 불, 빨간색, 이리	88	별
	ㅁ	100		83	사람
	ㄴ	90	원숭이	100	
	ㅇ	67	빵, 왕, 원숭이	100	
Vowels	ㅏ	100		67	산
	ㅣ	94	피아노	67	시장
	ㅔ/ㅐ	100		100	
	ㅓ	100		67	커피
	ㅗ	100		33	손, 호주
	ㅜ	100		100	
	ㅡ	100		100	
Glides	/w/ (와, 워)	50	귀, 왼쪽, 돼지, 위, 워	88	가위
	/j/ (요, 의)	100		100	

Figure 3.5. KPD Beta score report.

- T1_17 돈 (money): The original image included only paper money/bills, and some non-target responses reflected this. I replaced this with an image including both paper bills and coins, aiming to elicit the more general *money*.
- T1_25 용 (dragon): Upon careful inspection of the non-target responses, I noticed that several non-Western participants had difficulties coming up with the right word. I added an image of a dragon from East Asian cultures to make this item more cross-culturally effective.
- T1_29 레몬 (lemon): I replaced this item with 라면 (*ramyeon*, Korean ramen noodles).
- T1_32 초콜릿 (chocolate): Although the National Institute of the Korean Language (2015) maintains that the penultimate phoneme is /i/, I decided not to score (i.e., ignore/delete from specifications) the /i/ in the last syllable due to substantial NS and learner variation.

Task 2 – Nonword Reading

- Two /j/ glide items (T2_57 and T2_59) were replaced with items targeting /k*/: 까 (/k*ɑ/) and 이끼 (/i.k*i/).

Task 3 – Pronunciation Judgment

- OK_30 과일 (fruit, filler item), T3_01 비 (rain), T3_39 사람 (person), T3_41 하나 (one), T3_71 원 (Korean *won* currency): Re-recorded
- T3_33 싸움 (a fight): Changed to 비싸다 (/pi.s*ɑ.tɑ/, *expensive*), with the audio as 비사다 (/pi.sɑ.tɑ/)
- T3_44 창문 (window): Changed audio from 찬문 (/tɕʰɑn.mun/) to 차문 (/tɕʰɑ.mun/)

- T3_50 시장 (market): Changed audio from 쉬장 (ɕwi.ʨɑŋ) to 섀장 (/ɕje.ʨɑŋ/)
- T3_68 예쁘다 (pretty): Changed audio from 에쁘다 (/e.p* ɯ.tɑ/) to 왜쁘다 (we.p* ɯ.tɑ)

Task 4 – Sound Identification

- T4_40 니 (/ni/), T4_44 웅 (/uŋ/): Re-recorded

Conclusion

This chapter documented the design and development of the KPD, highlighting the linguistic and psycholinguistic bases for the design of the test as well as incremental efforts to better represent the underlying constructs and reduce sources of irrelevant variance in test-taker performance. This documentation will be revisited in Chapter 9, where evidence for the validity of the KPD is considered alongside the proposed validity argument from Chapter 1.

CHAPTER 4: METHODS

This dissertation is a test development project. In test development, developers typically go through several stages, beginning with setting a purpose for developing a test and ultimately producing an operational form of the test with supporting documentation (Irwing & Hughes, 2018). Previous chapters have detailed several of the early stages, including defining the test purpose and developing items. In this chapter, I outline the methods I used to carry out the validation stage of test development, that is, collecting evidence that relevant to the inferences and assumptions of the KPD's validity argument.

I adopted a mixed-methods research design for the validation stage of test development. Specifically, I used a mixed-methods design that is closest to a *convergent parallel design* in Creswell and Plano Clark's (2011) widely-used typology. I collected both quantitative and qualitative data at roughly the same time, with the KPD validity argument as the nexus for integrating sources of information and making interpretations. The quantitative component involves the collection of field test data and other relevant measures from a large sample of L2 Korean learners. The qualitative component entails interviews with L2 Korean learners and a teacher of two of those learners. These two components complement one another primarily by providing evidence relevant to different inferences or assumptions in the KPD's validity argument. In language testing, interviews are commonly used to explore, in some detail, stakeholder test score interpretations (e.g., Dimova & Kling, 2018) and interfaces between tests and teaching and learning (e.g., Allen, 2016; Tan & Turner, 2015).

This study makes uses of instruments with Korean-English bilingual directions, with Korean being the target language for participants and English being a global lingua franca which could support participants at earlier stages of Korean learning. Interviews utilized Korean and/or

English, with either language being used to various degree to support meaning-making and mutual understanding between interviewer and interviewees.

Participants

For the quantitative component of the study, which I refer to as *field testing*, I collected KPD test data from a large number of adult Korean language learners in Seoul, South Korea. I also collected data from a small number of Korean NSs.

For the qualitative component of the study, which I refer to as the *interview study*, I interviewed a subset of 21 learners from the field testing sample. In addition, I interviewed one Korean instructor who had taught two of these learners.

Field Testing

For field testing of the KPD, a large sample of Korean learners and a small number of Korean NSs participated.

Learners. In total, 198 learners of Korean participated in the field testing of the KPD (Table 4.1). A large majority (174) were female. A total of 24 L1s and 36 nationalities were represented in the sample. A plurality of these learners were L1 Mandarin speakers from Mandarin-dominant countries (i.e., China and Taiwan). Most learners were affiliated with Korean universities in some way, as intensive program language students, undergraduate, or graduate students. A small number were currently working in Korea in various capacities (e.g., embassy staff, English teacher).

Table 4.1

Field Testing Sample Characteristics: Demographic Categories

Category	n	Category	n
Gender			
Male	24		
Female	174		
Nationality	n	Circumstances in Korea	
China	59	Language Student***	79
Taiwan	30	Level 1 (Lowest)	5
Japan	14	Level 2	28
USA	14	Level 3	12
Russia	9	Level 4	21
Vietnam	7	Level 5	11
Hong Kong	6	Level 6 (Highest)	1
Kazakhstan	6	Other/Specialized Program	1
France	5	Undergraduate	39
Malaysia	5	Graduate Student	63
Other* (less than 5 per country)	43	Other (not a student)	17
First (Most Dominant) Language		Korean as a *j*th Language (median)	3
Chinese – Mandarin	88	1st	1
English	19	2nd	34
Russian	19	3rd	119
Japanese	13	4th	30
Spanish	11	5th or later	13
Chinese – Cantonese	8	NA	1
Vietnamese	7		
French	5		
Other** (less than 5 per language)	28		

Note. *Includes Azerbaijan, Bangladesh, Belarus, Bermuda, Brazil, Chile, Columbia, Ecuador, El Salvador, Germany, Indonesia, Iran, Italy, Kyrgyzstan, Mexico, Mongolia, Peru, Philippines, Singapore, Spain, Thailand, Turkey, Turkmenistan, Sri Lanka, Ukraine, and Uzbekistan.
**Includes Azerbaijani, Bangla, German, Indonesian (Bahasa), Italian, Kazakh, Kyrgyz, Mongol, Malay (Bahasa Malay), Persian (Urdu), Portuguese, Tagalog, Turkish, Taiwanese, and Sinhala.
*** "Language Student" refers to learners enrolled in a university-affiliated intensive Korean program. Throughout Korea, instruction in these institutes is almost universally divided into six levels, with 1 being appropriate for (true) beginners and 6 designed for learners at/approaching advanced levels of overall Korean proficiency.

The average age of participants was 24.17 years (median = 23 years, Table 4.2). The average participant began learning Korean at roughly the age of 19 years old, and most participants were learning it as their third or later language. On average, participants had spent a total of roughly one to one and a half years in Korea but varied considerably in their total time spent in-country. Of that time, approximately six months to one year was spent in in-country language study on average, but again, there was considerable variation (SD = 14.77 months). Outside of Korea, most likely in their home countries, participants had spent one to one and a half years studying Korean as a foreign language, yet again there was considerable variation (SD = 24.22 months).

Table 4.2

Field Testing Sample Characteristics: Age and Exposure

	n	M	SD	Median	Min	Max
Age (years)	198	24.17	4.46	23	19	48
Age of Onset (years)	198	19.35	4.79	19	0	39
Time Living in South Korea (months)	198	17.76	19.72	12	0	130
Time Living with a Korean-Speaking Family (months)	196	9.67	44.63	0	0	360
Time Studying Korean in South Korea (months)	198	11.01	14.49	6.5	0	130
Time Studying Korean as a Foreign Language (months)	198	17.31	24.22	12	0	216
Total Korean Study Time (months)	198	28.33	30.63	22.5	0	296

Participants self-reported their Korean proficiency in two ways: self-assessment of the four macroskills (speaking, listening, writing, and reading) and self-report of proficiency test results (Table 4.3). The self-assessment was based on a scale of 0 (“none”) to 10 (“perfect”), with each point having a simple descriptor (e.g., 5 = adequate). The means and median self-ratings for productive scales were roughly 5, and receptive skills were roughly 6.

Table 4.3

Self-Assessment of Macroskills

Skill	n	Mean	SD	Median	Min	Max
Speaking	198	5.01	1.94	5	1	10
Listening	198	5.82	2.03	6	1	10
Writing	198	4.84	1.84	5	1	10
Reading	198	5.80	2.06	6	1	10

For self-reported proficiency test results, a majority of participants reported Test of Proficiency in Korean (TOPIK) results (n = 140) as their most recent standardized proficiency test; the only other standardized test reported was the ACTFL Oral Proficiency Interview (n = 2; one participant reported a score of Novice Low and another reported a score of Intermediate High). The TOPIK exam has two levels, with a lower-level form (TOPIK I) that yields results in major bands 1 and 2, and a higher-level form (TOPIK II) that yields results in major bands of 3 to 6 (www.topik.go.kr). One-hundred twenty-nine participants reported results from the TOPIK II. The average TOPIK band score reported was 4.25 (SD = 1.09), with a median of 4.

Participants also reported on the contribution of extracurricular activities to their Korean learning, their current level of Korean use for common activities, and their motivations to learn Korean (Table 4.4). Relatively few participants reported having any family members who spoke Korean, explaining the low number of responses to questions about interacting with family in the first two parts of Table 4.4. However, as motivation may be more future-oriented or aspirational, most participants did respond to the motivation question about family. In general, participants had relatively high engagement in a variety of extracurricular activities. Motivation-wise, instrumental goals such as getting a job or going to university were of similar importance as integrative goals such as having friendships with Koreans or appreciating Korean culture.

Variation in responses to these questions was rather large, highlighting the diversity of participant learning practices, current Korean use, and their strong motivations for learning.

Table 4.4

Korean Learning, Use, and Motivation

	n	M	SD	Median	Min	Max
Contribution to Learning Korean* by…						
Interacting with Friends	198	7.16	2.61	8	0	10
Interacting with Family	198	0.76	2.02	0	0	10
Reading	198	5.95	2.35	6	0	10
Self-Study	198	6.76	1.93	7	0	10
Watching TV or Movies	198	6.77	2.31	7	0	10
Listening to Music	198	5.17	2.85	5	0	10
Level of Current Korean Use** when…						
Interacting with Friends	198	6.11	2.44	6	0	10
Interacting with Family	198	0.53	1.78	0	0	10
Reading	198	5.74	2.54	6	0	10
Self-Study	198	6.74	2.30	7	0	10
Watching TV or Movies	198	6.22	2.60	7	0	10
Listening to Music	198	5.69	2.98	6	0	10
Motivation for Learning Korean* due to…						
Getting a Job	193	6.61	3.12	8	0	10
Earning More Money	194	5.81	3.25	6	0	10
Going to University or Other Training	198	6.76	3.31	8	0	10
Impressing Friends and Family	198	4.09	3.06	5	0	10
Korean-Speaking Family	198	1.46	2.52	0	0	10
Korean-Speaking Spouse or Partner	198	2.75	3.35	1	0	10
Friendship with Koreans	188	6.27	2.76	6.5	0	10
Korean Culture	188	6.60	2.45	7	0	10

Note. *Scale: 0 = not at all, 1 = minimally, 5 = moderately, 10 = most importantly. **Scale: 0 = none, 1 = almost never, 5 = 50% of the time, 10 = always.

Native Speakers. In total, 6 Korean NSs completed field testing procedures. NS participants were recruited from the Seoul area, and all were connected to universities in some way (3 undergraduate students, 3 graduate students). Of the 6 NSs, 5 were female. Their average age was 23.5 years old (median = 23, min = 19, max = 31). All NS participants reported English

as their second most-dominant language; participants additionally reported lower levels of proficiency in Japanese (n = 3), French (n = 1), and Spanish (n = 1). On average, participants reported using Korean 76.5% of the time (median = 75%, min = 60%, max = 97%).

KPD Production Task Scoring Reliability Study

Six Korean NSs (female = 5), all enrolled in or recent graduates of a master's degree program in teaching Korean as a second/foreign language, participated in the scoring reliability study. These participants were not the same individuals as the previously described NSs who participated in the field testing. Participants varied in their teaching experience; at one end a participant had only minimal experience tutoring while on the other end another participant had been teaching Korean classes for immigrants at a cultural center for one year.

I gave all participants an introduction to the test and training on how to score the production tasks. All information was given in Korean. For each of the two tasks, training included examples of scoring (i.e., listen to real responses and see what score was given), detailed explanation of scoring criteria, and a selection of items from different test-takers to practice score (i.e., isolated items and responses) with feedback. Then, participants scored the entire production section for one sample examinee. After completing scoring for the sample examinee, they were given a copy of the scores given by the expert rater who scored all of the test-takers who completed the KPD in field testing. Participants had the opportunity to ask questions throughout the training.

After the introduction and training, all participants scored a subset of 20 randomly selected KPD tests from field testing. The 20 tests were a stratified random sample; two NSs KPD tests and 18 learner KPD tests were randomly selected to compose the subset used in the rater reliability study.

Interview Study

A total of 22 participants took part in the interview study. Among these 22, 21 were L2 learners of Korean, all of whom had completed the field testing procedures before their initial interview, and one was a teacher of Korean. Five of the L2 learners were graduate students, four were undergraduate students, and 12 were language students (i.e., currently studying in an intensive 20-hours-per-week Korean language program housed at a university). Learner interviewee L1 backgrounds included Chinese (Mandarin), Cantonese, French, German, Japanese, Russian, Spanish, and Vietnamese. More details on these participants can be found in Chapter 8.

I invited these learners to participate in the interview study primarily on the basis of representativeness and having potentially interesting perspectives on L2 Korean pronunciation. I looked for individuals representing a range of interesting KPD score profiles (different weaknesses, having relatively many or few weaknesses) and those who made interesting comments when chatting before/during/after their field testing appointment; I made brief notes about small talk with participants about jobs, learning experiences, interest or struggles in pronunciation, etc., during field testing appointments. I also considered learners' backgrounds (current circumstances in Korea, linguistic background, and Korean proficiency level), as I believed having diverse perspectives is important (Friedman, 2012). On a more practical level, I considered potential interviewees' linguistic ability to participate in an interview (i.e., sufficient Korean or English proficiency to understand and respond to open-ended interview questions). With just two exceptions, all participants who I invited to participate in the interview study accepted (one simply had no interest, and another cancelled her appointment due to illness and could not reschedule before departing Korea). Learner interviews took place in Korean and/or

English depending on the interviewee's and my own linguistic capabilities; most interviews were conducted entirely in Korean with minimal code-switching to English.

The Korean instructor who participated in an interview had taught two of the language students who participated in the interview study in a university intensive Korean language program. I recruited this teacher based on my personal network: He was one of my teachers in an intensive Korean course I took before starting data collection. Through informal observation of his teaching and informal chats about L2 research, I thought he would be interested in participating in the study. The interview with the Korean instructor was conducted in Korean with minimal English code-switching.

Materials

In addition to the KPD, described in detail in the previous chapter, the following instruments and materials were used.

Language Background Questionnaire

The language background questionnaire (LBQ, Appendix E) collects information on participants general linguistic background (L1, other L2s and associated proficiency levels) and elicits more detailed information on experiences with the Korean language. I used Marian, Blumenfeld, and Kaushanskaya's (2007) Language Experience and Proficiency Questionnaire (LEAP-Q) as a basis for the language background questionnaire, adding some items about current class level, Korean proficiency test results, prior instruction, and heritage status. Additionally, I removed some of the accent items from Marian et al., as these aspects are covered by the self-assessment. The LBQ was presented bilingually in Korean (the learners' target language) and English (a widely-known lingua franca).

Pronunciation Self-Assessment

The pronunciation self-assessment (SA, Appendix F) was intended to capture (a) perceptions of global pronunciation abilities and attitudes, and (b) awareness of pronunciation strengths and weaknesses at the level of individual phonemes. All self-assessment items utilize positively-oriented left-to-right numerical scales, i.e., the leftmost point indicates the least/worst and the rightmost point indicates most/best. Like the LBQ, the SA was presented bilingually in Korean and English. The first part of the SA contains items targeting self-perceived comprehensibility and accentedness, following Derwing and Munro's (1998, 2015) widely-used, simple 9-point scales. Learners were directed to focus on how others react to their speech, and to focus more on how they produce speech rather than what they are able to say (i.e., make judgments primarily based on their articulation rather than their knowledge of vocabulary or syntax). Additionally, 9-point scales targeting satisfaction with current pronunciation abilities and value placed on pronunciation were included.

The second part of the instrument deals with the difficulty in (a) production and (b) perception of each phoneme in Korean's inventory (28 phonemes in 2 modalities = 56 total items). When self-assessing, learners indicated how often they have difficulty with a sound in either modality on a 7-point scale, ranging from 1 ("Always") to 7 ("Almost never"). For production items (k = 28), reliability (Cronbach's alpha) was .95. For perception items (k = 28), Cronbach's alpha was also .95.

Independent Speaking Task

To elicit naturalistic, spontaneous speech, I created an independent speaking task following Kim et al.'s (2016) description, which they based on the TOEFL independent speaking task. The prompt (in English) was: "Some people prefer to live in a small town. Others prefer to

live in a big city. Which place would you prefer to live and why?" I produced a Korean translation, which was copy edited by a Korean-English bilingual with Korean teaching experience. This task should have been accessible to advanced beginners and higher. Much of the vocabulary (e.g., descriptive adjectives, places) and grammar structures (e.g., present tense, patterns to express like/dislike, comparatives) are covered within the first semester or two of coursework in most Korean programs.

The task directions and prompt were presented bilingually on paper (Appendix G). I gave oral directions in the participant's preferred language (Korean or English) and I always read the prompt aloud in Korean. After the directions and reading of the prompt, participants were given 15 seconds to think about their responses and then up to 1 minute to speak. Participants were not cut-off immediately at the one-minute mark, I allowed them to continue until a natural stopping point.

A team of three coders, all native speakers of Korean with training in linguistics, completed broad phonemic transcriptions of all 198 Independent Speaking task responses collected during Field Testing. A set of six tasks were transcribed by all three coders. At the phoneme level, agreement was achieved when all three coders indicated the same phoneme; where one coder differed a partial agreement (assigned a conservative value of 0.5) was recorded. The agreement among coders across a total of 2,136 phonemes was 92%.

While intercoder agreement was high, I noticed some inconsistency among coders in how they applied transcription conventions. For example, the common verb 있다 (/it.t* ɑ/, *to exist, to have*) in some of its inflectional variants was sometimes inappropriately transcribed with the 씨 letter, corresponding to the phoneme /s*/, even though the speaker clearly did not articulate that sound. This likely arose due to /s*/ (written as 씨) always changing to /t/ in coda positions in

Korean. Accordingly, I carefully reviewed audio files and corrected all phonemic transcriptions used in analyses to increase the consistency across each speech sample.

Elicited Imitation Test

The Korean elicited imitation test (EIT) developed by Kim, Tracy-Ventura, and Jung (2016) served as an independent measure of learners' oral language proficiency.

The Korean EIT consists of 30 items and takes approximately 10 minutes to administer. Each item requires the learner to listen to a spoken sentence, wait 2 seconds, and then repeat the sentence orally. Sentences range from 7 to 19 syllables in length; the length of sentences increases as the test progresses. Learner responses are recorded. Each item is scored on a 0-4 scale as follows, with 120 total points possible:

- 4: Perfect repetition without any discrepancy
- 3: Accurate content repetition with minor changes in form allowed
- 2: Features changes to content and/or form that affect meaning
- 1: Includes half of the sentence or less
- 0: Any of the following: no response, only one word repeated, or unintelligible repetition

Kim et al. (2016) reported 95% exact agreement between two raters, and an internal consistency of .96 (Cronbach's alpha). Based on a sample of 66 Korean learners living in Korea and had an average of 3 years residence (min = 2 months, max = 7 years), Kim et al. found a mean score of 52.82 (SD = 24.10).

For this study, the directions of the test and practice items were translated into Korean, drawing on Park's (2014) Korean-language instructions for an English EIT, and simplified in order to make the task more accessible to lower-proficiency Korean learners who do not have a strong command of English (Appendix H). No changes were made to the test items from Kim et

al. (2016). A Korean NS research assistant scored participant EIT responses. The internal consistency (Cronbach's alpha) of the EIT was .96. I scored a subset of 20 randomly-selected EIT responses. The total-score interrater reliability was r = .97.

Semi-Structured Interviews

For the interview and retesting study, I conducted face-to-face semi-structured interviews (Brinkmann, 2013). Learners participated in one or two interviews, and the teacher participated in one interview. Interviews involved responses to stimuli (KPD results and self-assessment results) and a set of pre-defined questions. As the interviews were only semi-structured, I probed further when participants made interesting comments and/or did not directly or elaborately answer questions. I also gave the floor to participants at the end of the interview, encouraging them to ask their own questions or bring up anything else that was on their minds related to the KPD and/or Korean pronunciation. All interviews were recorded and transcribed.

I designed the structure of interviews for learners and the teacher was to facilitate connections across time and perspectives. Figure 5.1 outlines the general structure of these interviews, and complete interview protocol can be found in Appendix I. The first interview for learners included four phases: Orientation and Reflection, Interpreting KPD Results, Learning Activity, and Progress. The first phase directs learners to think about their pronunciation and reviews their self-assessment responses. The second phase focuses on how learners understand their KPD results and elicits differences between self-assessment and KPD results. The Learning Activity phase explores learners' pronunciation learning practices (and those found in their classes) along with their immediate thoughts on what they might try after seeing the KPD results. The last phase, Progress, has learners reflect on their pronunciation learning history and future pronunciation goals.

The second interview for learners included three phases: KPD Results, Learning Activity, and Progress. The KPD Results phase connects to the Interpreting KPD Results phase from the first interview but shifts the focus to what aspects of the KPD Results remained salient for learners after some time has passed. The Learning Activity phase of this second interview is aligned with the phase of the same name in the first interview, but this time focused on what the learner has done in the interval between interviews. Similarly, the Progress phase touches on learner perceptions of learning progress over the time interval between the first and second interviews.

The interview for the teacher included four phases: Pronunciation Teaching, Teacher's Observations of Students, Interpreting KPD Results, and Utilizing Results. The Teacher's Observation of Students phase aligns with the Orientation and Reflection phase of the first learner interview, showing another perspective on informal observations of student's strengths and weaknesses. Similarly, the Interpreting KPD Results phase is parallel to the learner interview phase of the same name. The Pronunciation Teaching and Utilizing Results sections elicit information on current, typical teaching practices and ways in which the KPD results could be applied in a classroom setting, respectively.

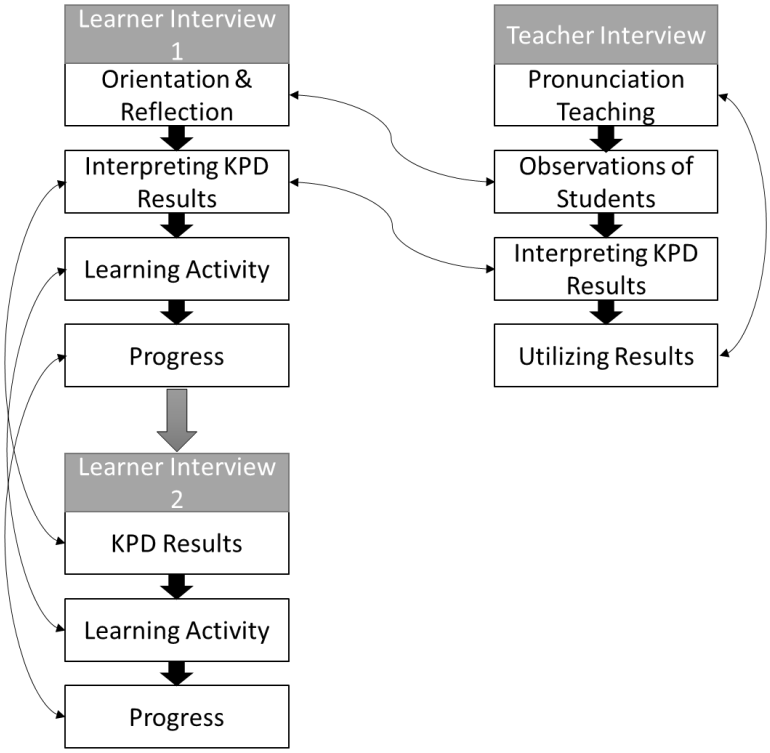


Figure 4.1. Structure of interviews. Single-headed arrows indicate sequence, double-headed arrows indicate content relationships.

Procedures

Study procedures are divided into two sets: Field Testing, which all learner participants will undergo, and Interview Study with Retesting. The order of study activities is listed below for each set.

- Field Testing
    1. Informed Consent (3 minutes)
    2. Language Background Questionnaire (10-15 minutes)
    3. Self-Assessment (5-10 minutes)
    4. Independent Speaking Task (3 minutes)
    5. KPD (15-20 minutes)
    6. EIT (10 minutes)

- Interview Study and Retesting
    1. Interview 1 (learners and teacher)
    2. Interview 2 & Retest (subset of learners)
        - Part 1: Interview
        - Part 2: Independent Speaking Task
        - Part 3: KPD

Analyses

This dissertation makes use of several quantitative analyses. When examining test score data, classical test theory (Crocker & Algina, 1986) and Rasch measurement (Rasch, 1960/1980) approaches were used. Correlations were used to examine the relationships among tasks and between instruments (e.g., between the KPD and self-assessments). I used cluster analysis (Yan & Ginther, 2017; Staples & Biber, 2015) to explore learner profiles based on KPD results. Additionally, I qualitatively analyzed interview data to investigate content related to participant understanding and application of KPD results. For coherence and readability, detailed descriptions of analyses can be found immediately preceding the results of each analysis in subsequent chapters. Basic analytical details for each RQ are outlined below:

- RQ1a: How reliable is the KPD?

  Analysis: Cronbach's alpha and Rasch-based reliability estimates were computed based on individual items for each task, each modality, and the whole test. Items in each modality parceled according to target phoneme were created and Cronbach's alpha and Rasch-based reliability estimates were calculated.

- RQ1b: How reliably are production items evaluated by different scorers?

  Analysis: Interrater reliability for production tasks were analyzed via computation of Cohen's kappa.

- RQ2a: What is the internal structure of test tasks?

  Analysis: Pearson correlations were run between the KPD total score and each task. Correlations were also be run among all tasks.

- RQ2b: Do item difficulty hierarchies align with expectations and previous research?

  Analysis: Item facility (percentage correct) and Rasch item difficulty estimates were computed.

- RQ3: Do scores indicate distinct test-taker profiles in terms of mode, articulatory features, and/or mastered phonemes?

  Analysis: Cluster Analysis was used to investigate the presence of clusters that represent distinct profiles in pronunciation ability.

- RQ4: Do overall results show expected relationship with Korean oral proficiency?

  Analysis: Correlations between KPD total scores and EIT results were computed.

- RQ5: Do KPD Results reflect difficulties test-takers show in spontaneous, meaning-focused speech?

  Analysis: Independent speaking task responses were phonemically transcribed by NS coders. Total error errors were tallied and normed to a standardized rate (per 100 words). Additional tabulations were made for individual phonemes, features, and contexts. Pearson correlations between total phonemic error rates in the independent speaking task and KPD scores were run. Additional correlations were run for target phonemes and features between the KPD and independent speaking task.

- RQ6: To what degree do KPD Results reflect self-assessments of pronunciation ability and difficulties?

  Analysis: Correlations were run between KPD results and self-assessments.

- RQ7: To what extent do a) learners and b) teachers understand score reports and/or learn anything new from them?

  Analysis: Interview data were coded for alignment and discrepancies between a) learners' understanding and KPD results and b) teachers' understanding and KPD results.

- RQ8a: Do learners report any changes in their self-study routines and/or their attention to phonological form in formal or informal learning situations?

  Analysis: Interview data were coded for pronunciation-related study activity and pronunciation awareness/attention. Codes were analyzed within subjects across time, allowing for analysis of changes in study activity and awareness/attention.

- RQ9: To what degree do learners show improvements in a) overall and/or b) weak areas after receiving KPD Results?

  Analysis: A subset of interviewed students were retested roughly 2-3 months after receiving their initial KPD score report. Initial KPD and post-test KPD scores were compared. Within-groups t-tests were used at the group level, and descriptive statistics were tallied to examine changes for individual learners on phonemes and features. These results are considered alongside interview data on pronunciation-focused learning activities.

CHAPTER 5: MEASUREMENT

In this chapter, I present results related to the measurement properties of the KPD. This includes basic summary statistics of whole test, section, and task scores as well as more detailed item analyses, reliability analyses, and analyses related to the internal structure of the test (i.e., part-total and part-part correlations). Results are primarily focused on learner test data, but NS test data is also considered where relevant and appropriate.

Research Questions

As a convenience to readers, the RQs addressed by the results in this chapter are as follows:

- RQ1a: How reliable is the KPD?
- RQ1b: How reliably are production items evaluated by different scorers?
- RQ2a: What is the internal structure of test tasks?
- RQ2b: To what extent do item difficulty hierarchies align with expectations?

Analysis Details

Brief descriptions of analyses were provided in the Methods chapter (Chapter 4). In what follows, I provide more detailed descriptions of the measurement analyses.

Measurement Models

A measurement model (or scale) can be simply defined as the way in which scores are assigned to objects of measurement (Hand, 1996; Stevens, 1946). In this case, I am concerned with how scores from the KPD are assigned to L2 speakers of Korean. All individual KPD items are scored dichotomously, and all items reflect some facet of a learner's phonological competence in Korean. Thus, a simple measurement model would simply be the sum of all KPD items as a reflection of phonological competence. However, this approach is of limited use and

relevance in the present context. Rather, theory and empirical research support the idea that productive and perceptive phonology, while related, are distinct. In turn, it would be defensible and more informative to calculate separate total scores for the production and perception sections of the test, where a learner's production ability is reflected by the sum of all production items and perception ability is reflected by the sum of all perception items; the two abilities are expected to be correlated because these two skills are related in their development; growth in one can support the growth in the other (most often, growth in perception aids growth in production). In the measurement models for production and perception abilities, item analyses for diagnosing poorly-performing items and examining the expected hierarchy of item difficulties would occur at the level of individual items.

However, KPD results are not intended to be used as simple sums reflecting an overall level of phonological competence. Rather, sub scores for each phoneme in production and perception, each based on the subtotal of several individual items, are the primary unit of interpretation and intended use (Dorans, 2018). Furthermore, due to variation among phonemes in the number of critical allophones and their overall frequency of occurrence in real words (Shin, Kiaer, & Cha, 2012), each phoneme is represented by non-uniform numbers of individual items. In other words, raw phoneme subtotals are not tau-equivalent (i.e., phonemes are not equally weighted by default), making some phonemes more important than others when using a simple sum of item scores to represent overall production or perception ability. Thus, I found it appropriate to consider the use of measurement models in which (a) subscores are aggregated at the phoneme level, such as within *item parcels*, and (b) scale weights of individual phonemes are uniform. To accomplish this, I computed item parcels for each phoneme in production (Task 1 and Task 2) and perception (Task 3 and Task 4) by summing all of the individual items that

target a given phoneme (refer to Appendix A). In measurement analyses, parcels are made tau-equivalent by (a) converting to percentage scores in CTT analyses or (b) specifying equal parcel weights in Rasch analyses. Thus, the measurement model of parcels maps test-takers' overall abilities by considering equally performance on each of the 28 Korean phonemes.

The creation of item parcels, also called *item bundles* or *super items*, warrants further discussion. Item parceling typically involves the principled summing of multiple individual items into one polytomous item. Instead of considering the dichotomous items A, B, and C separately in analyses, the responses to all three items are summed and considered as Parcel X with a scale of 0-3 points. This effectively reduces the total number of items on a test, potentially reducing the reliability of scores (Marais & Andrich, 2008), but this is mitigated by the increased amount of information about test-taker abilities provided by a parcel compared to any single item. This is referred to as a *score-based approach* to item parceling (Eckes, 2014). *Item-based approaches* to parcel measurement also exist but are excluded here due to technical complexity and concomitant sample size requirements.

There are two main reasons for parceling items: content and context. Parceling by content groups items that tap into the same aspect of a larger, overarching construct, e.g., items on a test of receptive phonological knowledge which target the same phoneme. Parceling by context groups items that share a context which influences responses across items. For example, consider a reading comprehension test where a test-taker must read a passage and then answer a main idea question followed by a question about the author's purpose: If a test-taker does not correctly identify the main idea of the passage, they might be less likely to subsequently identify the author's purpose for writing it. Marais and Andrich (2008) discuss these phenomena in terms of *local item dependence*, that is, sets of items with stronger than expected relationships in

responses, and refer to two types of dependence: *trait dependence* (corresponding to content) and *response dependence* (corresponding to context). Accounting for local dependence is critical to the application of many measurement models (e.g., Rasch, IRT) and can lead to better measurement of underlying test-taker ability. One common application of item parceling is in the creation of *testlets* for several dichotomous items which share a common stimulus, e.g., a text followed by several comprehension questions (e.g., Eckes, 2014). Parceling has also been applied to C-tests for items which have several dichotomous items embedded in the same paragraph (e.g., Lee-Ellis, 2009).

I chose to run and report analyses for both measurement models, individual items and item parcels, due to the quality assurance benefits of examining individual items and the necessity of considering the way scores are actually intended to be interpreted and used (i.e., item parcels).

Two Statistical Approaches to Measurement

In addition to there being more than one measurement model relevant to the analysis of the KPD, there are also multiple statistical approaches available for analyzing the measurement properties of the test. In the field of measurement, a general distinction is made between classical test theory (CTT, Crocker & Algina, 1986; DeMars, 2018) and item response theory (IRT, Brown, 2018; Meijer & Tendeiro, 2018). Tests with dichotomously scored items as well as tests with polytomously scored items can be analyzed with CTT and IRT. In short, CTT maintains that an observed score on a test is an examinee's 'true' ability, plus or minus a constant amount of measurement error. Thus, CTT aligns well with theory of measurement known as operationalism, which holds that an attribute, essentially, is defined as the score on the test (Hand, 1996). In the present context, this would be akin to saying that a learner's pronunciation

accuracy is one and the same with their KPD score. In contrast, IRT is based on the notion that what is really being measured (i.e., the attribute possessed by examinees) can only be measured indirectly: This *latent* attribute (or trait) is not something that can be directly measured, but its level can be inferred through analysis of observable responses to items. This approach is better aligned with the theory known as *representational measurement*, which aims to establish accurate links between the test scores from people who vary in their relative levels of the attribute (Hand, 1996). With reference to the KPD, this theoretical approach holds that a learner's underlying pronunciation (or perceptual) abilities are represented by scores on the KPD; this representation is mediated by the content and technical qualities of the test.

While these two statistical (and theoretical) approaches to measurement analysis differ in several other ways (see Embretson, 1996, and DeMars, 2018, for summaries), they do share several important features: (a) Tests should measure a single dimension, (b) scores from several items may be summed or otherwise combined, and (c) several statistical analyses are available to investigate flaws in individual items. Usefully, in the simplest of IRT models (i.e., the dichotomous Rasch model and some variations of it), raw sum scores of all items or item parcels will correlate nearly perfect with model estimations of person ability. This is helpful because a simple total of raw scores is easier to explain to and be interpreted by test users who are not savvy in quantitative measurement techniques, and it facilitates comparisons of information about the same dataset obtained by the two approaches.

Despite sharing some basic similarities, IRT offers several practical advantages over CTT. For one, IRT places the ability of examinees and the difficulty of items (or parcels) onto the same interval scale of measurement, allowing for the direct and meaningful comparison of item difficulty and person ability statistics. This can be useful for interpreting what typical and/or

particular examinees know or can do. Additionally, IRT allows for a more robust consideration of measurement error. Whereas CTT considers error as constant throughout the range of person ability, in IRT error can be examined conditionally along the continuum of person ability (through a calculation of *information* aggregated at the test level) as well as at the level of individual items (through calculations of *information* at the item level). Thus, IRT facilitates consideration of measurement error at critical score ranges, such as around cut-points for interpretation or decision making.

Nonetheless, IRT approaches do have some drawbacks. For one, they typically require large(r) sample sizes to estimate model parameters. For the simplest dichotomous models (i.e., 1-parameter models), tests of at least 30 items and sample sizes of 200 to 250 examinees meet minimum recommendations, though in one variation of 1-parameter models, the Rasch model (see below for details), Linacre (1994) has argued that meaningful results can be obtained for 30 item tests with fewer examinees. Linacre advised an absolute minimum of 30 examinees for dichotomously scored tests and 50 for polytomously scored tests, and further suggested that Rasch analyses conducted with 100 to 150 examinees will yield estimates of item and person ability within a reasonably narrow confidence range (0.5 logits).

Beyond sample size considerations, IRT models have stricter stances on the relationship between model estimates and response data. For most IRT approaches, models must be adjusted to fit a given set of response data. This can be done by adding additional parameters to the model to be estimated freely, such as item discrimination and/or a guessing parameter. However, doing so requires even larger sample sizes (e.g., 500 examinees in order to include a discrimination parameter, and 1,000 examinees to include both discrimination and guessing parameters), and is thus not considered further here. In the Rasch family of models, all item discriminations are

uniformly constrained, and in line with a view of the Rasch model as prescriptive, item response data must fit the model (rather than the other way around). What this view dictates is that elements of measurement (items, persons) which demonstrate poor fit to the model should be removed. When there are a large number of individual items, removing a handful of poorly-fitting items is usually not a grave concern. However, it is often difficult to justify removing examinees and large numbers of items or a whole content-based item parcel. After all, people who do not fit the Rasch model still may wish to receive diagnostic feedback on their pronunciation! Similarly, from a content perspective, it is often unreasonable, if not absurd, to remove substantial portions of content from a test due to poor fit statistics.

Bowles, Skibbe, and Justice (2011) illustrated this problem in their Rasch analysis of an assessment of letter name knowledge (LNK) for 909 children in the early stages of literacy development. In their LNK test, which featured one item for each letter in the English alphabet that children must point to and name, several items (i.e., letters) were found not to fit the Rasch model. Bowles et al. noted the absurdity, from a content perspective, of effectively removing letters of the alphabet to satisfy Rasch model fit demands. In the case of the KPD, removing a phoneme-based item parcel would not be justifiable, as all phonemes are undeniably part of the attribute being assessed and potentially relevant to making subsequent instructional decisions.

I elected to conduct both CTT and Rasch measurement analyses. Doing so allowed for the examination of converging or diverging evidence of measurement qualities. The inclusion of Rasch measurement provided the previously discussed benefits over CTT, while CTT served as both as an additional perspective on the data and as a "back-up" in the event that the data showed unignorable misfit to the Rasch model. In the following subsections, I provide relevant technical details for the present analyses conducting using each approach.

Classical Test Theory Analyses. In CTT, the relationship between test scores and the "true" score associated with the attribute of measurement is defined through the following equation (1):

(1) Observed Score = True Score + Error

Where the observed score is typically the sum of all item/task scores and error is typically estimated via standard error of measurement (SEM), which is calculated based on test reliability (e.g., Cronbach's alpha) and the standard deviation of test scores (see Brown, 1999, for the formula). Thus, an examinee's true score is estimated as falling somewhere within an interval defined by the observed score plus or minus the SEM. In practice, such as when using test scores for subsequent statistical analyses, the observed score is taken as a good estimate of an examinee's ability level on the attribute.

In CTT, statistics used for characterizing the qualities and performance of items include item facility (P) and item discrimination (D). P is the proportion of correct responses across all examinees for dichotomous items, or the averaged scores from all examinees for polytomous items. D is the association between test-takers' responses on an item and their overall scores on the test, typically estimated via correlation (the approach taken here) but sometimes as the difference in P between the examinees in the top and bottom third of total scores (see Carr, 2011, for ways to calculate P and D). Item discrimination, which is a value that runs from -1 to 1, is useful as an indicator of an item's technical quality; larger and positive discrimination values indicate that more able examinees responded correctly than less able examinees (which is desirable), while negative values indicate the opposite, which is obviously undesirable. Values at or near zero mean that the item did not discriminate, which means that the item provides no information and is not useful for measuring the underlying construct, at least from a

psychometric perspective. This can happen, for example, when everyone gets the item correct, or when everyone gets the item wrong (which is information that may be useful to teachers), or when responses on the item are seemingly random (which is information that may not be immediately useful to teachers). For dichotomous items, I used point-biserial correlations to calculate discrimination, and for polytomous item parcels I used Pearson correlations between the parcel score and the total score minus the parcel.

Rasch Analyses. Rasch analyses yield estimates of ability for each person, difficulty estimates for each item, fit statistics for persons and items, and estimates of reliability for both person ability and item difficulty (Bond & Fox, 2015; Boone, Staver, & Yale, 2014). Person ability and item difficulty are both expressed in *logits* (log-odds units), which relate to the probability that a given examinee will produce a correct response to a given item. For the PCM, the Rasch-Andrich difficulty threshold between each step of the scale (e.g., the boundary between a sum score of 3 or 4 on all items targeting /k*/) is estimated, based on the point along the person ability continuum where an examinee would have 50-50 odds of scoring in the higher or lower category, with the average difficulty of all thresholds reported as the overall item parcel difficulty. The measurement quality of person ability at the level of the whole test or individual items can also be examined using information functions; more information means more robust and precise measurement of ability. Test information functions (TIF) represent the amount of information yielded for examinees along the ability continuum; information is maximized where there are more items (or partial-credit scale steps) at or near a given person ability level. Similarly, item information functions (IIF) are maximized where item difficulty is equal to person ability. For the dichotomous Rasch model, all items have the same IIF, but IIFs for

polytomous items may take different shapes based on information associated with each scale-step (Linacre, 2005).

For both production and perception KPD items, I used two Rasch models to analyze response data: (1) the dichotomous Rasch model (Rasch, 1960/1980) for individual item analyses, and (2) the Rasch partial-credit model (PCM, Masters, 1982) for item parcels. All Rasch models were estimated using the Winsteps software (version 4.3.4). For both models, item response data from all 198 examinees was included. For the dichotomous Rasch model, this sample size was expected to yield highly-accurate model parameters per Linacre (1994). For the Rasch PCM, where the partial-credit scale thresholds for each item parcel is estimated separately from all other parcels, the sample size of 198 participants should be sufficient (Linacre, 1994, p. 328 noted that "100 responses per item may be too few").

Aside from sample size and precision considerations, the Rasch models used also assume unidimensionality (see Chapter 1 for conceptual discussion of unidimensionality). Assessments of unidimensionality, within the framework of a Rasch analysis and without performing more data-intensive item factor analyses, entail the analysis of model residuals via fit statistics and principal components analysis.

Fit statistics, at the level of individual observations or aggregated at measurement elements (i.e., persons, items) provide information on how frequently and significantly the item response data do not fit the unidimensional Rasch model. At the level of individual observations, model predicted values are compared to empirical values and the difference is standardized, allowing for interpretations following a Z distribution (i.e., critical values of ≥ 2 or ≥ 3 are considered statistically significant at the .05 and .01 alpha levels, respectively). Linacre (2019) proposed that when fewer than 5% of residuals exceed the $Z \geq 2$ threshold and fewer than 1%

exceed $Z \geq 3$. For person and item fit statistics, it is common to examine both *infit* (information-weighted fit, reported as a mean-square) and *outfit* (outlier-sensitive fit, also reported as a mean-square). The former is sensitive towards deviations from model expectations in observations near the estimated measure (e.g., observations from persons with ability near the difficulty of an item), while outfit is sensitive to deviations in observations where there is greater distance between measurement elements (e.g., when a high-ability person responds incorrectly to a low-difficulty item, or when a low-ability person responds correctly to a high-difficulty item). Both statistics may range from 0 (representing "overfit", where responses are *too* predictable) to infinity (representing increasingly large and frequent deviation in responses), with 1.0 indicating perfect fit. Common guidelines for interpreting infit and outfit state that values between 0.7 and 1.3 are acceptable for most purposes, with values between 0.5 and 1.5 acceptable in low-stakes assessment contexts (Wright & Linacre, 1994).

Principal components analysis (PCA) of the Rasch model residuals allow for the detection of systematic patterning among residuals which may indicate additional measurement dimensions of substance that could interfere with measurement of the primary Rasch dimension. One typically looks for eigenvalues greater than 2 in the first one or two contrasts when determining whether any additional measurement dimensions might be substantial enough to interfere with the unidimensionality requirement. When contrasts have generally small eigenvalues, it is relatively safe to assume any patterning in Rasch residuals to simply be reflective of noise.

As discussed previously, it is common to remove (or at least revise) persons or items that do not fit the Rasch model. However, because I went into this research uncertain of whether a unidimensional Rasch model is appropriate for the KPD, I considered these analyses exploratory

and did not engage in the typical subsequent trimming of items/parcels or people who did not fit the Rasch model. My main interests were (a) obtaining useful information on the reliability of the KPD, performance of KPD items/parcels, and the hierarchy of KPD items/parcels, and (b) determining the general suitability of applying Rasch measurement to the KPD.

Reliability Analyses

Reliability was considered from two perspectives: (1) conventional test reliability indices from CTT (internal consistency) and Rasch (person reliability) analyses, and (2) the inter-scorer reliability among several teachers (scorers) for the production section of the KPD.

From the perspective of conventional test reliability, all 198 KPD responses were scored by an experienced instructor of Korean and submitted to Cronbach's alpha analyses in *R* using the *psych* package (version 1.8.12, Revelle, 2018). Cronbach's alpha is a flexible, although conservative, method of estimating test reliability, and is able to accommodate dichotomously and polytomously scored items. For individual dichotomously-scored items, alpha was calculated for the whole test, production and perception sections separately, and separately for each task. For polytomously-scored item parcels, alpha was calculated for all parcels together and production and perception parcels separately. A commonly-used Rasch correspondent to Cronbach's alpha is the person separation index (Linacre, 2019); this was estimated in Winsteps separately for production and perception items/parcels.

To investigate reliability among several scorers, I recruited six additional scorers, all of whom were Korean NSs pursuing graduate degrees related to teaching Korean as a foreign language. These six scorers varied in their teaching experience; some had only limited tutoring experience while others had up to a year of formal classroom teaching experience. These six scorers rated a random sample of 20 KPD responses, in which I deliberately included two

(randomly-selected) Korean NS responses. The dichotomous scores from all seven scorers (including the primary scorer) for each item were submitted to calculations of interrater agreement and reliability using the *R* packages *irr* (version 0.84.1; Gamer, Lemon, & Singh, 2019) and *ragree* (version 0.0.4; Redd, 2019) including percent agreement, Fleiss' Kappa (a variant of Cohen's Kappa for more than two scorers), and Gwet's AC1. Percent agreement is a crude measure of interrater agreement that does not account for agreement by chance, and values closer to 100% are considered more desirable. Kappa ranges from -1 to 1 and adjusts for chance agreement, making it a superior estimate of interrater reliability to percent agreement, and is commonly interpreted according to Landis and Koch's (1977) benchmarks: kappa < 0.0 = poor, 0.0 ≤ kappa < 0.2 = slight, 0.2 ≤ kappa < 0.4 = fair, 0.4 ≤ kappa < 0.6 = moderate, 0.6 ≤ kappa < 0.8 = substantial, 0.8 ≤ kappa ≤ 1.0 = almost perfect. Gwet's AC1 accounts for both chance agreement and models random guessing by scorers (although truly random guessing is most likely not present in a context such as this one) and is a less-biased estimate of interrater reliability compared to kappa, especially when there is a high prevalence of one response option in the data (Gwet, 2008). It has the same range as Fleiss' Kappa and follows the same benchmarks for interpretation. For Fleiss' Kappa and Gwet's AC1, estimates for items with perfect agreement that all examinees responded correctly (or incorrectly) cannot be produced. In these instances, I manually recoded the indices to 1.0, representing perfect agreement. To summarize overall levels of interrater agreement/reliability across all items, I computed means, SDs, and ranges for each index.

Item parcel scores from each scorer were also calculated for each examinee and converted to percentages. I then examined the consistency of these parcel scores from all seven raters through calculation of interclass correlation coefficients (ICC). I used ICCs that modeled

random examinee and random rater effects (ICC(2,1) following the notation of Shrout and Fleiss, 1979), and ICCs that took into account consistency of examinee rankings (ICC_C) as well as absolute agreement in score levels (ICC_A) assigned by different coders (McGraw & Wong, 1996). ICCs may range from -1 to 1, with values closer to 1 desirable. Koo and Li (2016) offered the following guidelines for the interpretation of ICC values: ICC < 0.5 = poor, 0.5 < ICC < 0.75 = moderate, 0.75 < ICC < 0.9 = good, 0.90 < ICC = excellent. ICC values where all scorers were in perfect agreement cannot be estimated; in these cases I manually substituted a value of 1.0 to indicate perfect reliability. To summarize overall levels of interrater reliability of parcel scores, I computed means, SDs, and ranges of ICC values across all parcels.

Finally, to examine the reliability of interpretations and potential impact on decision making, I dichotomized all parcel scores from each rater using a threshold of 75% accuracy that represents the diagnostic flag criterion. This allowed for consideration of the reliability of diagnostic profiles of learners across several scorers. Like interrater reliability for the dichotomously scored items above, I calculated the same three statistics (percent agreement, Fleiss' Kappa, and Gwet's AC1), but this time only for the 28 dichotomized item parcels.

Correlations

To examine the internal structure of the various sections and tasks of the KPD, I ran Pearson product-moment or Spearman rank-order correlations as appropriate. When data were continuous and appeared to reasonably follow a normal distribution, I used Pearson correlations. When data had less variability and/or did not appear to follow a normal distribution, I used Spearman correlations.

Results

The following results provide information on the distribution of KPD scores, reliability of scores, detailed summary statistics of KPD items, and relationships among scores on the various parts of the KPD.

Measurement Summary

In this section, I report top-level summary information on individual item and item parcel measurement models analyzed with CTT and Rasch-based approaches. A brief summary of NS scores follows.

CTT Observed Scores. For individual dichotomously-scored items, sum score statistics are found in Table 5.1 based on all 198 L2 Korean learners who participated in the field testing. Relative to maximum scores, means were high for the whole task, each section, and each task. However, there was some nontrivial variation in sum scores, as shown by standard deviations (SD) and ranges. Figure 5.1 illustrates the distribution of sum scores at the level of individual tasks, sections, and all items of the KPD.

Table 5.1

Summary of Learner KPD Scores

Section	k	M	SD	Range
All	352	310.34	19.35	261 – 350
Production	217	201.21	8.86	178 – 217
Task 1 – Picture Naming	154	146.16	4.85	131 – 154
Task 2 – Nonword Reading	63	55.05	4.96	42 – 63
Perception	135	109.14	11.84	80 – 134
Task 3 – Pronunciation Judgment*	72	50.39	9.52	23 – 71
Task 4 – Identification	63	58.74	3.19	45 – 63

Note. *Excluding filler items.

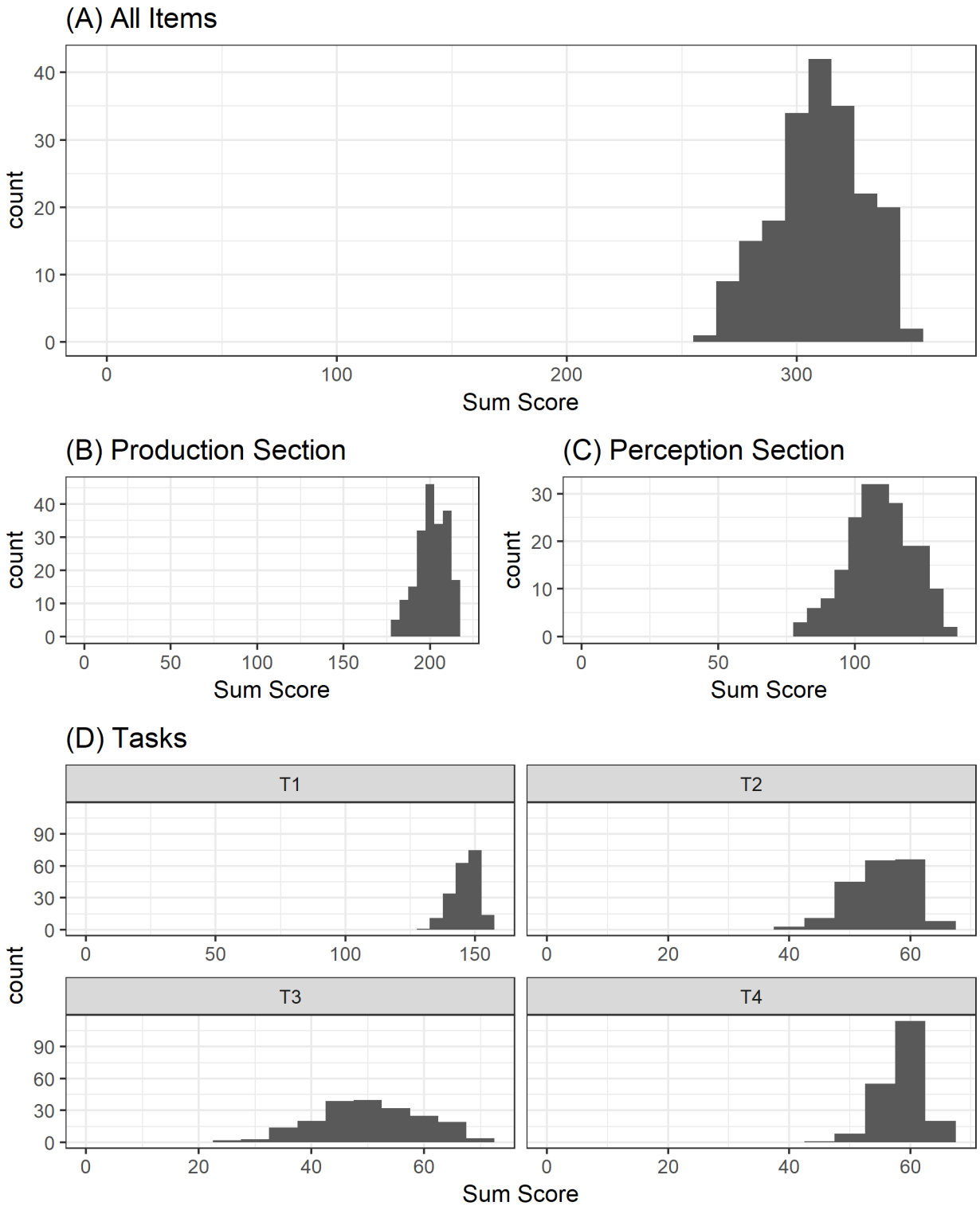


Figure 5.1. Histograms showing the distributions of sum scores for (A) all dichotomous KPD items, (B) all production KPD items, (C) all perception KPD items, and (D) all KPD tasks.

For individual items grouped into phoneme-based parcels (separately for production and perception) and converted to percentage scores to achieve tau-equivalence, the average learner production score was 90.7% (SD = 5.4%, Range = 73.3% – 100.0%) and the average perception score was 80.9% (SD = 9.3%, Range = 56.9% – 99.4%). Figure 5.2 illustrates the distribution of average phoneme accuracy scores for examinees in production and perception.

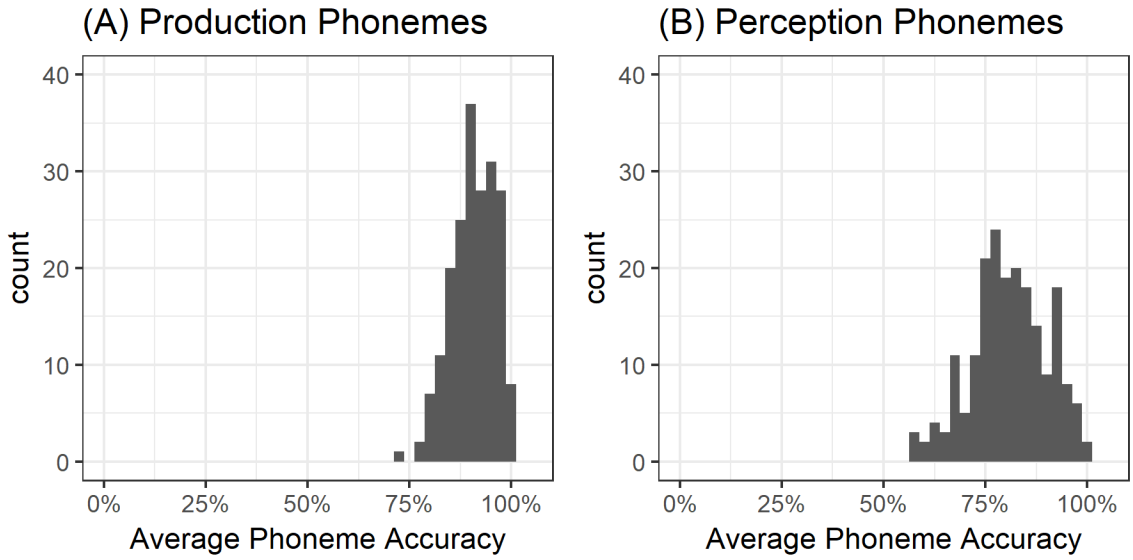


Figure 5.2. Histograms of average accuracy scores across all phonemes in (A) production and (B) perception.

Rasch Models. I estimated Rasch models based on individual items and phoneme-based item parcels for the production and perception sections of the KPD. For individual items, the dichotomous Rasch model was used, and for the item parcels, the Rasch partial-credit model (PCM) was used. Measurement summaries and indices of model fit are provided next.

Production Items. For the dichotomous Rasch model of production item responses, Rasch model parameters explained 18.1% of variance in observations. A total of 1442 observations, approximately 3.4% of the total number, were unexpected at the $Z \geq |2.0|$. At the $Z \geq |3.0|$ level, there were 709 unexpected observations (1.7%). A Principal Components Analysis (PCA) of model residuals found several contrasts with eigenvalues > 2.0 and explained variance in excess

of 2% (Linacre, 2019; first contrast eigenvalue = 6.24, proportion of variance explained = 2.7%). Examination of a scree plot (Figure 5.3) revealed a pronounced elbow at the third contrast. Due to a large number of items, it was difficult to extract meaningful patterns when examining biplots of the residual component loadings, but some informal observations could be made. For example, in the first contrast, I was able to observe some clustering of items targeting consonants, particularly tensed consonants. Thus, it appeared that there may be some dependence among phoneme targets.

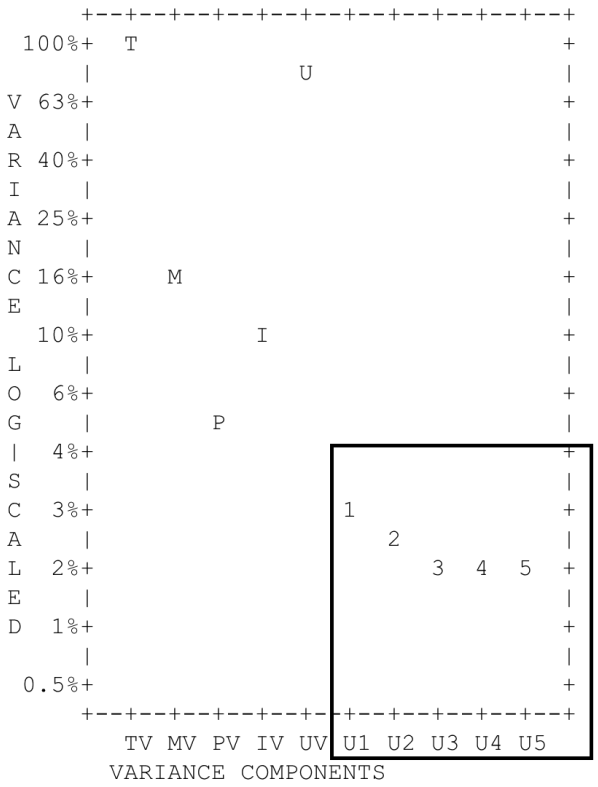


Figure 5.3. PCA of residuals for production items. TV = total variance, MV = variance explained by person & item measures, PV = variance explained by person measures, IV = variance explained by item measures, UV = unexplained variance, U1-5 = unexplained variance in PCA contrasts 1-5. Boxed region contains PCA contrast scree plot.

Table 5.2 contains summary statistics for person and item measures. As a group, examinees had generally high phoneme production ability compared to the difficulty of items.

Figure 5.4 illustrates the test information function (TIF), which provides information on where the most precise measurement occurs on the continuum of examinee abilities the test items. According to the TIF, the most accurate information is yielded from examinees with relatively low production abilities. In terms of infit, nearly all persons and items demonstrated good fit to the model. In other words, examinees with phoneme production ability near the difficulty of items tended to perform as expected. However, for outfit, many items showed overfit (values under 0.7) and underfit (values over 1.3). That is, examinees with generally high or low phoneme production ability performed unexpectedly on otherwise easy (or difficult) items with some nontrivial frequency.

Table 5.2

Rasch Measurement Summary for Production Items

	Avg.			Model				
Element	Measure	SD	Range	S.E.	Infit	Range	Outfit	Range
Persons*	3.36	0.87	1.86 – 6.19	0.33	1.00	0.73 – 1.39	0.99	0.23 – 3.69
Items**	0.00	0.20	-0.22 – 3.41	0.45	1.00	0.86 – 1.19	0.99	0.30 – 3.59

*Based on 197 examinees with non-extreme (i.e., not perfect) scores. **Based on 187 non-extreme items.

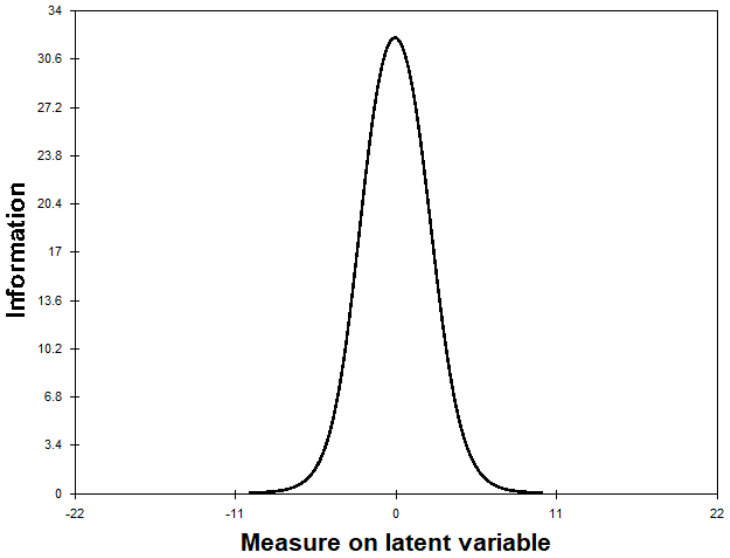


Figure 5.4. Test information function for production items.

Perception Items. For the dichotomous Rasch model of perception item responses, Rasch model parameters explained 37.2% of variance in observations. A total of 1034 observations, approximately 3.9% of the total number, were unexpected at the $Z \geq |2.0|$. At the $Z \geq |3.0|$ level, there were 409 unexpected observations (1.5%). A PCA of model residuals found several contrasts with eigenvalues > 2.0 and explained variance in excess of 2% (first contrast eigenvalue = 6.47, proportion of variance explained = 3.4%). Examination of a scree plot (Figure 5.5) revealed a pronounced elbow at the third contrast. Similar to the production items, it was possible to informally observe some clustering of items with related targets in the first contrast. For example, I observed negative loadings for several items targeting the /s*/ phoneme in the first contrast. Thus, it again appeared that there may be some dependence in residuals related to phoneme targets.

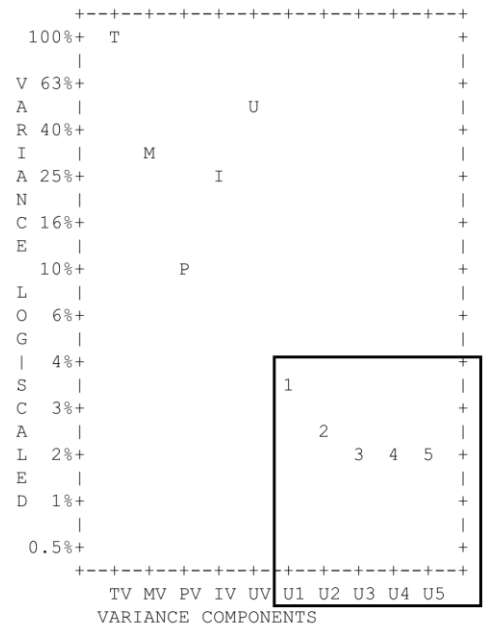


Figure 5.5. PCA of residuals for perception items. TV = total variance, MV = variance explained by person & item measures, PV = variance explained by person measures, IV = variance explained by item measures, UV = unexplained variance, U1-5 = unexplained variance in PCA contrasts 1-5. Boxed region contains PCA contrast scree plot.

Table 5.3 contains summary statistics for person and item measures. As a group, examinees had generally higher phoneme perception ability compared to the difficulty of items, but compared to production items there was more overlap. Figure 5.6 illustrates the test information function (TIF), which provides information on where the most precise measurement occurs on the continuum of examinee abilities the test items. According to the TIF, the most accurate information is yielded from examinees with low to moderate perception abilities. In terms of infit, nearly all persons and items demonstrated good fit to the model, with a few exceptions (8 misfitting persons, 1 misfitting item). In other words, examinees tended to perform as expected on test items whose difficulty levels were closely matched to the examinees' ability levels. However, as seen in the outfit values, many persons and items showed overfit (values under 0.7) and underfit (values over 1.3): 83 misfitting persons and 36 misfitting items. Most outfit issues for people were associated with overfit (n = 57), which indicated that their responses were *too* predictable based on item difficulties.

Table 5.3

Rasch Measurement Summary for Perception Items

	Avg.			Model				
Element	Measure	SD	Range	S.E.	Infit	Range	Outfit	Range
Persons*	2.24	1.03	0.31 – 6.33	0.30	0.99	0.66 – 1.37	0.94	0.07 – 4.46
Items**	0.00	1.91	-3.46 – 4.38	0.33	1.00	0.80 – 1.52	0.94	0.22 – 2.49

*Based on all 198 examinees. **Based on 121 non-extreme items (out of 135 total items).

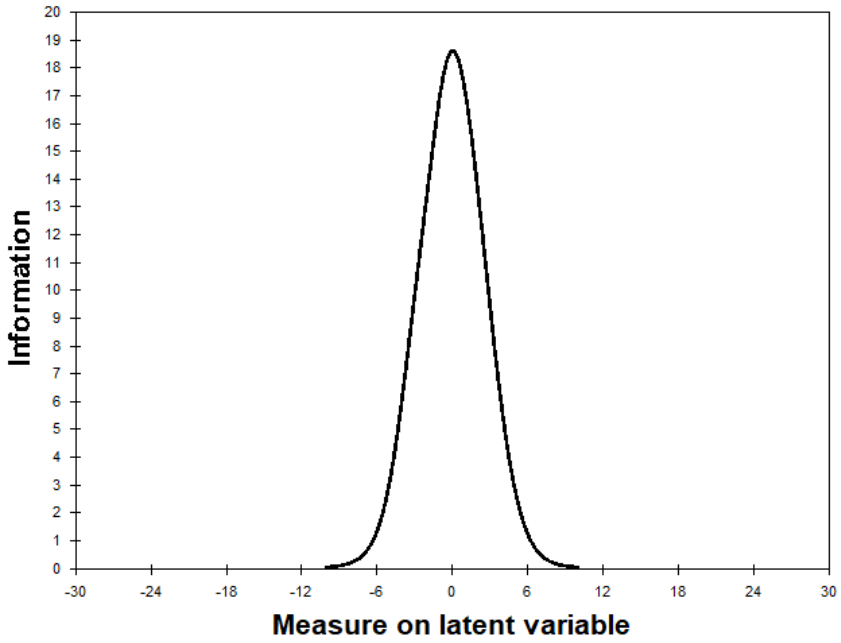


Figure 5.6. Test information function for perception items.

Production Parcels. For the Rasch PCM model of production parcel scores, Rasch model parameters explained 28.9% of variance in observations. A total of 285 observations out of 5,544 (5.1%) were unexpected at the $Z \geq |2.0|$. At the $Z \geq |3.0|$ level, there were 93 unexpected observations (1.7%). A PCA of model residuals found two contrasts with eigenvalues greater than 2.0 and explained variance in excess of 2% (first contrast eigenvalue = 2.62, proportion of variance explained = 6.7%). Examination of a scree plot (Figure 5.7) revealed a pronounced elbow at the third contrast. With the smaller number of parcels (compared to individual items), patterns in contrast loadings were more interpretable: The first contrast was defined primarily by a cluster of tense and aspirated consonants with positive loadings. The second contrast appeared to be characterized mostly by a cluster of lax stops (/k, p, t/). Thus, it appeared that there may be some parcel dependence based on articulatory features associated with phonemes.

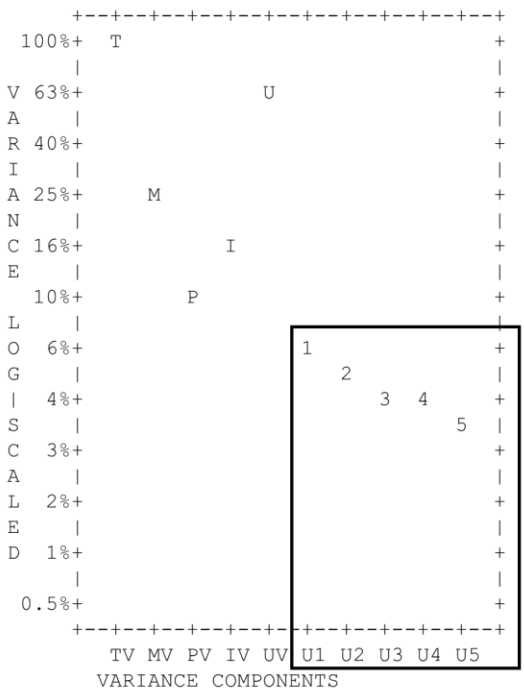


Figure 5.7. PCA of residuals for production parcels. TV = total variance, MV = variance explained by person & item measures, PV = variance explained by person measures, IV = variance explained by item measures, UV = unexplained variance, U1-5 = unexplained variance in PCA contrasts 1-5. Boxed region contains PCA contrast scree plot.

Table 5.4 contains summary statistics for person and parcel measures. As a group, examinees had generally higher phoneme perception ability compared to the difficulty of items, but compared to production items there was more overlap. Figure 5.8 illustrates the test information function (TIF), which provides information on where the most precise measurement occurs on the continuum of examinee abilities the test items. According to the TIF, the most accurate information is yielded from examinees with lower production abilities. In terms of infit, nearly all parcels and most persons demonstrated good fit to the model (58 misfitting persons; 34 overfitting and 24 underfitting). For outfit, more persons and items showed misfit: 93 misfitting persons and 5 misfitting items. Most outfit issues for people were associated with overfit (n =

60), which indicated that their responses were *too* predictable; outfit issues for parcels were slight. Full, detailed information on parcel statistics are found in the following sections.

Table 5.4

Rasch Measurement Summary for Production Parcels

Element	Avg. Measure	SD	Range	Model S.E.	Infit	Range	Outfit	Range
Persons*	1.71	0.76	0.54 – 4.36	0.31	0.98	0.50 – 2.27	0.95	0.18 – 2.67
Parcels	0.00	0.71	-1.54 – 0.93	0.13	1.00	0.88 – 1.19	0.95	0.33 – 1.44

*Based on 197 examinees with non-extreme (i.e., not perfect) scores.

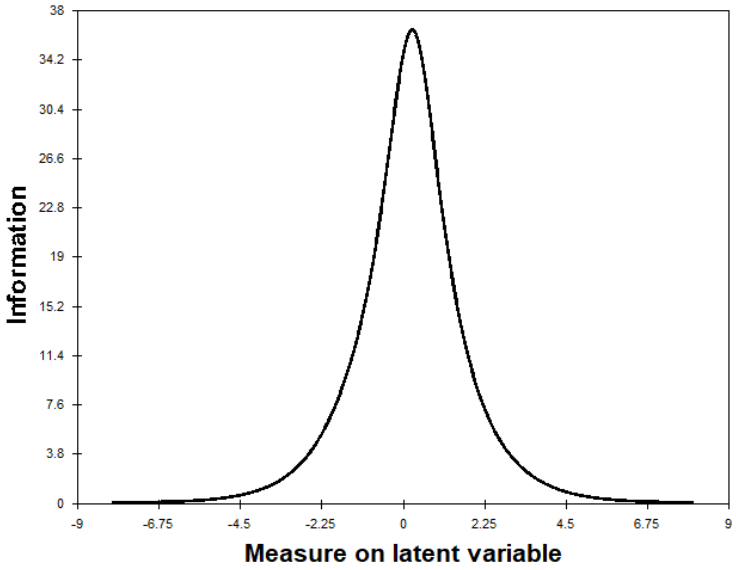


Figure 5.8. Test information function for production parcels.

Perception Parcels. For the Rasch PCM model of perception parcel scores, Rasch model parameters explained 47.5% of variance in observations. A total of 260 observations out of 5544 (4.7%) were unexpected at the $Z \geq |2.0|$. At the $Z \geq |3.0|$ level, there were 50 unexpected observations (1.0%). PCA of model residuals found one contrast with an eigenvalue > 2.0 (first contrast eigenvalue = 3.45, proportion of variance explained = 6.5%). Examination of a scree plot (Figure 5.9) suggests an elbow at the second or third contrast. First contrast loadings suggest that some relation among phonemes with similar articulations may influence measurement. For example, the phonemes affricate stops /tɕ, tɕ*, tɕ^{h}/ (lax, tense and aspirated, respectively) had

large positive loadings (all > .40). Similar patterns are observable for other stop consonants with similar place and manner of articulation, although sometimes lax stops had negative loadings.

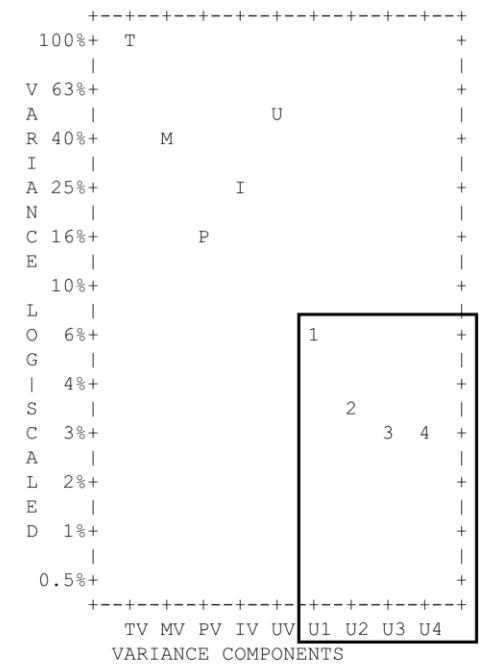


Figure 5.9. PCA of residuals for perception parcels. TV = total variance, MV = variance explained by person & item measures, PV = variance explained by person measures, IV = variance explained by item measures, UV = unexplained variance, U1-4 = unexplained variance in PCA contrasts 1-4. Boxed region contains PCA contrast scree plot.

Table 5.5 contains summary statistics for person and parcel measures. As a group, examinees had generally higher phoneme perception ability compared to the difficulty of items, but compared to production items, there was more overlap. Figure 5.10 illustrates the test information function (TIF), which provides information on where the most precise measurement occurs on the continuum of examinee abilities the test items. According to the TIF, the most accurate information was yielded from examinees with low to moderate perception abilities. In terms of infit, nearly all parcels (except one, for /s*/) and most persons demonstrated good fit to the model (61 misfitting persons; 30 overfitting and 31 underfitting). For outfit, more persons

and items showed misfit: 71 misfitting persons and 3 misfitting items. Most outfit issues for people were associated with overfit (n = 45), which indicated that their responses were *too* predictable; outfit issues for parcels were slight. Full, detailed information on parcel statistics are found in the following sections.

Table 5.5

Rasch Measurement Summary for Perception Parcels

	Avg.			Model				
Element	Measure	SD	Range	S.E.	Infit	Range	Outfit	Range
Persons*	1.45	1.04	-0.40 – 5.80	0.30	1.00	0.42 – 2.00	0.99	0.06 – 2.90
Parcels**	0.00	0.76	-1.34 – 1.82	0.12	1.00	0.77 – 1.36	0.99	0.68 – 1.57

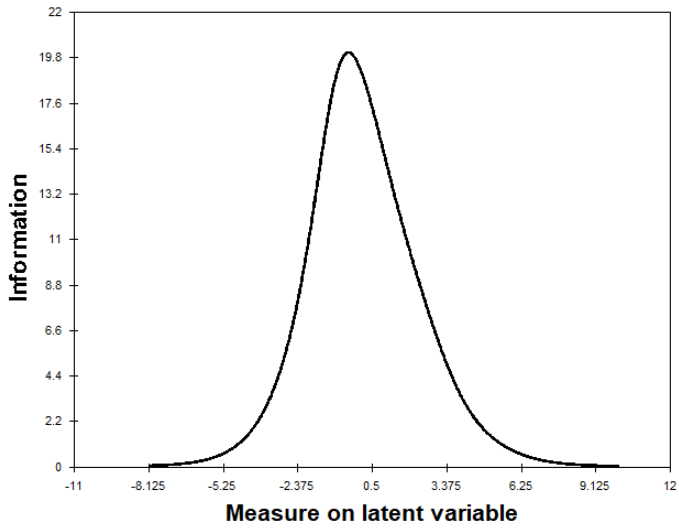


Figure 5.10. Test information function for production parcels.

Native Speakers. Summary statistics for total scores from 6 NSs of Korean are in Table 5.6. NS performance on the KPD was at or very near ceiling; this was true for individual tasks as well. For phoneme parcels, the average NS production score was 99.9% (SD = 0.2%, range = 99.6% – 100%) and the average perception score was 98.5% (SD = 1.1%, Range = 96.5% – 99.4%).

Table 5.6

Summary of NS KPD Scores

Section	k	M	SD	Range
All	352	349.5	1.38	348 – 351
Production	217	216.67	0.52	216 – 217
Task 1 – Picture Naming	154	153.67	0.52	153 – 154
Task 2 – Nonword Reading	63	63.00	0.00	63 – 63
Perception	135	132.83	1.17	131 – 134
Task 3 – Pronunciation Judgment*	72	70.17	0.75	69 – 71
Task 4 – Identification	63	62.67	0.52	62 – 63

Reliability

This section details the reliability of the KPD, including estimates of internal consistency for all parts of the KPD and estimates of inter-scorer agreement for the production section.

Internal Consistency. Internal consistency estimates (Cronbach's alpha) for the KPD are in Table 5.7. Across the board, most estimates exceed recommended thresholds for low-stakes testing. The lowest reliability estimate, .65, comes from an item-level analysis of the Identification task. Many of the alpha values obtained are similar to those from the pilot study (Chapter 2), and once again it appeared that item parcels sacrifice little in terms of internal consistency.

Table 5.7

Internal Consistency of the KPD

Section	k	alpha (Items)	k	alpha (Parcels)
All	352	.92	56	.91
Production	217	.83	28	.78
Task 1 – Picture Naming	154	.72		
Task 2 – Nonword Reading	63	.74		
Perception	135	.89	28	.89
Task 3 – Pronunciation Judgment*	72	.89		
Task 4 – Identification	63	.65		

Note. *Excluding filler items.

The Rasch person reliability figures for the KPD production and perception sections (Table 5.8) are similar to corresponding Cronbach's alpha estimates. Little to no reliability in distinguishing overall production and perception ability appears to be lost when parceling items.

Table 5.8

Rasch Person Reliability Estimates for the KPD

Section	k	Person Reliability (Items)	k	Person Reliability (Parcels)
Production	217	.82	28	.78
Perception	135	.90	28	.90

Production Items – Inter-Scorer Agreement. For each individual item assessing production of phonemes (including Task 1 – Picture Naming and Task 2 – Nonword Reading), percent agreement, Fleiss' Kappa, and Gwet's AC1 were computed based on the scores assigned by the seven scorers. Summary statistics for these agreement indices, based on all 217 items, are presented in Table 5.9.

Table 5.9

Inter-Scorer Agreement for Individual Production Items

Index	Mean	SD	Range
Percent Agreement	85.39	15.55	30 – 100
Fleiss' Kappa	0.48	0.40	-0.11 – 1.00
Gwet's AC1	0.93	0.09	0.49 – 1.00

While the average Fleiss' Kappa indicates only moderate agreement among coders, the average percent agreement and Gwet's AC1 tell a different story. Due to a high prevalence of intelligible pronunciation (i.e., correct responses), the reduced negative bias of Gwet's AC1 statistic better reflects the simple percent agreement. Figure 5.11 shows the distribution of the three agreement indices across items. While many items have Kappa values interpretable as

"none" or "slight" (Landis & Koch, 1977), the large majority of items have AC1 values associated with substantial or near-perfect agreement among raters.

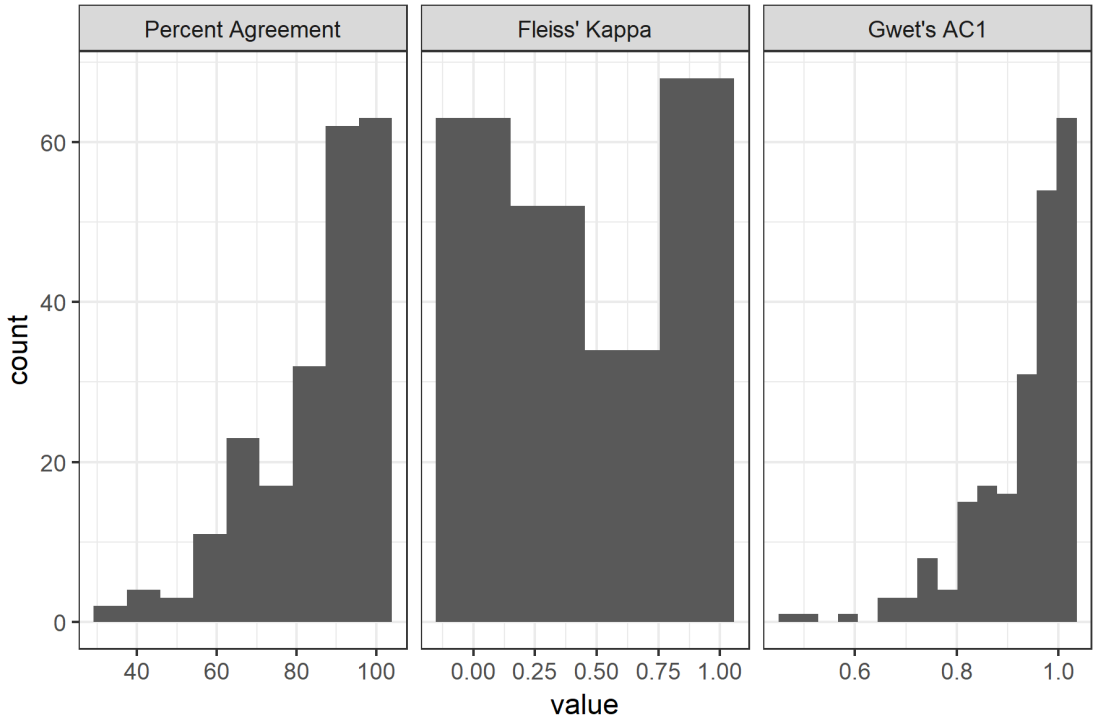


Figure 5.11. Histograms of item agreement indices for individual items based on all seven scorers.

Given the high prevalence of correct responses and a closer alignment of Gwet AC1 values and the intuitive percent agreement figures, Gwet's AC1 values were examined in further detail, which revealed that three items had less than substantial agreement, following Landis & Koch's (1977) guidelines per Gwet (2008): T2_08 (target phoneme: /t/), T2_23 (target phoneme: /ʨ*/), and T1_30-3 (target phoneme: /ʨ*/; the ㅉ in 왼쪽, *left*). An additional 18 items (8% of all items) had AC1 values between 0.60 and 0.80, indicating substantial agreement. All other values obtained for individual items exceeded 0.80, indicating almost perfect agreement.

Production Parcels – Inter-Scorer Reliability. The mean ICCs across all 28 item parcels are shown in Table 5.10. The ICC focused on consistency of ratings and the ICC focused

on absolute agreement in terms of parcel scores were similar in magnitude and dispersion. Parcel scores across the seven raters ranged from essentially no agreement to perfect agreement.

Table 5.10

Inter-scorer Reliability for Item Parcel Scores

Index	Mean	SD	Range
ICC_C	0.50	0.22	-0.02 – 1.00
ICC_A	0.48	0.23	-0.02 – 1.00

Table 5.11 contains ICC estimates for each phoneme parcel. Many of the ICC values fell into the 'poor' range, with 12 phoneme parcels in the 'moderate' or 'good' range. In some cases, closer inspection of the data revealed extremely high prevalence of high accuracy rates for some phonemes leading to low variability among the 20 test-takers, which would result in low ICC values despite generally similar scores being given to each examinee. For example, the phoneme /m/ (ㅁ) had an ICC_A and ICC_C of -0.02, the lowest among all phonemes and a figure that essentially indicates no interrater reliability. Out of the 140 parcel scores assigned to the 20 test-takers by the seven scorers, 129 were 100%, 10 were 87.5%, and one was 75%. The standard deviations of scores for the 120 /m/ parcel scores was 3.8%.

Table 5.11

Inter-Scorer Reliability/Agreement Indices for all Parcel Scores and Diagnostic Flags

	Parcel Accuracy Scores		Diagnostic Flags		
Phoneme	ICC_A	ICC_C	Percent Agreement	Fleiss' Kappa	Gwet's AC1
ㄱ /k/	0.61	0.64	80	0.10	0.91
ㅋ /k^h/	0.66	0.69	80	0.47	0.91
ㄲ /k*/	0.46	0.51	75	0.17	0.87
ㄷ /t/	0.61	0.62	60	0.35	0.73
ㅌ /t^h/	0.48	0.49	90	0.38	0.95
ㄸ /t*/	0.83	0.84	90	0.84	0.97
ㅂ /p/	0.25	0.30	65	0.12	0.82
ㅍ /p^h/	0.68	0.68	80	0.60	0.92

Table 5.11 (cont'd)

ㅃ /p*/	0.56	0.57	90	-0.01	0.97
ㅈ /tɕ/	0.14	0.15	100	1.00	1.00
ㅊ /tɕʰ/	0.45	0.48	80	0.15	0.91
ㅉ /tɕ* /	0.55	0.57	45	0.44	0.63
ㅅ /s/	0.20	0.35	60	0.12	0.84
ㅆ /s*/	0.34	0.42	50	0.26	0.73
ㅎ /h/	1.00	1.00	100	1.00	1.00
ㄹ /l/	0.44	0.48	80	0.03	0.93
ㅁ /m/	-0.02	-0.02	100	1.00	1.00
ㄴ /n/	0.66	0.66	70	0.26	0.84
ㅇ /ŋ/	0.36	0.41	80	0.03	0.93
ㅏ /ɑ/	0.57	0.56	100	1.00	1.00
ㅣ /i/	0.69	0.69	100	1.00	1.00
ㅖ /ɛ/	0.86	0.87	100	1.00	1.00
ㅓ /ʌ/	0.45	0.49	50	0.25	0.77
ㅗ /o/	0.37	0.38	100	1.00	1.00
ㅜ /u/	0.16	0.16	100	1.00	1.00
ㅡ /ɯ/	0.27	0.30	100	1.00	1.00
/w/	0.48	0.51	70	0.03	0.88
/j/	0.61	0.63	80	0.50	0.92

Production Parcels – Identification of Diagnostic Weaknesses across Scorers. Parcel scores from each scorer were dichotomized using a 75% accuracy threshold, where scores below the threshold were flagged as targets requiring further instruction (see Chapter 3 for discussion of this approach). Summary statistics for item parcel diagnostic flags agreement are contained in Table 5.12. Overall, the average agreement of diagnostic classifications across phonemes was 82.25%, which yielded an average Fleiss' kappa that would be considered moderate and a Gwet's AC1 that would be considered near perfect following Landis & Koch (1977).

Table 5.12

Inter-Scorer Agreement for Diagnostic Flags

Index	Mean	SD	Range
Percent Agreement	81.25	17.35	45 – 100
Fleiss' Kappa	0.50	0.39	-0.1 – 1.00
Gwet's AC1	0.91	0.10	0.63 – 1.00

Table 5.11 also contains diagnostic flag agreement index values for each phoneme parcel. As with the agreement index values for individual items, percent agreement and Gwet's AC1 point in the same direction; toward rather high levels of inter-scorer agreement for most phonemes. Due to a preponderance of phonemes not being flagged diagnostically, the same phenomenon of large, negative bias in Kappa values is present here, too. Informally, I observed that many of the phoneme parcels with relatively lower agreement also tended to be those phonemes which were among the most difficult, on average, in the full field-testing sample.

Item Analyses

In this subsection, analyses of individual items and parcels utilizing both CTT and Rasch analyses are presented in detail. More detail is provided for item parcels, as these are the primary units of score interpretation and use for the KPD (i.e., learners and teachers will make instructional decisions based on phoneme difficulties, not difficulties with individual items on the KPD). Finally, the item and parcel stats (CTT only; sample size too small for Rasch analysis) of NSs are covered in brief.

CTT Item Analyses. Item analyses for individual items and item parcels follow.

Individual Items. Item analyses based on all 198 test-takers for perception and production items indicated a generally low level of item difficulty and minimal levels of discrimination (point-biserial). For production items, the mean item difficulty was 0.93 (k = 217, SD = 0.10, range = 0.48 – 1.00), with 30 items answered correctly by all 198 test-takers. The most difficult items tended to target tense consonants, though the seventh most difficult item targeted the vowel /ʌ/. Mean discrimination for the production items where at least one examinee was scored as 0 was 0.14 (k = 135, SD = 0.11, range = -0.09 – 0.46). Production items with poor discrimination tended to have very low difficulties, with nearly all examinees having earned

scores of 1, while items with stronger discrimination values tended to be relatively more difficult.

For perception items, mean item facility was 0.81 (SD = 0.22, range = 0.14 – 1.00), with 14 items answered correctly by all learners. The most difficult items targeted the phonemes /s, s*/, and higher difficulty items were generally diverse in targets, including consonants, vowels, and glides. The easiest items tended to be found in Task 4, and targeted phonemes such as /w, o, e, l, m, n/. Mean discrimination, based on 121 items, was 0.24 (SD = 0.14, range = -0.17 – 0.58). Perception items presenting at least a relatively moderate degree of difficulty tended to have better discrimination, similar to the production items, with the notable exception of item T3_34 targeting /s*/, which was the most difficult and least discriminating item (d = -0.17). Complete CTT (and Rasch) item statistics based on all 198 test-takers are available in Appendix J for all individual production (Table J1) and perception items (Table J2).

Parcels. For CTT parcel analyses, all raw parcel scores were converted to percentages. The mean production parcel difficulty was 90.7% (SD = 8.5%, range = 67.2% – 99.7%). For perception parcels, mean difficulty was 80.1% (SD = 10.8%, range = 56.22% – 97.22%). Parcel difficulty statistics are visually displayed in Figure 5.12. As can be seen in the figure, production parcels were generally easier than perception parcels. The most difficult phonemes to produce were tensed consonants, which were also among the most difficult phonemes to perceive. The easiest phonemes to produce included cross-linguistically common vowels like /ɑ, i/ and consonants like /m, h/. Some sounds were noticeably more difficult to perceive than produce, such as /s, s*/. For parcel discrimination, production parcels had a mean discrimination (*r*) of 0.29 (SD = 0.13, range = 0.08 – 0.54). Consonants including the tenseness feature and the nasal /ŋ/, the vowel /ʌ/, and the glide /j/ had the strongest discrimination while other vowels tended to

have low discriminatory power. Perception parcels had a mean discrimination of 0.45 (SD = 0.15, range = 0.08 – 0.67). While vowels such as /ɑ, i, u/ had relatively lower discrimination, similar to production parcels, there were no clear patterns in terms of perception parcels with strong discriminatory power; a wide range of phonemes had high discrimination values. Complete parcel statistics for production and perception phonemes are in Table 5.13 and Table 5.14, respectively.

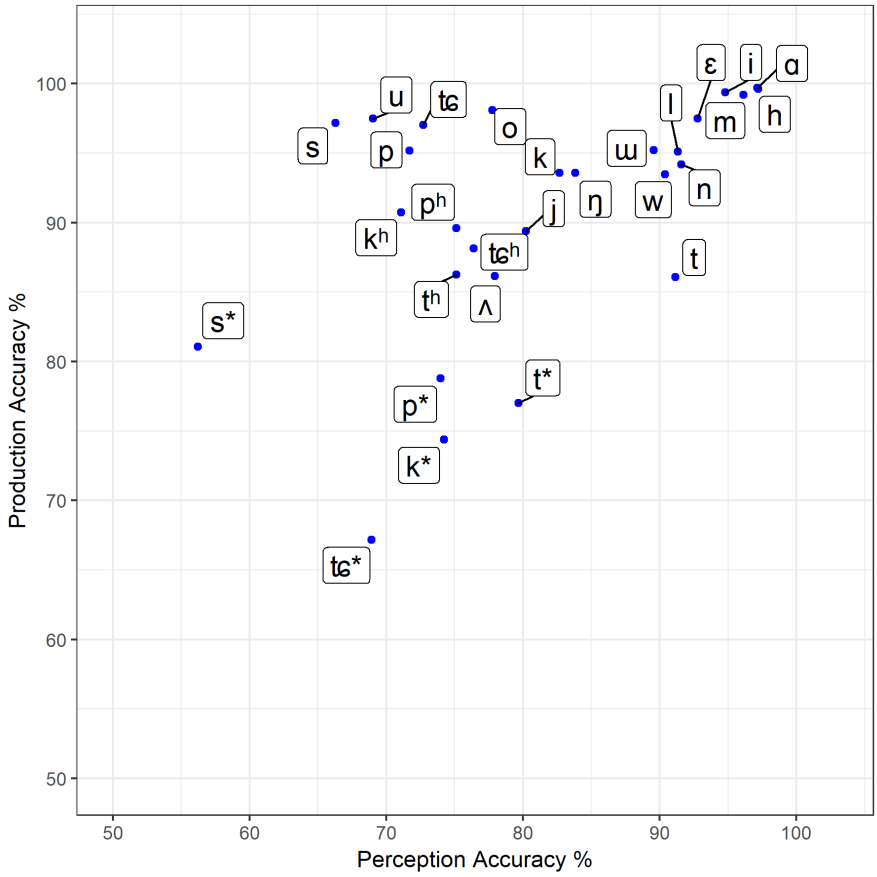


Figure 5.12. Average accuracy (inverse of difficulty) for each phoneme parcel on the production (y-axis) and perception (x-axis) sections of the KPD.

Table 5.13

Production Parcel Statistics

Phoneme	k	Mean	SD	Min	Max	Mean %	SD %	*r**	Rasch Measure	Rasch S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
ㄱ /k/	14	13.10	1.10	9	14	94	8	.32	0.52	0.07	1.04	0.35	0.95	-0.28
ㅋ /kʰ/	6	5.44	0.89	1	6	91	15	.30	0.25	0.09	1.09	0.59	1.12	0.71
ㄲ /k*/	4	2.97	0.97	0	4	74	24	.45	0.34	0.08	0.98	-0.24	0.95	-0.43
ㄷ /t/	9	7.75	1.23	4	9	86	14	.38	0.76	0.07	0.96	-0.34	0.97	-0.19
ㅌ /tʰ/	5	4.31	1.05	0	5	86	21	.34	0.3	0.08	1.11	0.83	1.08	0.51
ㄸ /t*/	4	3.08	1.07	0	4	77	27	.54	0.71	0.08	0.93	-0.64	0.83	-1.32
ㅂ /p/	7	6.66	0.63	4	7	95	9	.32	-0.33	0.12	0.93	-0.45	0.82	-0.9
ㅍ /pʰ/	5	4.48	0.85	1	5	90	17	.38	0.25	0.09	0.96	-0.22	0.88	-0.61
ㅃ /p*/	4	3.15	1.12	0	4	79	28	.55	0.72	0.07	0.92	-0.71	0.85	-0.95
ㅈ /tɕ/	8	7.76	0.59	4	8	97	7	.23	-0.06	0.13	0.94	-0.16	0.77	-0.81
ㅊ /tɕʰ/	4	3.53	0.80	0	4	88	20	.33	0.05	0.1	1.03	0.27	0.96	-0.15
ㅉ /tɕ* /	4	2.69	0.94	0	4	67	23	.44	0.93	0.09	0.94	-0.56	0.93	-0.68
ㅅ /s/	10	9.72	0.57	7	10	97	6	.14	-0.13	0.13	1.05	0.32	1.04	0.26
ㅆ /s*/	7	5.67	1.12	2	7	81	16	.38	0.75	0.08	1.07	0.67	1.02	0.21
ㅎ /h/	4	3.98	0.16	2	4	100	4	.20	-1.29	0.45	0.97	0.24	0.33	-0.2
ㄹ /l/	12	11.41	0.93	8	12	95	8	.14	0.5	0.08	1.13	0.9	1.31	1.63
ㅁ /m/	8	7.93	0.27	6	8	99	3	.11	-1.25	0.27	0.99	0.06	0.78	-0.44
ㄴ /n/	10	9.42	0.84	6	10	94	8	.24	0.28	0.09	1.03	0.27	0.96	-0.21
ㅇ /ŋ/	10	9.36	1.15	3	10	94	12	.38	0.36	0.07	0.92	-0.45	1.00	0.06
ㅏ /ɑ/	14	13.95	0.21	13	14	100	1	.19	-1.53	0.34	0.97	0.00	0.68	-0.71
ㅣ /i/	15	14.90	0.33	13	15	99	2	.17	-0.87	0.22	0.99	0.04	0.94	-0.06
ㅔ /ɛ/	9	8.77	0.45	7	9	97	5	.09	-0.75	0.17	1.05	0.4	1.41	2.21
ㅓ /ʌ/	5	4.31	0.87	1	5	86	17	.41	0.16	0.09	0.88	-1.06	0.79	-1.58
ㅗ /o/	12	11.77	0.56	9	12	98	5	.19	-0.03	0.13	1.00	0.08	0.95	-0.08
ㅜ /u/	4	3.90	0.32	2	4	97	8	.17	-1.22	0.23	1.01	0.13	1.09	0.38
ㅡ /ɯ/	4	3.81	0.43	2	4	95	11	.20	-0.59	0.17	0.97	-0.13	0.91	-0.37
/w/	10	9.34	0.84	7	10	93	8	.12	0.59	0.09	1.19	1.66	1.44	2.89
/j/	9	8.05	1.04	5	9	89	12	.38	0.56	0.08	0.99	-0.03	0.95	-0.34

Note. *Parcel-total correlation with parcel dropped from the total score.

Table 5.14

Perception Parcel Statistics

Phoneme	k	Mean	SD	Min	Max	Mean %	SD %	*r**	Rasch Measure	Rasch S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
ㄱ /k/	6	4.96	0.90	3	6	83	15	.45	0.60	0.10	1.06	0.70	1.03	0.27
ㅋ /kʰ/	4	2.84	0.92	0	4	71	23	.59	0.04	0.09	0.91	-0.87	0.89	-1.01
ㄲ /k*/	4	2.97	0.83	0	4	74	21	.58	-0.04	0.10	0.87	-1.24	0.85	-1.36
ㄷ /t/	6	5.47	0.77	3	6	91	13	.50	-0.25	0.11	0.88	-1.06	0.72	-1.59
ㅌ /tʰ/	4	3.01	0.74	1	4	75	19	.52	0.30	0.11	0.95	-0.46	0.96	-0.41
ㄸ /t*/	4	3.19	0.94	0	4	80	24	.66	-0.01	0.09	0.78	-2.06	0.68	-2.47
ㅂ /p/	6	4.30	1.20	1	6	72	20	.55	0.54	0.08	0.99	-0.09	0.94	-0.54
ㅍ /pʰ/	4	3.01	0.82	1	4	75	20	.66	0.22	0.10	0.77	-2.73	0.74	-2.76
ㅃ /p*/	4	2.96	0.99	0	4	74	25	.50	0.07	0.09	1.08	0.85	1.07	0.58
ㅈ /tɕ/	4	2.91	1.07	0	4	73	27	.67	0.26	0.08	0.80	-2.09	0.76	-2.01
ㅊ /tɕʰ/	4	3.06	0.81	1	4	76	20	.64	0.31	0.10	0.83	-1.90	0.78	-2.16
ㅉ /tɕ*/	4	2.76	0.94	0	4	69	24	.53	0.37	0.09	1.01	0.17	1.00	0.05
ㅅ /s/	8	5.30	1.22	3	8	66	15	.43	1.58	0.08	1.28	2.62	1.27	2.54
ㅆ /s*/	6	3.37	1.02	1	6	56	17	.33	1.82	0.09	1.36	3.25	1.35	3.15
ㅎ /h/	4	3.89	0.32	3	4	97	08	.28	-0.96	0.23	0.97	-0.14	0.80	-0.51
ㄹ /l/	8	7.31	0.93	4	8	91	12	.33	-0.10	0.09	1.16	1.26	1.24	1.39
ㅁ /m/	6	5.77	0.48	4	6	96	08	.38	-0.79	0.16	0.93	-0.47	0.79	-0.83
ㄴ /n/	6	5.49	0.73	2	6	92	12	.33	-0.50	0.11	1.08	0.64	1.01	0.11
ㅇ /ŋ/	4	3.35	0.82	0	4	84	21	.41	-0.39	0.10	1.06	0.55	1.03	0.25
ㅏ /ɑ/	3	2.91	0.33	1	3	97	11	.20	-1.22	0.22	0.96	-0.05	0.99	0.16
ㅣ /i/	3	2.84	0.39	1	3	95	13	.08	-1.34	0.19	1.12	0.70	1.57	1.69
ㅔ /ɛ/	3	2.78	0.44	1	3	93	15	.39	-1.26	0.17	0.92	-0.62	0.87	-0.50
ㅓ /ʌ/	3	2.34	0.66	1	3	78	22	.50	0.47	0.12	0.91	-1.02	0.87	-1.22
ㅗ /o/	3	2.33	0.57	1	3	78	19	.39	0.19	0.14	1.00	0.01	1.01	0.09
ㅜ /u/	3	2.07	0.80	0	3	69	27	.27	0.24	0.10	1.26	2.54	1.25	2.25
ㅡ /ɯ/	3	2.69	0.61	0	3	90	20	.43	-0.90	0.13	0.95	-0.34	1.01	0.11
/w/	8	7.23	0.85	4	8	90	11	.44	-0.12	0.10	1.03	0.28	1.15	1.09
/j/	10	8.02	1.14	5	10	80	11	.58	0.87	0.08	0.99	-0.08	0.98	-0.16

Note. *Parcel-total correlation with parcel dropped from the total score.

Rasch Item Analyses. Rasch item analyses for individual items and item parcels follow.

Individual items. Figure 5.13 shows the relationships between item difficulty and person ability for the production and perception items through plots referred to as Wright maps (or variable plots). In each plot, the logit scale, which is used to describe both item difficulty and person ability, is indicated on the y-axis. The left side of each Wright map shows the distribution of person ability estimates, and the right side shows the distribution of item difficult estimates where items located higher up are more difficult. Where an item and a person are parallel on the map, that person has .50 odds of responding correctly to that item. As the Wright maps indicate, relatively few production items presented much of a challenge for most learners on average, but for perception roughly a third to a half of item difficulties were in the range where many learners would find them challenging on average. Among production items, those targeting tense consonants were frequent at the higher end of the item difficulty continuum, and items targeting vowels such as /ɑ, i, ɛ/ were common at the lower end. For perception items, items targeting /s, s*/ and aspirated stop consonants were common at the higher end while vowels and glides were common at the lower end.

Considering item fit, as previously mentioned there were relatively few issues with infit for either production or perception items. Among production items with outfit issues, many of the overfitting (outfit < .70) items were relatively easy and targeted vowels such as /ɑ, i, ɛ, o/ and the consonants /h, m/; examinees with abilities substantially greater than the difficulty of these items rarely produced them inaccurately. Underfitting items (outfit > 1.3) varied in their difficulty and target phonemes, but there was a noticeable preponderance of glides, nasals, and the liquid /l/ among underfitting production items. For perception items, the overfitting items were on the easier side but otherwise had little in common. Underfitting items varied in

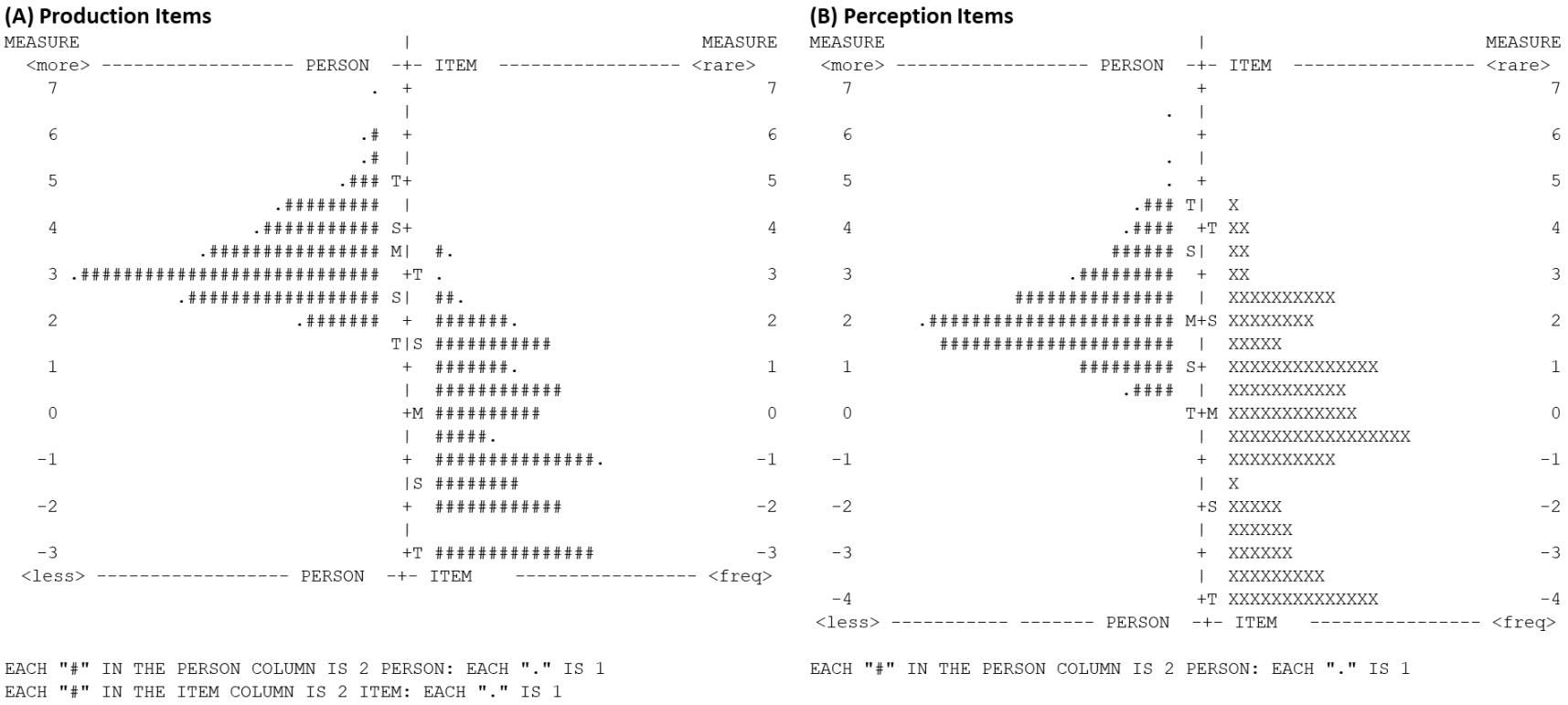


Figure 5.13. Wright maps for the KPD (A) production (Task 1 and Task 2) and (B) perception (Task 3 and Task 4) individual items. The left columns on each plot show test-taker ability (higher = more able) and while the right columns show item difficulty (higher = more difficult.

difficulty, but some patterns did emerge in terms of targets: several items for glides (/w, y/), /m/, and fricative consonants /s, s*/ were among the 16 underfitting perception items. Complete Rasch item statistics for production and perception items are found in Appendix J (Table J1 and Table J2, respectively).

Parcels. Figure 5.14 plots the parcel difficulties, in Rasch-scaled logits, of each phoneme in perception (x-axis) and production (y-axis). Phonemes closer to the diagonal had comparable perception and production difficulties, while those further from the diagonal were easier or harder in one modality. There were many similarities between the Rasch measures and the percentages based on observed scores (Figure 5.12). For example, among the easiest phonemes in both modalities were / ɑ, h, m/. The most difficult phonemes to produce were the tensed consonants, and also /t/. The most difficult phonemes to perceive were /s, s*/, but these were not the most difficult to produce (though /s*/ was among the most difficult). Complete Rasch parcel statistics for the production and perception sections of the KPD are contained in Table 5.12 and Table 5.13.

Figures 5.15 and 5.16 provide conventional Wright maps, showing the distribution of average parcel difficulties relative to test-taker abilities, and expected score category keyforms, which show the test-taker ability ranges associated with scores on each parcel (ranges divided by colons ":"), for production and perception parcels, respectively. Much like the observed score parcel analyses, the two Wright maps reveal that production parcels were relatively easy for most test-takers while perception parcels were more likely to present a challenge. The category keyforms allow for direct comparisons of scores across parcels, despite that many parcels contained a different total number of items. For instance, a learner with an overall Korean phoneme production ability of 1.0 logits would be expected to score a 2 out of 4 on /ʧ*/ and 15

out of 15 on /i/. Similarly, a learner with an overall phoneme perception ability of 1.0 logits would be expected to score a 3 out of 6 on /s*/ and 3 out of 3 on /i/. Thus, given the high abilities of examinees relative to parcel difficulties, individuals scoring lower on a particular phoneme would often be considered unexpected from the Rasch perspective. On the note of differing numbers of items, some parcel total scores were never achieved. For example, for the production parcel for /ɑ/ (Figure 5.15, near the bottom) only aggregated scores of 13 or 14 were observed, effectively rendering it a dichotomous item with only one threshold. This phenomenon is examined in closer detail in the following section.

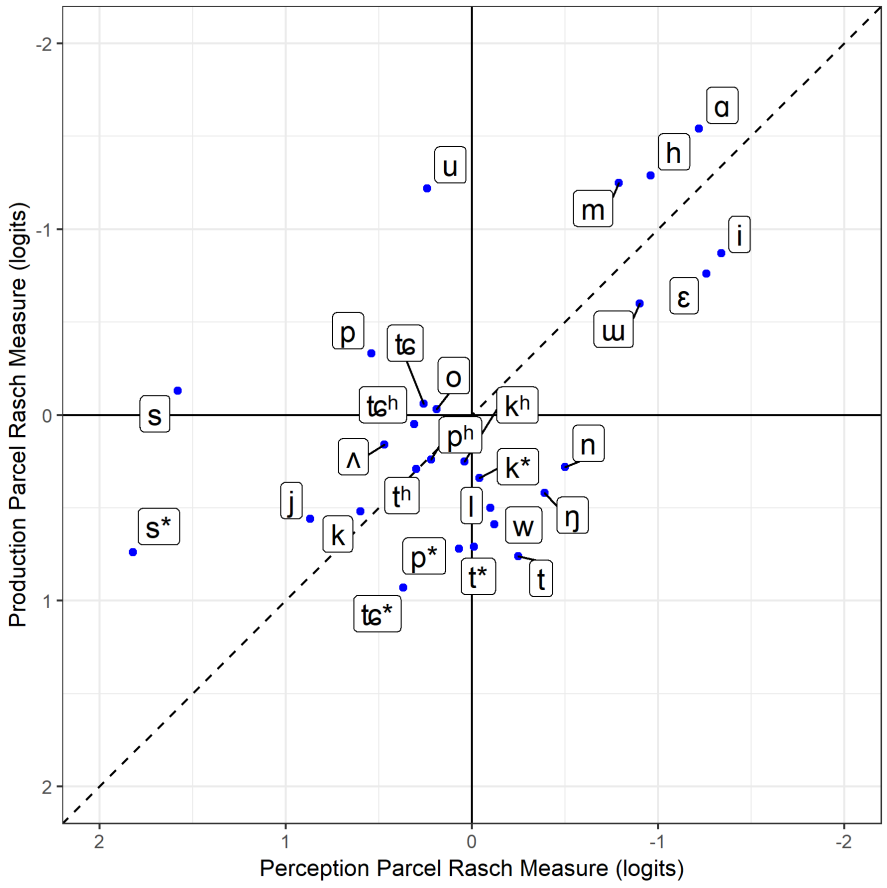
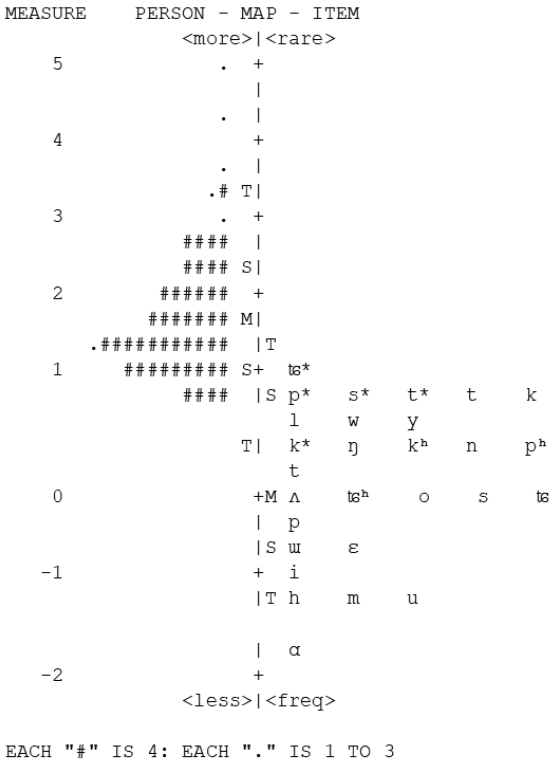


Figure 5.14. Rasch item difficulty measures for each phoneme parcel on the production (y-axis) and perception (x-axis) sections of the KPD. Axes inverted; easier parcels are located upward (production) and rightward (perception).

(A) Production Parcels - Wright Map



(B) Production Parcels - Category Keyform

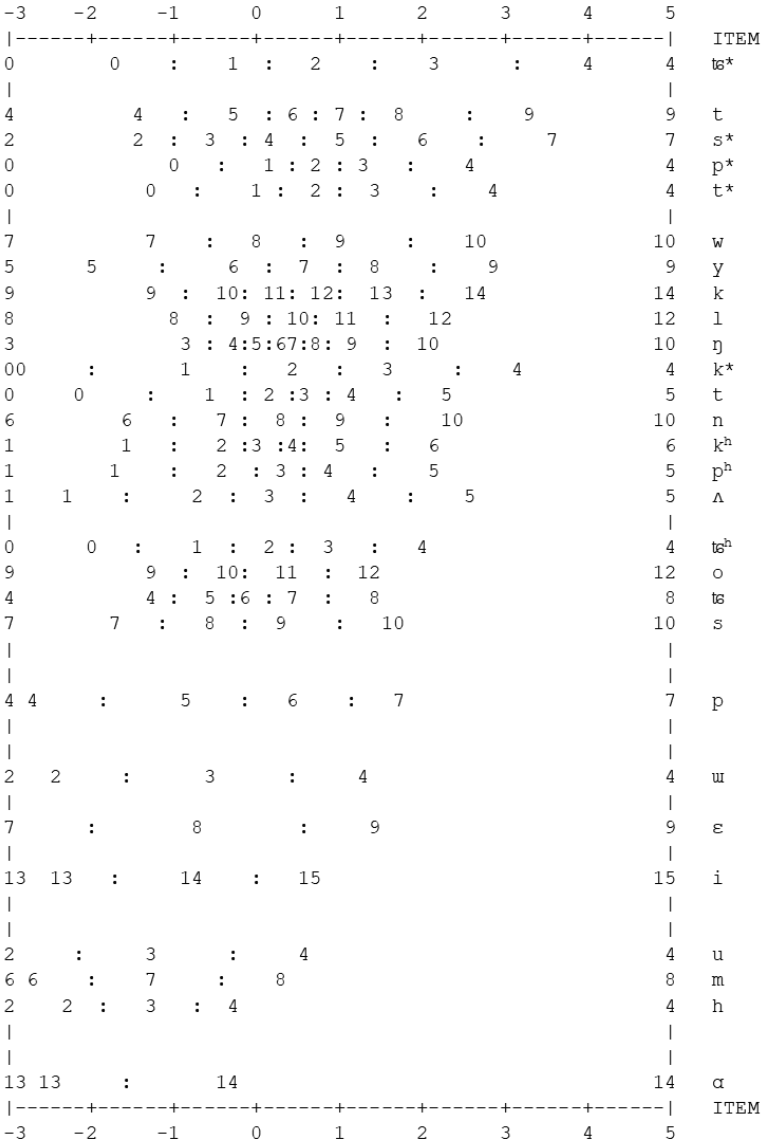
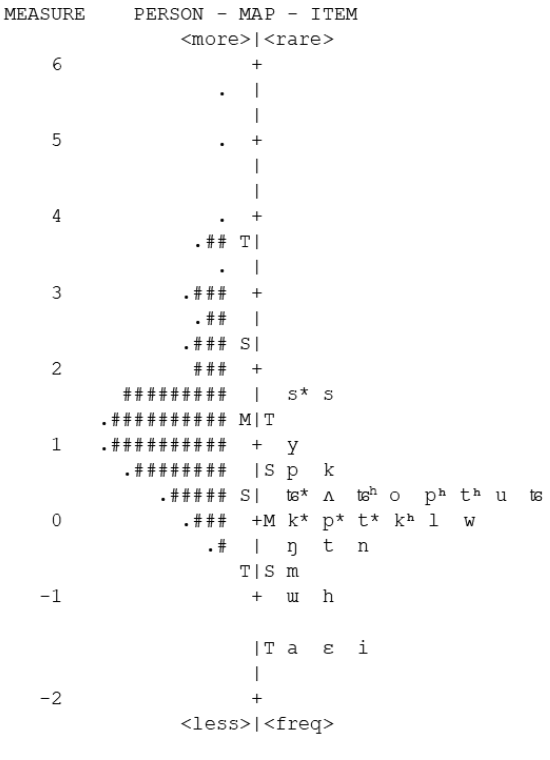


Figure 5.15. Visual summary of production parcel difficulties (A) and category thresholds (B).

(A) Perception Parcels - Wright Map



(B) Perception Parcels - Category Keyform

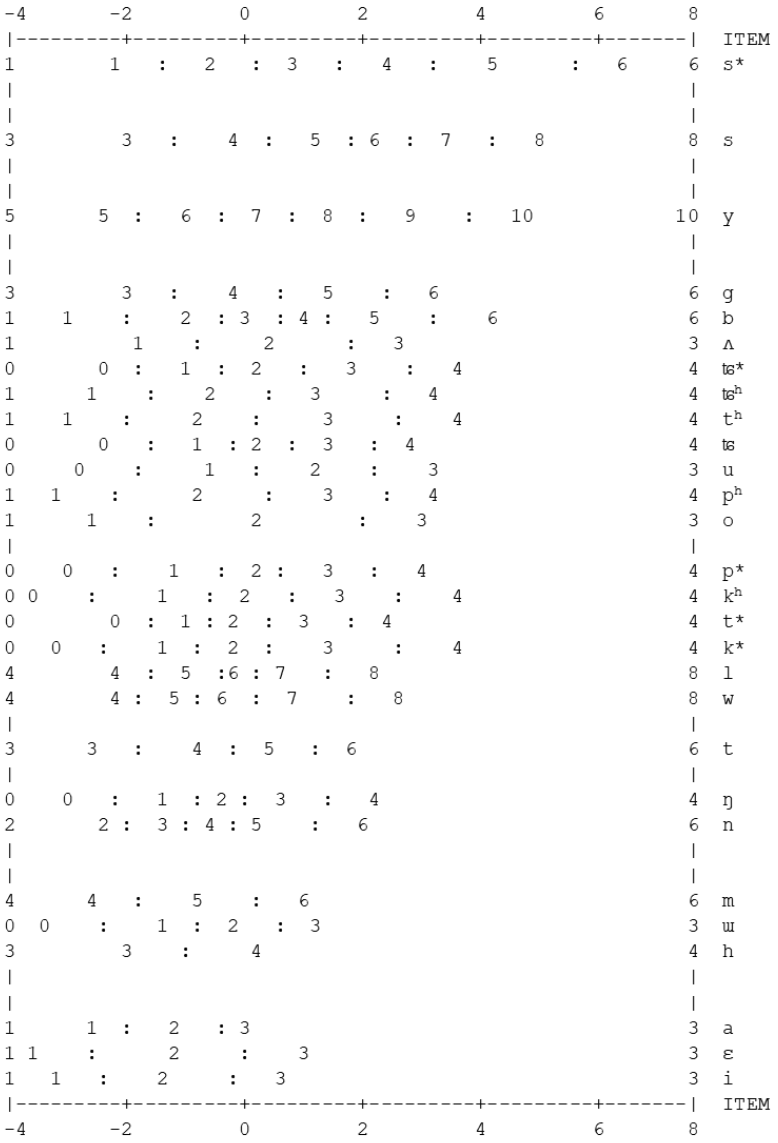


Figure 5.16. Visual summary of perception parcel difficulties (A) and category thresholds (B).

In terms of parcel fit, few phonemes in either modality exhibited any substantial misfit. For infit, no production parcels misfit and only one perception parcel demonstrated slight underfit: /s*/ infit = 1.36. For outfit, the production parcel /h/ considerably overfit (outfit = 0.33) while three parcels exhibited slight underfit: /l/ outfit = 1.31, /ʌ/ outfit = 1.41, /w/ outfit = 1.44. One perception parcel, /t*/ had slight overfit (outfit = 0.68) and two perception parcels underfit: /s*/ outfit = 1.35, /i/ outfit = 1.57. Although parcels generally had acceptable fit to the Rasch model, it is worth examining other technical qualities in further depth. Figures 5.17 and 5.18 display the item information function (IIF, blue regions) and the partial-credit step probability curves (black lines) for each phoneme parcel in production and perception, respectively. IIFs show where and how much information is gleaned about examinees, relative to the mean parcel difficulty. For example, the production parcel for /k*/ provides some information about test-takers across a range of ability, while the production parcel /m/ provides most information at lower ranges. The step probability curves represent the probability of an examinee of a given ability level obtaining a step score. As noted previously, some step categories are missing due to no examinees earning very low scores on some parcels, such as /o, s/. The curves for /k*/ in production have distinct peaks, which is generally desirable for score interpretation and indicates that test-takers with differing phoneme production abilities are likely to earn different observed scores. For many production phonemes, several step probability curves are highly overlapped or completed subsumed by other steps, e.g., /ŋ/. In these cases, differences in person ability at certain ranges are not reflected well by observed score differences. Such cases support the collapsing/combining of several categories for the purpose of measurement, not entirely unlike dichotomizing a parcel score to arrive at a diagnostic flag (i.e., 75% diagnostic flag threshold).

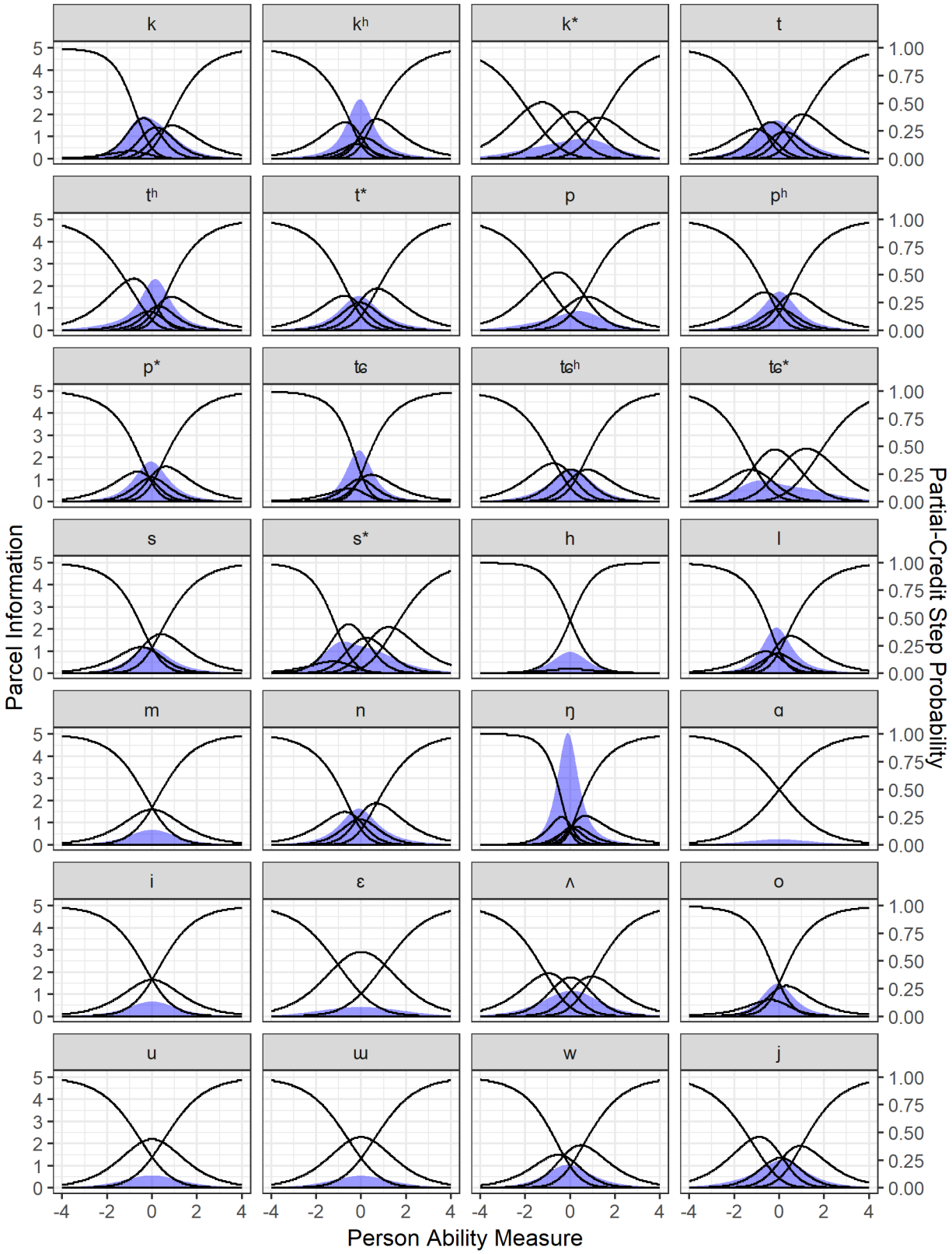


Figure 5.17. Item information and partial-credit step probability plots for production parcels.

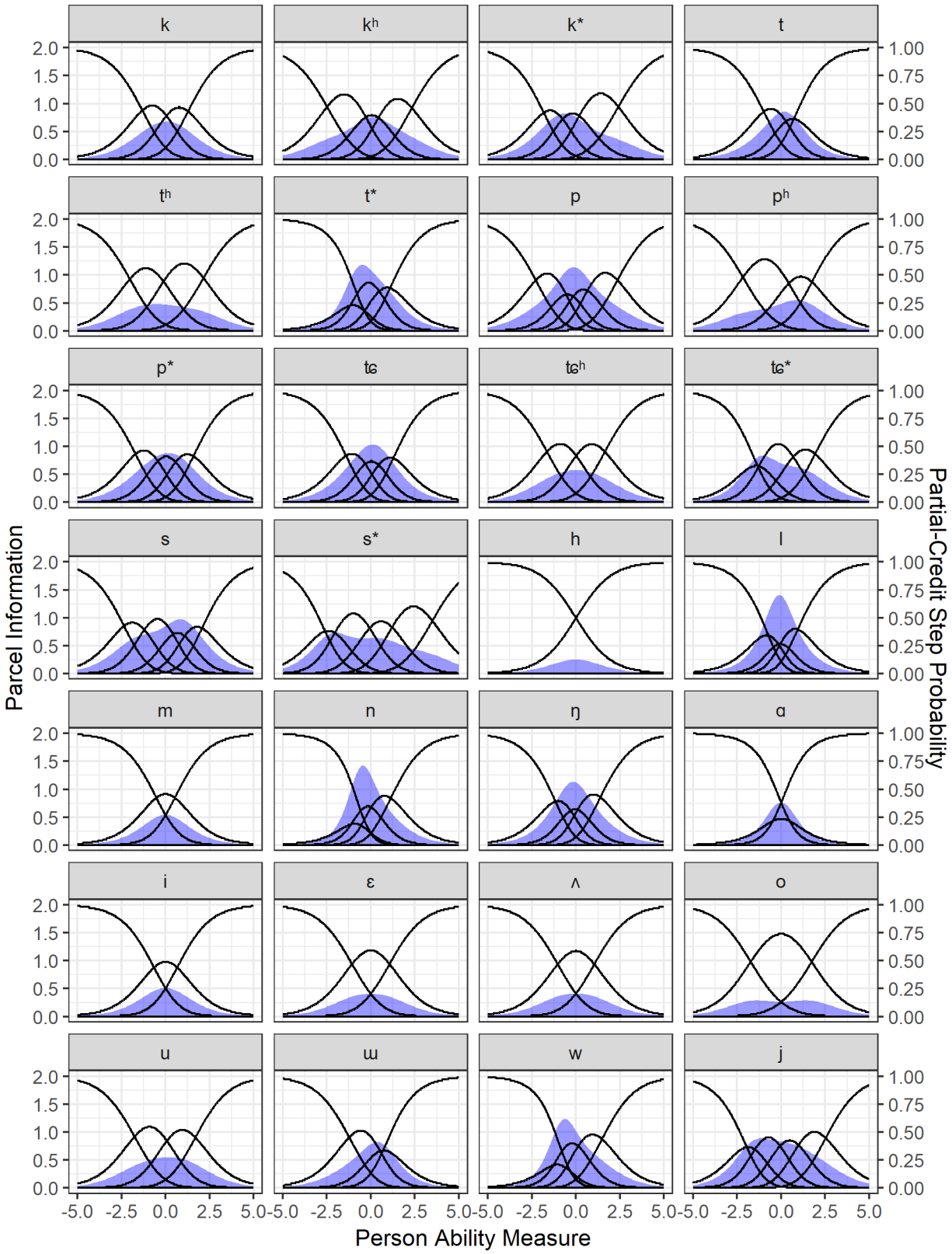


Figure 5.18. Item information and partial-credit step probability plots for perception parcels.

In general, the perception parcels provided more information across a wider range of learner abilities compared to the production parcels. Perception parcels also featured more distinct step probability curves. Much of this is due to the relatively higher difficulty of perception parcels; information is always increased where parcel difficulty is near examinee ability. On the other hand, the IIFs and step probability curves are suggestive of inflection points that might distinguish between learners who have strong control of a phoneme and those who do not. The overall production abilities of many examinees were far from the last 2-3 step thresholds for many parcels. Thus, higher scores on these parcels provided relatively little information of use, and instead more could be learned about learner abilities when they notched only middling or low scores on a parcel.

Native Speakers. The six NSs all earned scores of 1 on nearly every individual KPD item. For production items, the average item difficulty was 0.99 (SD = 0.02, range = 0.83 – 1.00). Only 2 out of 217 items were responded to incorrectly, each by just one person: T1_23-1 (the glide /j/ in 의자, *chair*) and T1_33-6 (the /n/ in the coda of the second syllable of 빨간색, *red*). Both items were found in Task 1. For perception items, the average item difficulty was 0.98 (SD = 0.10, range = 0.00 – 1.00). There were five perception items responded to incorrectly by NSs: T3_16 (4/6 correct responses, the final /k/ in 미국, *America*) , T3_34 (0/6 correct responses, the /s*/ in 접시, *dish*), T3_59 (5/6 correct responses, the /u/ in 눈, *eye*), T3_71 (4/6 correct responses, the /w/ in 원, *Korean Won*), T4_07 (5/6 correct responses, the /u/ in 우) T4_54 (5/6 correct responses, the /p^h/ in 이피).

For phoneme parcels, the average NS production score was 99.9% (SD = 0.5%, range = 98.1% – 100%) and the average perception score was 98.5% (SD = 3.9%, range = 83.3% – 100%). For production, the only parcels with less than perfect scores for all six NS participants

were /j/ (1 NS received a score of 89%) and /n/ (1 NS received a score of 90%). No NS was diagnostically flagged for any phoneme in production. For perception, the following 5 parcels had less than perfect scores for all NSs: /k/ (two NSs received scores of 83.3%), /p^h/ (1 NS received a score of 75%), /s*/ (all 6 NSs received scores of 83.3%), /u/ (1 NS received a score of 33%), /w/ (2 NSs received scores of 87.5%). One NS would have received a secondary diagnostic flag indicating difficulty hearing /u/, but they would not have received the primary flag for difficulty producing that phoneme. While these lower perception scores are not ideal, they are perhaps reflective of the high yet imperfect NS performance in speech perception research, even in favorable (i.e., quiet, lack of background noise) listening conditions (e.g., Broersma & Scharenborg, 2010; Cutler, Weber, Smits, & Cooper, 2004).

Internal Structure

Examining the internal structure of the various parts of a test can provide information on the degree to which test scores align with expectations about the relationship among (sub)constructs. For the KPD, theory strongly suggests that phoneme production and perception abilities should be related at least moderately. Mechanistic expectations of how and what knowledge and skills are elicited by the various KPD tasks, gleaned from psycholinguistic processing models, hold that scores from production tasks should be at least moderately related, and the same goes for perception tasks. Some degree of relationship among scores from all tasks would in turn be expected. Tasks that tap into orthographic knowledge (and/or sound-symbol correspondences) and tasks that tap into lexical knowledge (i.e., meanings and phonological forms of relatively common lexical items) were also expected to be correlated. At a more intricate level, scores for each phoneme in production and perception are expected to be moderately correlated, in line with theory and empirical findings from speech learning.

Production and Perception Total Score Correlations. The correlation between total Production (raw sum of correct Task 1 and Task 2 items) and Perception (raw sum of Task 3 and Task 4 items) scores was $r = .74$ ($df = 196$, $p < .001$). Figure 5.19 presents this relationship in a scatterplot.

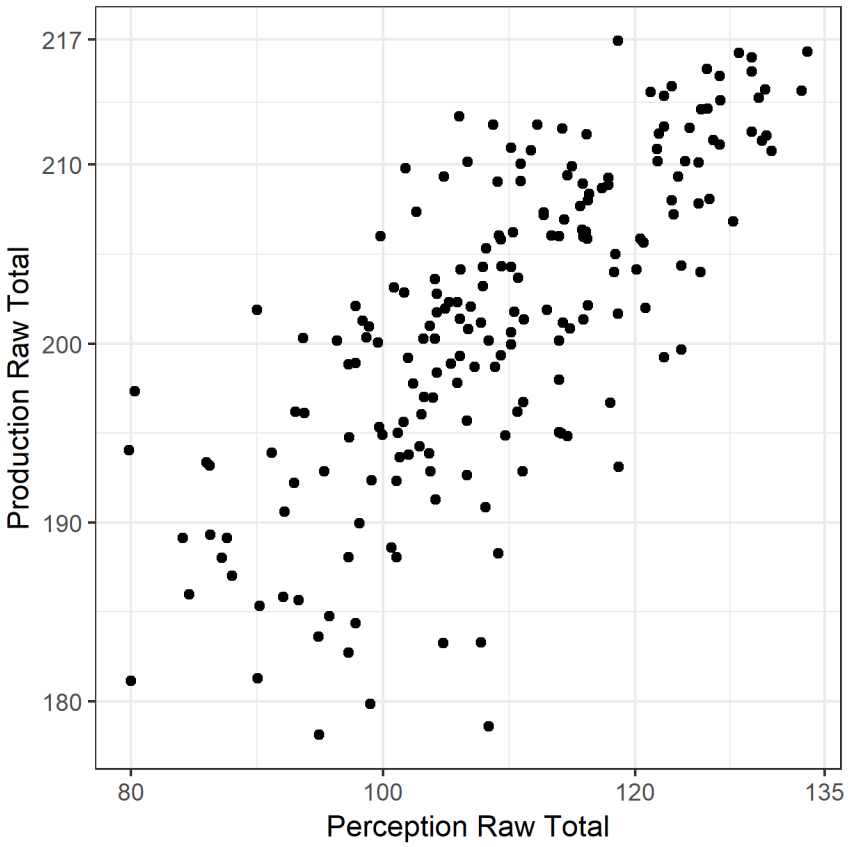


Figure 5.19. Scatterplot of production and perception raw total scores.

Task Total Correlations. I ran sum scores correlations among tasks, and correlations between each task and the total KPD score minus that task (Table 5.15). The KPD tasks largely correlated with one another and with the sum of all other tasks.

Table 5.15

Correlations Among KPD Task Sum Scores

	Task 1	Task 2	Task 3	Task 4	Total - Task
Task 1 – Picture Naming	1.00	.63	.66	.52	.71
Task 2 – Nonword Reading		1.00	.63	.59	.70
Task 3 – Pronunciation Judgment			1.00	.65	.76
Task 4 – Identification				1.00	.69

Production and Perception Phoneme Parcel Correlations. The correlation between learners' average production parcel accuracy and average perception parcel accuracy was $r = .73$ ($df = 196$, $p < .001$). This relationship is shown visually in Figure 5.20. Within each learner, the average correlation between all 28 production and perception phoneme parcels was Spearman's $\rho = 0.20$, with a standard deviation of 0.21 and a range of -.28 to 0.70. Some smaller (and small negative) individual correlations may be attributable to lack of variability (e.g., learners with very high scores in production and perception across all or most phonemes). In other cases, idiosyncratic differences in learner phonological systems may have yielded small negative correlations.

Focused at the phoneme level across all 198 learners, Table 5.16 contains the Spearman correlations between production and perception phonemes. These values ranged from -0.11 (/u/) to 0.52 (/t*/) with an average of 0.20. Phonemes with higher correlations tended to also have higher difficulties and higher discrimination or information across a wider range of examinees as shown by the CTT and Rasch analyses; very small correlations appeared to be a product of limited variability (e.g., nearly all examinees earning maximal scores for /h/).

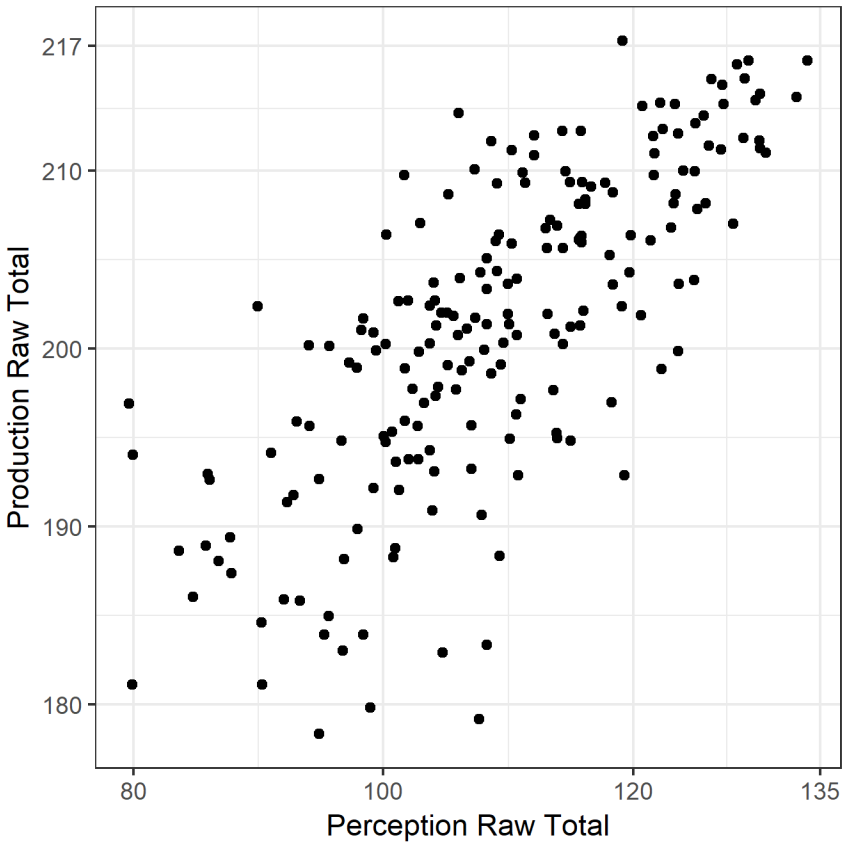


Figure 5.20. Scatterplot of production and perception parcel average accuracy scores.

Table 5.16

Phoneme Production and Perception Parcel Spearman Correlations

Phoneme	ρ	Phoneme	ρ
ㄱ /k/	0.21	ㅎ /h/	-0.04
ㅋ /kʰ/	0.31	ㄹ /l/	0.36
ㄲ /k*/	0.29	ㅁ /m/	0.07
ㄷ /t/	0.25	ㄴ /n/	0.24
ㅌ /tʰ/	0.27	ㅇ /ŋ/	0.26
ㄸ /t*/	0.52	ㅏ /ɑ/	0.04
ㅂ /p/	0.29	ㅣ /i/	0.03
ㅍ /pʰ/	0.35	ㅔ /ɛ/	0.00
ㅃ /p*/	0.39	ㅓ /ʌ/	0.31
ㅈ /tɕ/	0.27	ㅗ /o/	0.03
ㅊ /tɕʰ/	0.34	ㅜ /u/	-0.11
ㅉ /tɕ*/	0.16	ㅡ /ɯ/	0.23
ㅅ /s/	0.05	/w/	0.11
ㅆ /s*/	0.21	/j/	0.23

Discussion

In this chapter, I presented results pertaining to the measurement of Korean phoneme production and perception. I analyzed individual item and item parcel measurement models via CTT and Rasch analyses. I also reported additional analyses including descriptive (technically CTT) analyses of a small sample of NS test data and inter-scorer reliability analyses for a subset of KPD learner test data. In many ways, the results of these several analyses point in similar directions. Broadly speaking, the KPD appeared to have many desirable measurement qualities: the test was of appropriate overall difficulty for diagnostic purposes, the KPD scores and diagnostic flags were adequately reliable when considering the low-stakes of decision-making, the vast majority of items performed as intended, and the KPD sections and tasks were related in accordance with expectations. In what follows, I review these results in more detail as they pertain to the research questions.

RQ1a: How Reliable is the KPD?

By the standards of lower-stakes tests, such as classroom achievement tests, the KPD exceeds acceptable thresholds of reliability and in some cases meets the standards typically expected for high-stakes tests, such as standardized large-scale language proficiency tests. In terms of test reliability (Cronbach's alpha), the individual KPD items exceeded alpha values of .80 for both production and perception sections; the perception section (alpha = .89) approached levels of reliability more commonly associated with high-stakes standardized tests. Each KPD task also had acceptable reliability, with Task 4 – Identification showing the lowest overall level of reliability (.65). Little to no reliability was lost by parceling items according to target phonemes; perception reliability did not appreciably drop, and production reliability fell slightly to a still-respectable .78. Rasch estimates of person separation reliability, which is

normally viewed as similar to the internal consistency of Cronbach's alpha, told a nearly identical story, as expected.

High levels of internal consistency might naturally be expected for long tests. However, as mentioned, even collapsing 100+ or 200+ items into 28 parcels, some of which had no observations at lower parcel scores, and still obtaining adequately high levels of reliability provides some additional evidence in favor of the KPD's reliability. Of course, each parcel provides considerably more information about test-takers' abilities than a single dichotomously scored item, so in some ways the minimal loss of reliability is not so unexpected. Additionally, given that test reliability is maximized when items are well-targeted to the range of examinee ability, the reliability indices obtained for the generally low-difficulty KPD items and parcels is also positive in terms of the interpretation of test scores.

RQ1b: How Reliably are Production Items Evaluated by Different Scorers?

In additional to test reliability, inter-scorer (intercoder) agreement results for the human scored production section were also favorable, though not quite as robust in all indices. The classic index of inter-scorer agreement, kappa, showed very poor levels of agreement due to a high prevalence of correct responses (i.e., intelligible articulations). However, the intuitive percent agreement and Gwet's AC1, an intercoder reliability index designed to reduce bias in such contexts, found high levels of agreement among the 7 scorers at the level of individual items. According to Gwet's AC1, all items had at least moderate levels of agreement and most fell into the range of very good or nearly perfect. For parcels, ICC values were moderate on average, with some showing essentially no agreement or consistency among coders. However, this too appeared to be related to a high prevalence of correct responses/high parcel scores where a small number of slight deviations could yield a low ICC value. When parcels were

dichotomized for the purpose of assigning diagnostic flags, the average Gwet's AC1 was 0.91 and all parcels were in the range of very good to nearly-perfect agreement. In sum, this provides reasonably compelling evidence that the KPD can be scored consistently by different teachers after minimal training; levels of consistency found here are adequate for low-stakes, localized decision-making.

RQ2a: What is the Internal Structure of Test Tasks?

For overall scores (raw sums of individual items), there was a large correlation (Plonsky & Oswald, 2014) between learner production and perception abilities. The total scores for each task were also highly intercorrelated. These correlations aligned with general expectations for production and perception abilities to be substantially related (Flege, 1995; Isbell, 2017).

At the level of individual phonemes, correlations between production and perception were extant and positive, but smaller. Across all phonemes measured for all test-takers, the correlation between phoneme production and perception scores was .32, which may be interpreted as small to medium following Plonsky and Oswald (2014). Within each learner, the average correlation (Spearman's rank-order) was small at .20. Interestingly, there was substantial variation in terms of these within-learner correlations; some learners had almost no correlation or even negative correlations between their perception and production of phonemes. Some cases of no correlation seem plausibly connected to very little variation in both production and perception scores (e.g., a learner with very high scores across the board). In other cases, it may well be learner idiosyncrasies at work, though undesirable influences on measurement cannot be entirely ruled out (e.g., measurement error attenuating correlations).

At the specific phoneme level (across all learners), correlations between production and perception ranged from essentially nothing to medium-sized. Here, it was clear that the generally

easier sounds had weaker correlations due to ceiling effects/restriction of range. Otherwise, the results speak positively to expected relationships between phoneme perception and production (Flege, 1995).

RQ2b: To What Extent Do Item Difficulty Hierarchies Align with Expectations?

Specific to L2 Korean phonology, several phonemes were expected to be the most difficult to produce and among the most difficult to perceive: all tensed consonants and /l/ (Kim, 2015; Lee, Moon, & Long). All tensed consonants were indeed among the most difficult for learners to produce and perceive, on average, but somewhat surprisingly, /l/ was not. In fact, in observed score analyses, /l/ was among the most accurately produced phonemes (95%), though Rasch measures placed /l/ more in the middle of the difficulty continuum. One explanation for this, and a desirable one at that, is that the scoring criteria for the productive task – i.e., unambiguous, not necessarily native-like, intelligible pronunciation – made it possible for learners to be relatively successful with /l/, thanks to Korean lacking any other liquids or phonemes with similar qualities to /l/. In other words, even if an /l/ was substituted with a phone like [ɻ] (which is not present in Korean) by a Chinese or American English speaker, it was unlikely to be misheard, or heard with uncertainty, by the scorer. This apparent phenomenon also relates to the previous RQ, whereby the production accuracy tended to exceed perception accuracy in many cases.

Aspirated consonants were of moderate difficulty in production and perception, which also finds support in the literature (e.g., Holliday, 2014). Phonemes which were easier to produce and perceive tended to be cross-linguistically common vowels such as /i/ and consonants such as /m/, which makes intuitive sense along the lines of cross-linguistic influence and aligns generally with speech learning theories (e.g., Best & Tyler, 2007; Flege, 1995). Additionally, NSs

generally performed at ceiling for all phonemes, which was also expected. In sum, the hierarchy of item difficultly general to the sample of learners largely aligned with theory and previous findings once due consideration was given to the Intelligibility Principle-based scoring criteria for production items.

Additional Considerations

Beyond the specific RQs that motivated the analyses in this chapter, the results also motivate additional discussion and consideration of issues of measurement models and analytical approaches for diagnostic assessments such as the KPD. While both item and parcel measurement models appeared to work adequately, I favored the parcel model due to its more direct relation to the way scores were intended to be interpreted and used, even though this resulted in some minor loss of reliability in the production section. Additionally, the question of which measurement analytical approach—CTT or Rasch—is best suited for a diagnostic like the KPD with the sample available is unsettled. Both analyses yielded generally similar information, though the estimation of item difficulty differed in some cases. It was also not clear that traditional Rasch conceptualizations of overall ability and unexpectedness of observations could be clearly applied to the task of diagnostic flagging, though this avenue was not explored in depth and could not be ruled out entirely. These issues will be revisited in more depth in the Conclusion chapter.

CHAPTER 6: PRONUNCIATION PROFILES

In this chapter I focus on learner pronunciation profiles, which relates to the *explanation inference* of the KPD's proposed validity argument. I present cluster analyses of learners' production and perception scores for Korean phonemes followed by description of the pronunciation strengths and weaknesses of learner clusters. Then, I present descriptive statistics for within-cluster learner L1 backgrounds and oral proficiency. Finally, I discuss these results in relation to the relevant research question.

Research Question

The primary research question I address in this chapter is:

- RQ3: Do scores indicate distinct test-taker profiles in terms of phoneme production and perception abilities?

This question bears on the explanatory power of KPD scores. While there were general trends in phoneme difficulties in perception and production (see Chapter 5, Measurement), individual learners exhibited variation in their phoneme accuracy scores. Given the influence of learner L1 (and other known languages), proficiency, exposure to Korean, and phonological aptitude on phonological development in an L2, it is unlikely for that variation to simply be reflective of a single-path, deterministic range of L2 Korean phonological ability. Along these lines, one would also expect to see some commonalities emerge across subsets of learners, i.e., profile groupings. The emergence of several such shared profiles would offer some positive evidence that KPD is sensitive to distinct, and meaningful, differences in pronunciation difficulties.

Additionally, the identification of test-taker profiles has implications for the utilization and overall usefulness of the KPD. Namely, if nearly all learners with pronunciation difficulties

had similar profiles, or if all learners from shared L1 backgrounds had nearly identical KPD score profiles, then it would make little sense to use the KPD at all: To guide instruction or to raise awareness of a learner's pronunciation difficulties, simply knowing that a learner is struggling with pronunciation, or knowing the learner's L1 (which would predict certain pronunciation difficulties), or, better yet, knowing both, would be more than sufficient. A diagnostic test such as the KPD would not be needed.

Analysis Details

The primary analysis I used to investigate learner profiles was cluster analysis (Hastie, Tibshirani, & Friedman, 2009; Kassambara, 2017; King, 2015; Staples & Biber, 2015). In many respects, cluster analysis can be viewed as a counterpart to factor analysis (especially exploratory factor analysis and principal components analysis). Factor analysis groups variables (or items) into *factors* (that is, groups of variables or items that share an underlying construct), while cluster analysis sorts people (or other objects of interest) into *clusters* (that is, groups of people that share similar characteristics). In a data matrix where variables are columns and people are rows, factor analysis combines similar columns while cluster analysis groups similar rows.

Although the consideration of individual profiles is crucial in DLA, for the purposes of broadly considering the diversity of profiles that might emerge in KPD results, a means of finding and describing relatively common profiles is useful. Cluster analysis "provides a bottom-up way to identify *new* groups that are better defined with respect to target variables" (Staples & Biber, 2015, p. 243). My intent with the analysis was to identify groups of individual learners who shared similar diagnostic profiles. This also allowed me to consider the differences in pronunciation difficulties between these groups and their backgrounds. In this sense, I did not

seek to make claims about theoretically-motivated, generalizable profiles of pronunciation difficulties, rather I simply aimed to describe profiles that emerged among the study's sample.

Ginther and Yan (2018) innovatively applied cluster analysis to language testing data for the purpose of enriched score interpretation and decision-making. By considering the TOEFL subscores of Chinese international students at an American university, Ginther and Yan found four distinct score profiles, each of which fared differently in terms of first-year academic performance. Of note here is that their cluster analysis was able to meaningfully distinguish shared profiles among language learners of the same L1 background. Like Ginther and Yan (2018), my interest in identifying groups of learners is the interpretation and use of their test scores. I wished to consider how the groupings of learners, including those from similar backgrounds, pointed to different profiles that would lead to different instructional foci.

Due to high dimensionality in the dataset (28 phoneme parcels for each modality), the sample size in the present study (n = 198) would be considered relatively small for cluster analysis. To deal with this limitation, I adopted two strategies for dividing and paring down the data. First, I elected to run separate cluster analyses for production and perception. Second, I excluded phoneme parcels that exhibited little variation across the entire sample; details on which phonemes follow.

Cluster Analysis

Cluster analysis is used to classify *i* objects (in this case, test-takers) into groups based on (a) similarity within groups and (b) dissimilarity between groups across a set of *j* variables. If all objects are highly similar in respect to a particular variable or variables, inclusion of those variables in the analysis adds little information for classification and may inflate the level of within-groups similarity. Thus, I elected to remove phoneme scores from several phonemes from

the cluster analysis of production phonemes: ㅁ, ㅎ, ㅏ, ㅣ, ㅗ, ㅖ (/m, h, ɑ, i, o, ɛ/). These phonemes all had mean accuracy ratings > 90%, SDs ≤ 5%, and minimum accuracy ≥ 75% (i.e., no test-taker was flagged for a pronunciation weakness for these phonemes). This left 22 production phoneme parcel scores as variables in the cluster analyses.

For a cluster analysis of perception phonemes, I removed several phoneme scores (ㅎ, ㅁ, ㅏ, ㅣ, ㅖ; /h, m, ɑ, i, ɛ/) which had mean accuracy ratings > 90% and SDs ≤ 10%, indicating that most test-takers had similarly high scores and that inclusion of these phonemes in a cluster analysis would have relatively little benefit.

Carrying out cluster analyses from a strictly descriptive perspective, that is, where there was no theory on the number of clusters, I used three techniques to evaluate the most appropriate number of clusters in the data: The three techniques are described in Chapter 14 of Hastie, Tibshirani and Friedman (2009). I used *R* and support functions from the *factoextra* package (version 1.0.5, Kassambara & Mundt, 2017) for all cluster analyses. First, I conducted a hierarchical cluster analysis (HCA) using Ward's D2 criterion (which squares the input Euclidean distance matrix, Murtagh & Legendre, 2014). HCA differs from *k*-means CA primarily in that it starts from the bottom, with each individual as a cluster being joined to other highly-similar clusters. HCA yields a graphical representation of the hierarchy of similarities called a *dendrogram*. By examining forks in the dendrogram, and the distances between forks, it is possible to determine a likely number of clusters that exist in the data. The remaining techniques involve the computation of a set of *k*-means cluster analyses, usually from 2 to 10 or 15 clusters. The first of these techniques is known as the *elbow method*. This procedure involves examining a plot of total within-cluster variances for the set of *k*-means cluster solutions. Where the plot bends (i.e., where an elbow is visible) and levels off is considered a good indicator for a

suitable number of clusters, as adding additional clusters only minimally reduces the total amount of within-clusters variation. The final technique I used is called the *Gap statistic*. This statistic involves the comparison of actual data against a set of uniformly distributed data. The total within-cluster variance for the actual and uniform dataset are plotted for a set of *k*-means clusterings, much like the elbow method. Where the curve for the actual data deviates furthest from the uniform data is where clustering is likely to be most meaningful. Based on these techniques, I arrived at a suitable *k* number of clusters to include my final *k*-means cluster analyses.

Data Standardization

K-means cluster analysis depends on the distance, usually Euclidean, between observations. As such, the presence of variables that differ considerably in scale leads to cluster assignments that poorly represent patterns in the data: large-scale variables effectively drown out small-scale variables. To resolve this dilemma, it is common practice to standardize the scales of all variables entered in a cluster analysis (Steinley, 2004). Two commonly used methods for scale standardization are *z*-scores (i.e., subtracting the mean and dividing by the standard deviation) and min-max scaling (i.e., converting variable scores to percentages by dividing values by the maximum possible score). Conveniently, I had already converted the observed phoneme parcel scores for the KPD to a percentage scale to achieve tau-equivalence and easy interpretability, and I used these percentage scores in the present cluster analyses.

Results

In the following subsections, I present the results of cluster analyses on learners' production and perception KPD phoneme scores. This is followed by comparisons of production and perception clusters in terms of learner L1 and overall Korean oral proficiency.

Production Profiles

In this subsection, I present results pertaining to the identification of production parcel clusters and the description of the final solution clusters.

Determining the Number of Clusters. As an initial step, I ran a hierarchical cluster analysis, the results of which are shown as a dendrogram in Figure 6.1. Based on the dendrogram, which has color coding for five clusters, four or five clusters seemed likely; the yellow and green clusters in the middle of the plot are rather small compared to the others, and the last step going from four to five clusters has a small height.

The average within-cluster variation (sum of squares) of *k*-means cluster analyses for 1 through 10 clusters are in Figure 6.2. In this plot, there is a fairly distinct elbow at $k = 4$ clusters. After this point, reductions of within-cluster variance are fairly small.

Figure 6.3 plots the gap statistic for each of the first $k = 1$ through $k = 10$ cluster solutions. Here, the 4-cluster solution appeared to be the point at which differences in within-cluster similarity for the production parcel data became markedly different from within-cluster similarity of a simulated uniform dataset.

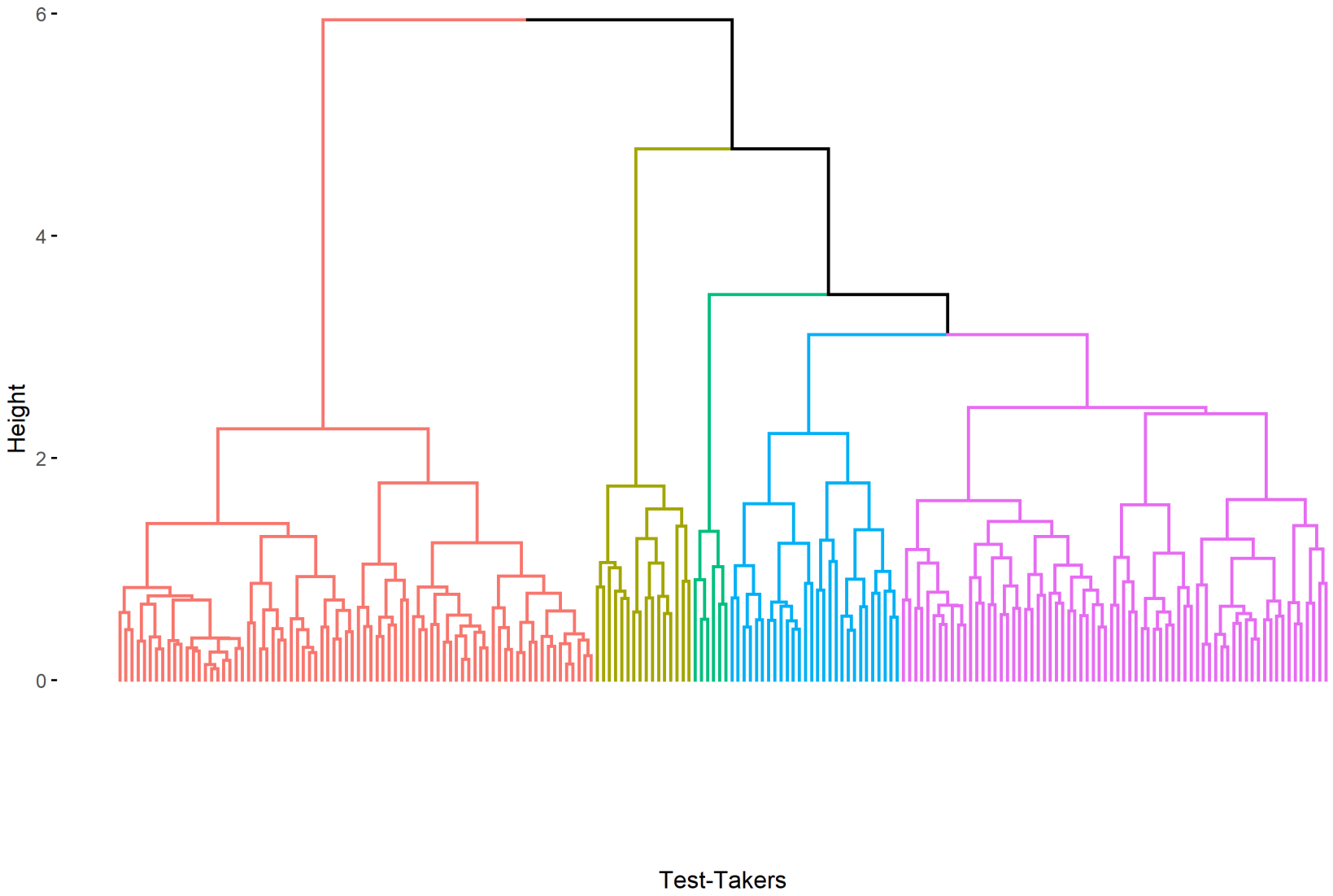


Figure 6.1. HCA dendrogram depicting suggested clustering of test-takers according to production parcel scores.

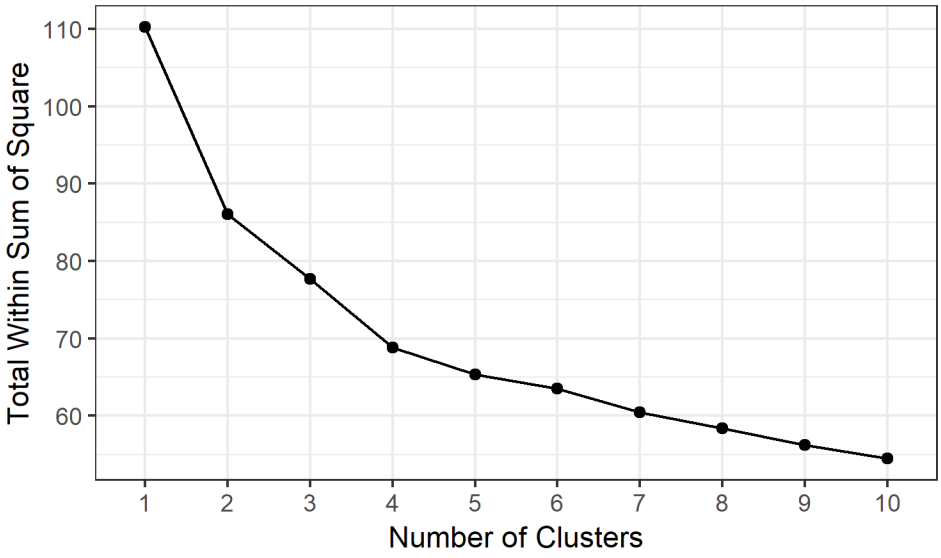


Figure 6.2. Plot of within-cluster sum of squares for k = 1..10 clusters based on production parcel scores.

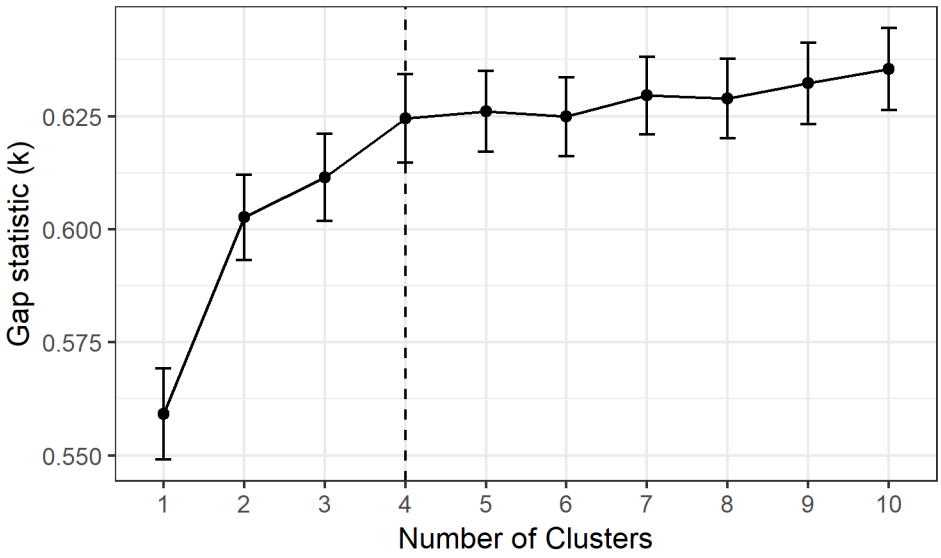


Figure 6.3. Gap statistic plot for k = 1..10 clusters based on production parcel scores. The vertical dashed line indicates where the number of clusters is optimal.

All in all, a 4-cluster *k*-means solution appeared to be sufficiently well-supported. The 4-cluster solution resulted in 19 test-takers in Cluster 1, 73 in Cluster 2, 76 in Cluster 3, and 30 in

Cluster 4. Figure 6.4 plots the four clusters in two dimensions based on a principal components analysis of the production parcel data; each test-taker's first contrast (Dim1) and second contrast (Dim2) scores are used for Cartesian coordinates. While the four clusters are not entirely distinct in this limited two-dimensional representation, some differences are visible.

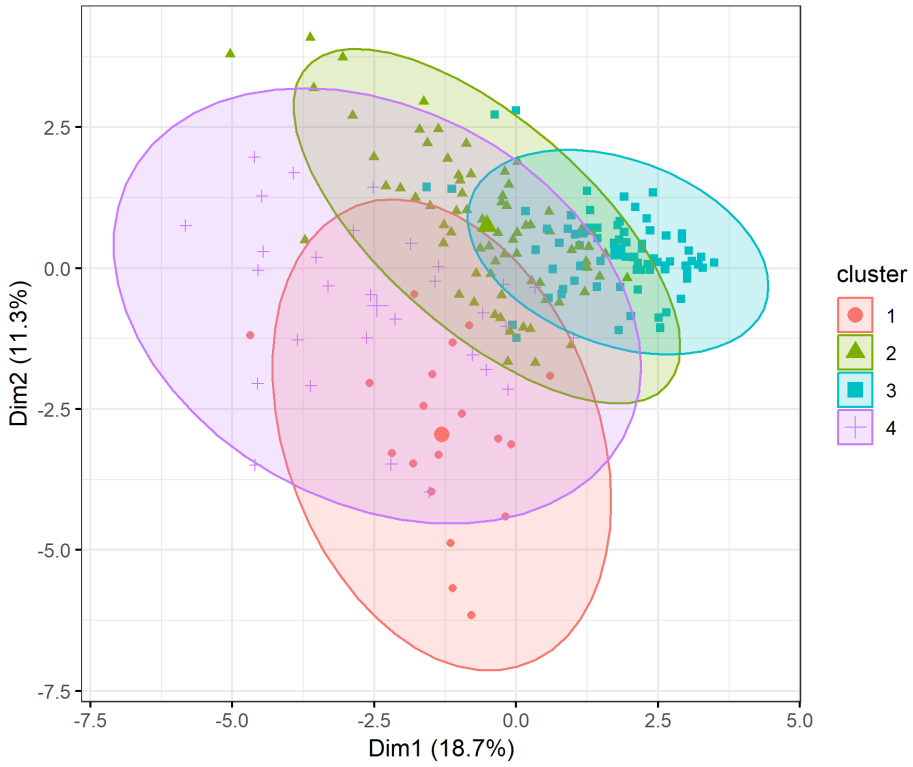


Figure 6.4. Plot of clusters along the first two principle components of the production parcel data.

Cluster Descriptions. For each cluster, I computed mean accuracy for each production parcel as well as the proportion of diagnostic flags (based on a < 75% criterion). These values are visually represented in Figure 6.5; detailed numeric values are in Table 6.1. In the Figure 6.5 heatmaps, red cells indicate very low accuracy and a high proportion of diagnostic flags, while green cells indicate high accuracy and a low proportion of diagnostic flags, with yellow representing middling accuracy and a split proportion of diagnostic flags. Based on the mean

accuracy rates and proportion of diagnostic flags, the four clusters can be summarized as follows in terms of production difficulties:

- Production Cluster 1: Difficulty with aspirated consonants, fricative and affricate tensed consonants.
- Production Cluster 2: Limited difficulties; difficulty with tensed affricate and fricative consonant (and to a lesser extent, /k*/), /ʌ/.
- Production Cluster 3: Few to no difficulties.
- Production Cluster 4: Major difficulties with tensed consonants. Some difficulty distinguishing /t, t^h, t*/, some difficulty with /ŋ, ʌ, j/.

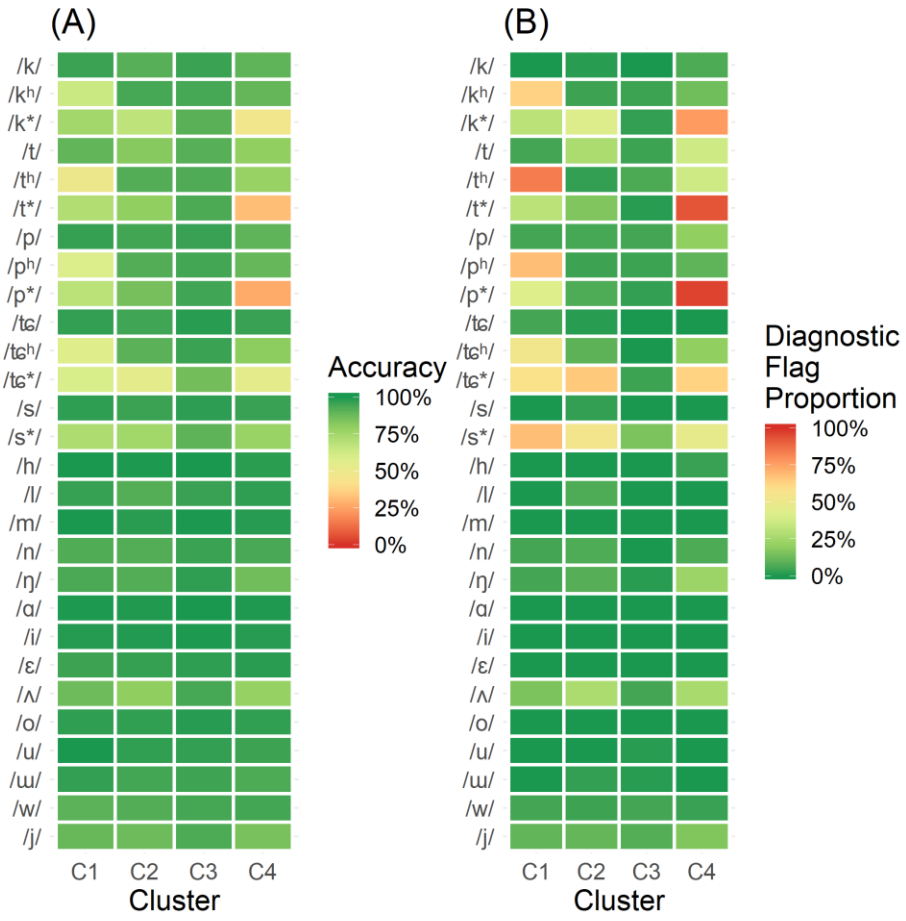


Figure 6.5. Heatmaps of phoneme production mean accuracy (A) and diagnostic flag proportion (B) by cluster.

Table 6.1

Phoneme Production Mean Accuracy and Diagnostic Flag Proportion by Cluster

	C1		C2		C3		C4	
	Acc.%	Flag%	Acc.%	Flag%	Acc.%	Flag%	Acc.%	Flag%
ㄱ /k/	96	0	91	1	96	0	90	7
ㅋ /kʰ/	64	63	95	4	95	4	89	13
ㄲ /k*/	75	32	68	42	91	3	48	77
ㄷ /t/	89	5	82	27	92	4	80	37
ㅌ /tʰ/	49	84	92	3	93	7	78	37
ㄸ /t*/	71	32	80	16	93	1	32	93
ㅂ /p/	97	5	95	5	97	5	90	20
ㅍ /pʰ/	59	68	92	4	95	4	89	10
ㅃ /p*/	68	42	86	7	95	3	27	97
ㅈ /ʧ/	97	5	95	1	99	0	97	0
ㅊ /ʧʰ/	57	53	91	10	96	0	81	20
ㅉ /ʧ*/	61	58	54	66	86	4	54	63
ㅅ /s/	98	0	96	3	98	0	97	0
ㅆ /s*/	72	68	75	53	90	16	78	47
ㅎ /h/	100	0	100	0	100	0	98	3
ㄹ /l/	97	0	92	7	97	0	98	0
ㅁ /m/	100	0	98	0	100	0	99	0
ㄴ /n/	93	5	92	7	96	0	94	7
ㅇ /ŋ/	94	5	92	8	98	1	86	23
ㅏ /ɑ/	100	0	99	0	100	0	100	0
ㅣ /i/	99	0	99	0	100	0	99	0
ㅔ /ɛ/	96	0	97	0	98	0	99	0
ㅓ /ʌ/	87	16	80	27	94	5	79	27
ㅗ /o/	98	0	98	0	99	0	98	0
ㅜ /u/	100	0	98	0	97	1	96	0
ㅡ /ɯ/	97	0	95	3	96	1	93	0
/w/	91	5	92	4	95	5	95	3
/j/	88	11	87	11	93	8	85	17

Perception Profiles

Like in the previous section, with the following results I detail how I identified a clustering solution and describe the phoneme perception clusters that emerged.

Determining the Number of Clusters. To start, I ran a hierarchical cluster analysis (Figure 6.6). Based on the dendrogram, which has color coding for four clusters, three to five clusters seemed likely; the fifth cluster would have split the purple cluster (towards the left-hand side of the Figure 6.6) while the three-cluster solution would have merged the red and green clusters on the left.

For the perception parcel scores, the location of an elbow plot of within-cluster variances for $k = 1$ through $k = 10$ cluster solutions (Figure 6.7) was not so clear. The most acute angle appeared to be centered on a 2-cluster solution, but the following 3-, 4-, and 5-cluster solutions also appeared to offer non-trivial reductions in within-cluster variability. More than five clusters seemed unnecessary. Finally, the gap statistic plot (Figure 6.8) suggested that $k = 4$ clusters would be optimal.

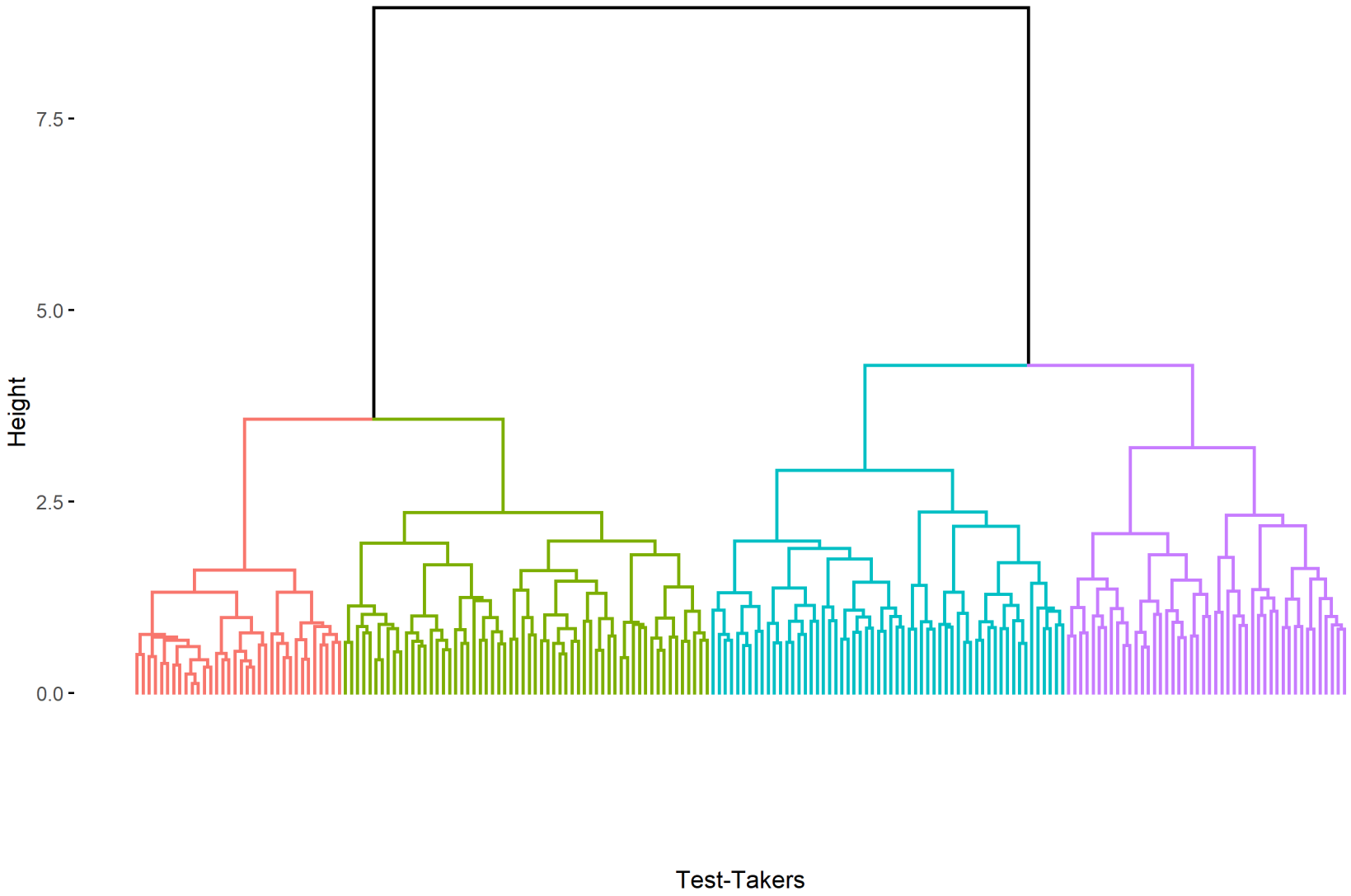


Figure 6.6. HCA dendrogram depicting suggested clustering of test-takers according to production parcel scores.

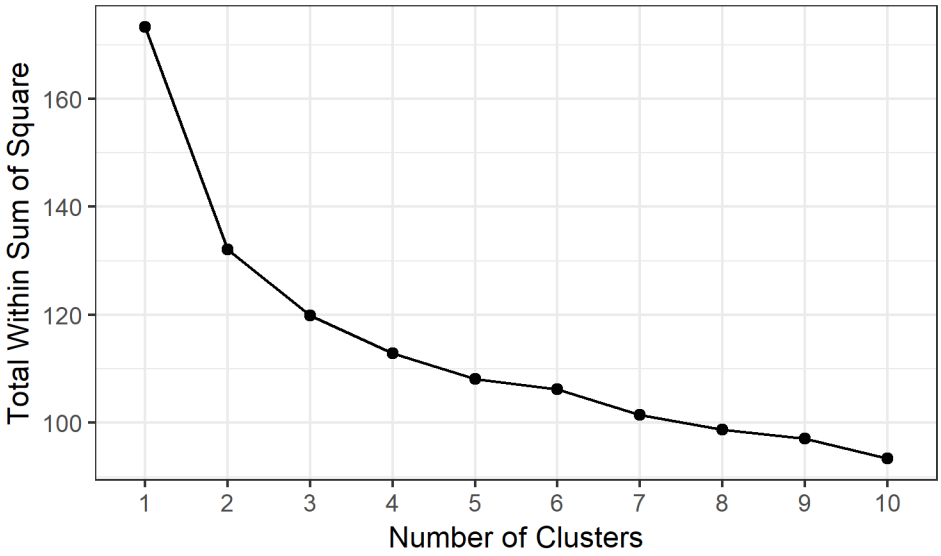


Figure 6.7. Plot of within-cluster sum of squares for k = 1..10 clusters based on perception parcel scores.

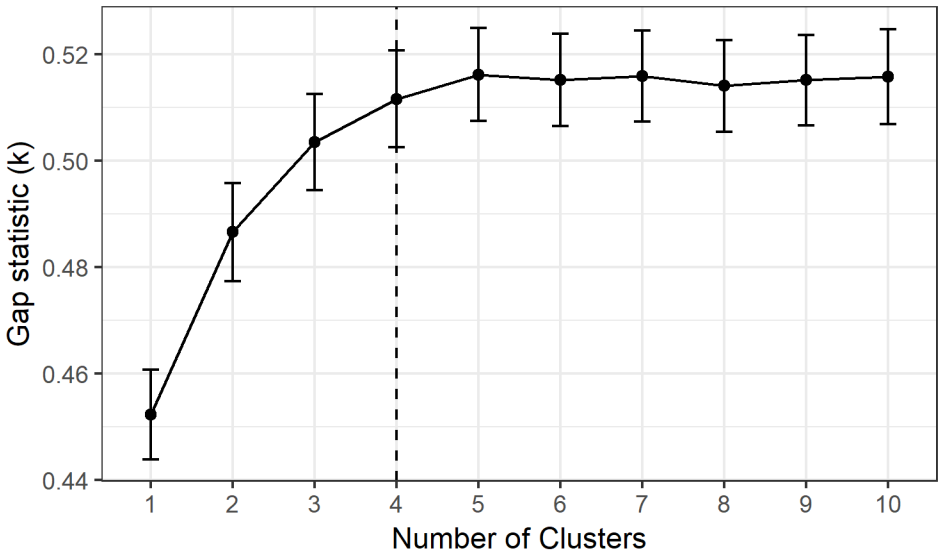


Figure 6.8. Gap statistic plot for k = 1..10 clusters based on perception parcel scores. The vertical dashed line indicates where the number of clusters is optimal.

Ultimately, I settled on a 4-cluster solution, based both on the evidence considered thus far and the descriptive utility of the emergent clusters (see the next subsection for details). All in all, a 4-cluster *k*-means solution appeared to be sufficiently well-supported. The 4-cluster solution resulted in 42 test-takers in Cluster 1, 46 in Cluster 2, 74 in Cluster 3, and 36 in Cluster 4. Figure 6.9 plots the four clusters in two dimensions based on a principal components analysis of the production parcel data; each test-taker's first contrast (Dim1) and second contrast (Dim2) scores are used for Cartesian coordinates. While the four clusters did overlap somewhat in this two-dimensional representation, some clear distinctions between pairs of clusters were quite apparent. For example, there is no overlap between Clusters 3 and 4.

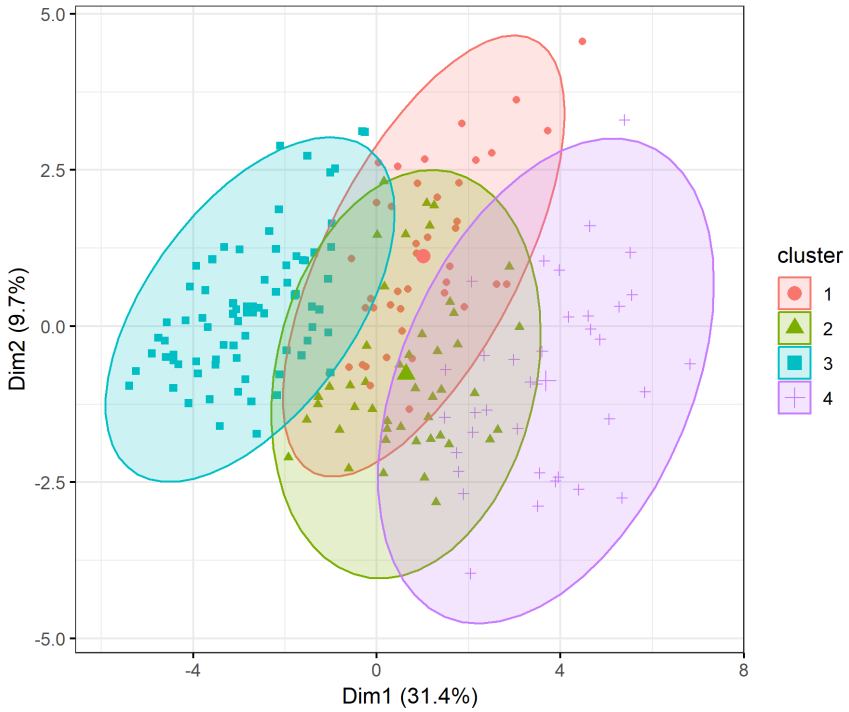


Figure 6.9. Plot of clusters along the first two principle components of the perception parcel data.

Cluster Descriptions. For each cluster I computed the mean accuracy for each perception parcel and estimated the proportion of diagnostic flags (based on a < 75% criterion).

These values are visually represented in Figure 6.10 (numeric values are available in Table 6.2). Based on the mean accuracy rates and the proportion of diagnostic flags, the four clusters can be summarized as follows in terms of phoneme perception difficulties:

- Perception Cluster 1: Difficulties with /p/, /s, s*/, some difficulty with back vowels, especially /u/.
- Perception Cluster 2: Moderate difficulties with many stop consonants and fricatives, difficulty with the /s, s*/ and /o-ʌ/ distinctions.
- Perception Cluster 3: Minimal difficulties outside of /u/ and /s*/.
- Perception Cluster 4: Considerable difficulties with most consonants, back vowels and glide /j/.

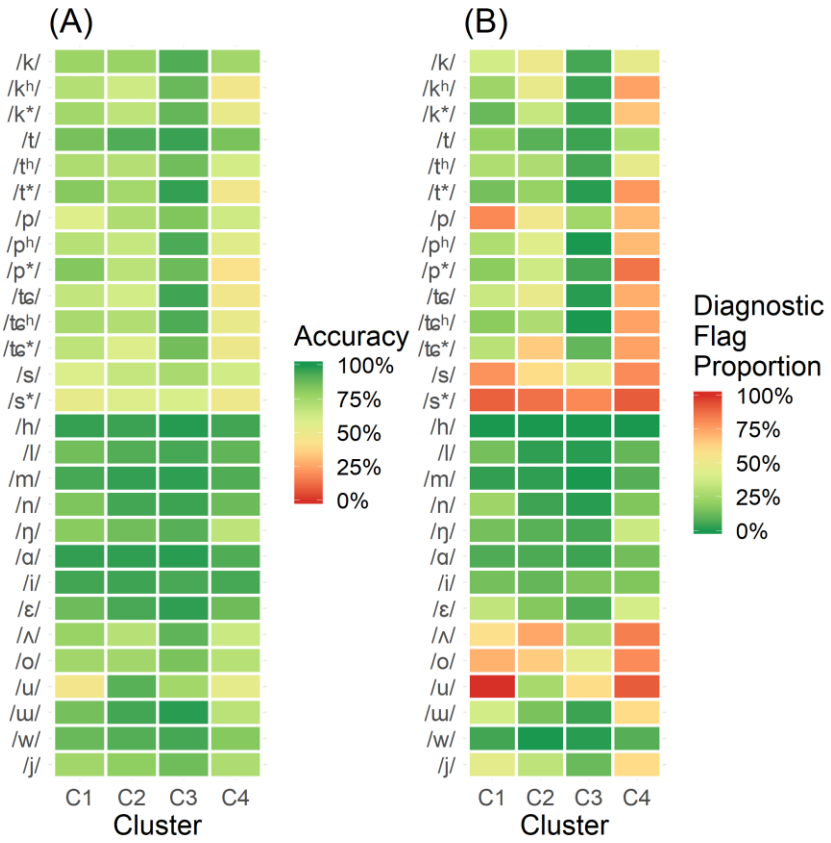


Figure 6.10. Heatmaps of phoneme perception mean accuracy (A) and diagnostic flag proportion (B) by cluster.

Table 6.2

Phoneme Perception Mean Accuracy and Diagnostic Flag Proportion by Cluster

	C1		C2		C3		C4	
	Acc.%	Flag%	Acc.%	Flag%	Acc.%	Flag%	Acc.%	Flag%
ㄱ /k/	77	38	78	50	93	5	75	47
ㅋ /kʰ/	70	24	63	48	88	4	47	75
ㄲ /k*/	75	12	67	35	89	4	52	67
ㄷ /t/	85	21	93	9	97	4	85	28
ㅌ /tʰ/	72	29	70	28	86	5	62	47
ㄸ /t*/	82	14	74	22	97	1	47	78
ㅂ /p/	58	81	72	52	83	24	63	69
ㅍ /pʰ/	70	29	65	43	94	0	56	69
ㅃ /p*/	83	19	68	37	88	5	43	86
ㅈ /ʧ/	66	36	62	48	96	1	47	72
ㅊ /ʧʰ/	73	19	71	28	93	0	52	75
ㅉ /ʧ*/	68	31	58	65	86	11	49	75
ㅅ /s/	59	79	65	61	73	45	62	81
ㅆ /s*/	53	90	57	87	61	81	50	92
ㅎ /h/	97	0	96	0	99	0	95	0
ㄹ /l/	86	14	92	2	94	1	90	11
ㅁ /m/	94	2	97	2	98	0	93	8
ㄴ /n/	84	24	95	4	96	1	88	17
ㅇ /ŋ/	82	14	86	9	92	5	67	36
ㅏ /ɑ/	98	7	98	7	99	4	93	14
ㅣ /i/	95	14	96	11	94	16	94	17
ㅔ /ɛ/	87	33	94	17	98	7	87	39
ㅓ /ʌ/	78	60	70	74	90	28	64	83
ㅗ /o/	75	71	75	65	85	45	69	81
ㅜ /u/	47	100	91	26	75	61	54	92
ㅡ /ɯ/	86	38	95	15	99	4	69	61
/w/	88	5	92	0	95	1	82	8
/j/	75	45	80	33	87	12	72	61

Profiles, L1, and Proficiency

I considered two relevant background variables, L1 (dominant language) and oral proficiency (as measured by the EIT), alongside production and perception cluster membership. While there were two dozen L1s represented in the sample, it was only meaningful to examine

the distribution of L1s across clusters for languages with several speakers; I selected 10 speakers as a cutoff for inclusion in these analyses.

Table 6.3 contains information on the L1 composition of the production clusters. Keeping in mind that Cluster 3 was indicative of few pronunciation difficulties, it is interesting to examine where among the remaining clusters learners from each L1 subgroup were concentrated. L1 Chinese (Mandarin) speakers, the most numerous L1 subgroup, were absent from Cluster 1 and mostly were split across Cluster 2 and 3, with a handful in Cluster 4. English speakers fell primarily into Cluster 3, but several were found in Cluster 2 and 4. This relatively even distribution across clusters was found for the Japanese and Spanish speakers as well. Russian speakers, like the Chinese speakers, were absent from Cluster 1, but nearly half of all Russian speakers fell into Cluster 4.

Table 6.3

L1 Composition of Phoneme Production Clusters

Cluster	Chinese	English	Japanese	Russian	Spanish	Others	Total
C1	0 (0%)	2 (11%)	*4 (31%)*	0 (0%)	*3 (27%)*	10 (21%)	19
C2	**48 (55%)**	4 (21%)	2 (15%)	*6 (32%)*	1 (9%)	*12 (25%)*	73
C3	*35 (40%)*	**8 (42%)**	**5 (38%)**	4 (21%)	**4 (36%)**	**20 (42%)**	76
C4	5 (6%)	*5 (26%)*	2 (15%)	**9 (47%)**	*3 (27%)*	6 (13%)	30
Total	88	19	13	19	11	48	198

Note. Percentages are based on L1 subgroups (columns). **Bold** indicates highest proportion of an L1 subgroup, *italics* indicate second highest proportion.

For oral proficiency (Table 6.4, based on EIT scores), participants in Cluster 3 had the highest mean oral proficiency. Examination of the 95% confidence intervals shows that Cluster 3 had significantly higher oral proficiency than Clusters 1 and 4, but Cluster 2 and 3 overlapped considerably, as did Cluster 1 and 4. However, a one-way analysis of variance (ANOVA) based on production cluster membership did not return a significant result ($F(1, 196) = 2.39$, $p =$

0.124), indicating that on a whole the null hypothesis (that clusters did not vary in proficiency) could not be rejected. Interestingly, oral proficiency standard deviations were similarly large across clusters, and each cluster featured at least one member with rather low or considerably high oral proficiency.

Table 6.4

Oral Proficiency of Phoneme Production Clusters

Cluster	Mean	SD	95% CI	Min	Max
C1	46.68	25.93	[34.18, 59.18]	13	97
C2	68.01	21.36	[63.03, 86.28]	21	106
C3	81.14	22.50	[76.00, 86.23]	20	116
C4	54.26	23.98	[45.31, 63.22]	9	109

Table 6.5 contains L1 composition for each perception cluster. Like the production clusters, Cluster 3 membership was indicative of few difficulties with Korean phonemes. Chinese speakers were primarily concentrated in two of the perception clusters (3 and 1), with a handful being grouped in each of the other two clusters. English speakers were concentrated in Clusters 2 and 3, with a few in Cluster 4, while Japanese speakers mostly fell into Clusters 4 and 1. Aside from one test-taker in Cluster 1, Russian speakers were evenly split across Clusters 2 and 4. Spanish speakers were also concentrated in these two clusters. No Russian or Spanish speakers were found in Cluster 3.

Table 6.5

L1 Composition of Phoneme Perception Clusters

Cluster	Chinese	English	Japanese	Russian	Spanish	Others	Total
C1	*26 (30%)*	1 (5%)	*4 (31%)*	1 (5%)	2 (18%)	8 (17%)	42
C2	7 (8%)	**7 (37%)**	1 (8%)	**9 (47%)**	*4 (36%)*	**18 (36%)**	46
C3	**50 (57%)**	**7 (37%)**	2 (15%)	0 (0%)	0 (0%)	*15 (31%)*	74
C4	5 (6%)	4 (21%)	**6 (46%)**	**9 (47%)**	**5 (46%)**	7 (15%)	36
Total	88	19	13	19	11	48	198

Note. Percentages are based on L1 subgroups (columns). **Bold** indicates highest proportion of an L1 subgroup, *italics* indicate second highest proportion.

For oral proficiency (Table 6.6), Cluster 3 had a visibly higher mean proficiency compared to all other clusters: An examination of the 95% confidence intervals suggests that this difference is statistically significant. However, there did not appear to be any reliable differences among the means of the other clusters, as indicated by their highly overlapping confidence intervals. An ANOVA did not return a statistically significant result for oral proficiency differences based on phoneme perception cluster membership ($F(1, 196) = 1.13$, $p = 0.289$), which did not permit me to reject the null hypothesis that clusters were not different in oral proficiency. Interestingly, all four clusters had at least one member with considerably high oral proficiency, and like the production clusters, standard deviations were rather large.

Table 6.6

Oral Proficiency of Phoneme Perception Clusters

Cluster	Mean	SD	95% CI	Min	Max
C1	62.48	24.52	[54.83, 70.12]	19	115
C2	60.89	21.45	[54.52, 67.26]	20	116
C3	84.66	18.41	[80.40, 88.93]	33	115
C4	54.36	27.30	[45.13, 63.60]	9	109

It was also informative to consider combinations of production and perception cluster membership (Table 6.7). Members of Production Cluster 1, who had notable difficulties producing aspirated stops, most commonly fell into Perception Cluster 4, indicating that they also had difficulty perceiving aspirated stops (alongside difficulties with many other consonants). A smaller number fell into Perception Cluster 2, which was also characterized by some difficulty with aspirated stops, though to a lesser degree. Production Cluster 2, which had generally intelligible pronunciation of Korean sounds but moderate difficulties with /s*, ʧ*/, and some difficulty with /k*, ʌ/, fell primarily into Perception Clusters 3 (minimal difficulties outside of /s*, u/and 1 (considerable difficulties with /s, s*/, difficulties distinguishing among /ʌ, o, u/), which seemed to align well. Interestingly, 11 members of Production Cluster 2 fell into Perception Cluster 4, which was characterized by a wide range of perception difficulties including those that were not salient problems for production. Production Cluster 3, which had good control of nearly all phonemes, mostly fell into Perception Cluster 3, as one might expect, though individuals fell into other perception clusters. Finally, Production Cluster 4, marked by the most severe and broad pronunciation difficulties, had no one fall into Perception Cluster 3; instead Production Cluster 4 members were concentrated in Perception Clusters 2 and 4.

Table 6.7

Cross-Tabs of Production and Perception Cluster Membership

		Perception			
		C1	C2	C3	C4
Production	C1	3	6	1	9
	C2	24	13	25	11
	C3	13	14	48	1
	C4	2	13	0	15

Discussion

In this chapter, I used cluster analysis to identify groups of learners with similar production and perception profiles. Separate analyses on phoneme parcel scores for production and perception each identified four reasonably well-defined clusters. In each set of clusters, the largest cluster consisted of few to no difficulties with Korean sounds. The remaining clusters were characterized by learners with varying weaknesses, both in terms of targets and degree of difficulty. In broad terms, the answer to Research Question 3 appears to be "yes": The KPD was able to detect substantive differences among test-takers' pronunciation profiles, spanning the production and perception of Korean sounds, including the identification of profiles that were common to subgroups of learners.

Of particular interest was Production Cluster 1. The individuals within it had greater difficulty producing aspirated consonants than some tensed consonants, bucking the general trend in phoneme production difficulty found in Chapter 5 (i.e., that tensed consonants were the most difficult phonemes to produce). This cluster interrupts potential interpretations of the clusters representing developmental stages. It is possible to trace a path of development from Cluster 4 to 2 to 3, where less difficult phonemes (Chapter 5) are mastered first followed by more difficult phonemes. However, Cluster 1 does not fit neatly into any sort of similar progression, as they showed better mastery of the generally difficult tense consonants compared to the moderate difficulty aspirated consonants. Interestingly, Cluster 1 and Cluster 4 had highly similar oral proficiency levels.

Production Cluster 4 was notable for its L1 composition. Dominated by Russian speakers (roughly 1/3 of the cluster, and nearly half of all L1 Russian test-takers), it also featured fair proportions of English, Spanish, and other L1 speakers. There were even a small handful of

Chinese L1 speakers in the cluster; in other words, those five L1 Chinese speakers of Korean had more in common with the members of Cluster 4 than they did with the majority of their L1-background peers, who fell into Cluster 2.

Perception Cluster 4 is another interesting cluster. Like Production Cluster 4, it breaks up a clean interpretation of clusters as developmental stages or solely proficiency related. Perception Cluster 4's acute difficulties with /p/ and the location of the most acute back-vowel difficulty (i.e., /u/ instead of /ʌ/) are distinguishing features that prevent interpretation of a neat Perception Cluster 4 → 2 → 1 → 3 progression or continuum. Roughly equal proportions of L1 Chinese and Japanese speakers ended up in Cluster 1, along with smaller proportions of Spanish, Other L1s, English, and Russian speakers. In terms of overall oral proficiency, perception Clusters 1, 2, and 4 were extremely similar, further dampening any clear developmental interpretation of cluster membership, though it can be said that most learners with advanced oral proficiency tend to have few phoneme perception difficulties (Cluster 3).

Cross-referencing production and perception cluster memberships revealed further differences among learners. Members of Production Cluster 2, who struggled with just a few tense consonants and had some difficulty with /ʌ/, were spread rather evenly across the perception clusters. This provides some support for the inclusion and instructional utilization of phoneme perception scores. For instance, those test-takers in Production Cluster 2 who fell into Perception Cluster 3 (i.e., the cluster with little to no phoneme perception difficulty) would be unlikely to benefit much from perception practice of the /ʨ*, k*/ (and to a lesser extent, /ʌ/) targets. However, those who fell in to Perception Cluster 4 would almost certainly benefit from such perception practice.

The L1 composition of the various clusters highlights a key point: Clustering was not L1-deterministic. Although there were some clear trends among L1 and cluster membership, which was to be expected to at least some degree, portions of test-takers from the every major L1 background were located in more than one cluster. Some of this L1 dispersal is due to the achievement of highly intelligible Korean pronunciation and the strong speech sound perception ability by learners from a wide range of backgrounds. However, even among learners who had not yet achieved those high levels of L2 phonological competence, dispersion across diagnostic profiles was evident. As mentioned, learners from several L1 backgrounds could be found in the peculiar Production Cluster 4 and Perception Cluster 1. L1 English and Japanese speakers were broadly dispersed across the four Production and Perception Clusters. Thus, the KPD appears to have some utility in identifying learner profiles beyond what could be guessed at by any L1-based generalizations or a contrastive analysis.

Similarly, although there was a clear trend for learners with advanced oral proficiency to fall into the production and perception clusters with few difficulties, there were not always clear differences among clusters in terms of their overall proficiency. For production clusters, it is notable that Clusters 2 and 3 did not appear to be reliably different in oral proficiency, as their confidence intervals overlapped to a considerable degree. Furthermore, in both production and perception, very high and very low oral proficiency members could be found in all clusters. This too suggests that the KPD is potentially useful for addressing a wide array of learners, including beginners who are having difficulty tuning in to Korean phonology and advanced learners who perhaps have a fossilized interlanguage phonology that persists to cause them difficulties in communication.

This utility potentially holds considerable value for pedagogy. Teacher training manuals and textbooks for pronunciation often rely on L1-based recommendations for addressing specific learner difficulties (e.g., Kwon, 2017, for an example of a L2 Korean teacher training text and Choi, Kim, Park, Jin, & Park, 2009a, 2009b for Korean pronunciation textbooks). While such recommendations may serve as a broadly useful starting point, the present findings suggest that they will fall short for addressing the needs of some learners, not to mention wasting the time of some others. For example, the Choi et al. (2009a, 2009b) textbooks provide recommendations of which sounds (and corresponding textbook units) to focus on for twelve different L1 background (e.g., English, Japanese, Chinese, Arabic). L1 Chinese speakers are advised to focus on the /s, s*/ and /tɕ, tɕ*, tɕ^{h}/ distinctions, which indeed proved challenging for many L1 Chinese speakers in this study, but the book also recommends focus on /p, p*, p^{h}/ and /t, t*, t^{h}/, which were not a common problem for the numerous Chinese speakers in Production Cluster 2. That latter advice would be better suited to the members of Production Cluster 1. Some of the L1-based advice for learning and teaching may also be at odds with the Intelligibility Principle and less relevant for instruction. Kwon (2017, p. 124) noted that English speakers often substitute the Korean mid vowels /ɛ/ and /o/ with English diphthongs [eɪ] and [oʊ], respectively, and suggested that teachers make L1 English learners aware of these mistakes. However, neither of these sounds were difficult to produce intelligibly for any cluster, much less clusters with larger concentrations of English speakers. The intelligibility-focused pronunciation diagnosis of the KPD may be able to point teachers and learners to more productive uses of limited time and energy.

In the specific context of this study, it is worth pointing out that many of the participants attended the same intensive Korean program or the same graduate program in Korean as a

foreign language (focusing on either education or translation/interpretation), and some of them attended specific classes together. Having learners from different backgrounds and with different pronunciation needs is more than a hypothetical.

Some important qualifications need to be made to the discussion of results thus far. I have so far pointed out where L1-based predictions of pronunciation difficulties fall short and where production and perception profiles have not lined up, each of which have potentially valuable instructional implications. However, I must clarify that the present results do not contradict or meaningfully call into question prevailing theory and findings in the fields of L2 pronunciation and speech learning. Indeed, there was considerable L1 patterning observable in the clusters: Of L1 Chinese speakers with notable pronunciation difficulties, virtually all of them fell into one production cluster (Production Cluster 2). Similarly, most of the adept articulators from Production Cluster 3 were also members of the highly-skilled Perception Cluster 3, in line with theoretical expectations (Best & Tyler, 2007; Flege, 1995), with relatively few falling into clusters characterized by substantial perception difficulties. But with almost any theory, especially complicated ones with many moving parts and difficult-to-observe processes, there are people who will fall somewhat to the wayside of group-level predictions. DLA can map out ability profiles for those individuals in ways that basic theory-driven expectations might not be able to, and in turn provide relevant support that might otherwise be missing from standard instructional materials, approaches, or curricula.

CHAPTER 7: EXTERNAL RELATIONSHIPS

The relationships between KPD scores and external variables, which are relevant to the explanation and extrapolation inferences in the KPD's proposed validity argument, are the focus of this chapter. Relevant to the explanation inference, I examined the relationship between KPD results and general oral proficiency. Relevant to the extrapolation inference, I examined the relationships between KPD results, pronunciation performance in spontaneous speech, and learner self-assessments. I utilized correlations and descriptive statistics to examine these relationships. In the discussion which follows the presentation of the results, I consider findings in relation to the primary research questions listed below.

Research Questions

The primary research questions addressed by the results in this chapter are:

- RQ4: To what extent do KPD results show an expected relationship with Korean oral proficiency?
- RQ5: To what extent do results reflect difficulties test-takers show in spontaneous, meaning-focused speech?
- RQ6: To what extent do results reflect self-assessments of pronunciation ability and difficulties?

For RQ4, Korean oral proficiency is a product of language experience and instruction, two factors known to influence L2 phonological development (Piske et al., 2001). This premise is well-grounded in SLA theory and empirical research, which has shown generally positive associations between both the amount of instruction and language experience (e.g., length of residence in an L2 environment) on proficiency outcomes (Isbell, Winke, & Gass, 2018). Compared to self-reports of language experience or the amount of instruction, oral proficiency as

measured by EIT scores is more directly comparable among subjects. In general, more proficient L2 speakers tend to have more intelligible and comprehensible L2 speech (Kang & Moran, 2014), and thus there is expected to be a small-to-moderate relationship between KPD results and overall oral proficiency. Similarly, higher proficiency learners are expected to have higher accuracy in the production and perception of individual phonemes, and vice-versa. As segmental production and phoneme identification are but pieces of speaking and listening processes and proficiency, respectively, an exceedingly strong relationship cannot be reasonably expected. Furthermore, local fossilization (i.e., the cessation of development over a considerable period of TL exposure and use, Han, 2004) and plateaus in interlanguage phonology are well-attested phenomena in the L2 pronunciation literature, whereby generally high-proficiency speakers' productions are characterized by non-target like and sometimes unintelligible articulation of L2 speech sounds (Derwing & Munro, 2007; Derwing et al., 2014). The presence of such individuals in the present analyses, such as one participant with over 10 years of residence in Korea currently pursuing a doctoral degree (see Chapter 8), would limit the strength of any quantitative relationship between pronunciation ability and overall oral proficiency.

For Research Question 5, KPD results, in terms of difficult to produce/inaccurate phoneme scores, ought to be reflected in spontaneous, meaning-oriented oral production. If the KPD results *do not* reflect difficulties in meaning-oriented oral communication, it would be hard to argue that the test reflects learners' actual pronunciation weaknesses, and in turn that the test has any utility at all. Arguably, this is the most important piece of evidence in support of the extrapolation inference. However, it may also be the most difficult evidence to adequately capture, as collecting and analyzing spontaneous speech is subject to a host of challenges, such as collecting a long and representative enough speech sample(s) that would facilitate rigorous

and generalizable analyses of a learner's complete segmental inventory in production. As such, the analysis and results presented here must be treated as preliminary.

Research Question 6 provides an additional perspective on the extrapolation of KPD scores to pronunciation and perception in meaning-oriented general oral language use via comparison with learner self-assessments. To at least some extent, learner's KPD scores should reflect their own observations of production and perception difficulties in their daily use of the language in Korea. However, the degree to which participants are (un)aware of their own specific phoneme-level difficulties is likely to limit the strength of the relationship between KPD scores and self-assessments. The findings related to this research question also bear on the utilization inference: Learners who may have poorer self-assessments (awareness) of their pronunciation and perception abilities stand to benefit the most from receiving KPD results. Similarly, if learners' self-assessments were in perfect alignment with KPD results, there would be little reason to use the KPD at all, as self-assessments require fewer resources.

Analysis Details

The analyses related to each external measure are detailed in the following subsections.

Oral Proficiency

I examined the EIT and KPD parcel scores for all 198 learners in the study. I correlated EIT scores with KPD scores (averages and individual phonemes) in perception and production. I also examined average KPD parcel accuracy scores and individual phoneme scores for learners in different quantiles of EIT scores.

Pronunciation in Spontaneous Speech

To explore the relationship between KPD results and pronunciation difficulties evident in spontaneous speech samples, I selected a subset of 21 learners' independent speaking (IS)

productions (with accompanying phonemic transcriptions) and conducted an error analysis. These 21 learners are the same learners who participated in follow-up interviews (see Chapter 8) who were originally selected for their diversity of background (proficiency level, academic status, L1) and pronunciation difficulties (numerous to minimal, with different strengths and weaknesses). By analyzing the IS responses of these learners, I ensured broad representation in a limited sample size and at the same time enriched the findings related to the interview study.

For the 21 transcriptions (originally transcribed by linguistically-trained NSs, and edited by me for consistency of conventions, see Chapter 4), I coded all deletion (removal of a phoneme) and substitution (replacement of a phoneme with a non-target sound) errors. Knowledge of the prompt and repeated careful listens of speech files helped me judge what a speaker intended to say, and in instances where I had difficulty or uncertainty in determining intended words I consulted a NS highly familiar with L2 Korean speech (the same NS Korean instructor who scored the KPD; see Chapter 4). I counted the total number of phonemes, excluding nonverbal sounds (e.g., 엄…, *umm…*) but including lexical word fillers (e.g., 뭐, *what*), repetitions (e.g., 그 그, *that that*), and false-starts/interrupted words (e.g., 살- 살아요, *li-live*). I also counted the total number of erroneous phonemes. For each error, I tallied which target phoneme was mispronounced.

Self-Assessment

I examined the relationship between KPD scores and self-assessments in three ways: computation of difference scores, correlations, and alignment with diagnostic flags. KPD parcel scores for each phoneme in production and perception were in the form of percentages (see Chapter 4 for details), and to facilitate the computation of difference scores, I converted the phoneme-level self-assessments to percentages. For each learner, I subtracted their self-

assessment (as a percentage) from their KPD parcel score to arrive at a difference score. Each learner's mean difference score in production and perception was computed.

For correlational analyses, I used Pearson correlations (no substantial differences were found when using Spearman correlations). I started with correlations among global measures of pronunciation ability: I computed correlations between global self-assessments (accentedness and comprehensibility), average phoneme-level self-assessments in each modality, and average KPD scores (i.e., average parcel scores across phonemes in each modality).

To investigate learner agreement with Korean phonemes most in need of remediation, I examined the alignment of KPD diagnostic flags and phoneme-level self-assessments. As the reader might recall, diagnostic flags were assigned to phoneme parcels with < 75% accuracy, and I dichotomized the self-assessments using the same criterion (< 75%, i.e., a rating of 5 or less out of 7).

Results

The following subsections feature results of analyses on the relationships between KPD results and oral proficiency, pronunciation in spontaneous speech, and self-assessment.

Relationship between KPD Results and Oral Proficiency

Among the 198 field testing participants, the average EIT score was 68.92 out of 120 (SD = 25.38, median = 70, min = 9, max = 116). Figure 7.1 shows the distribution of total EIT scores.

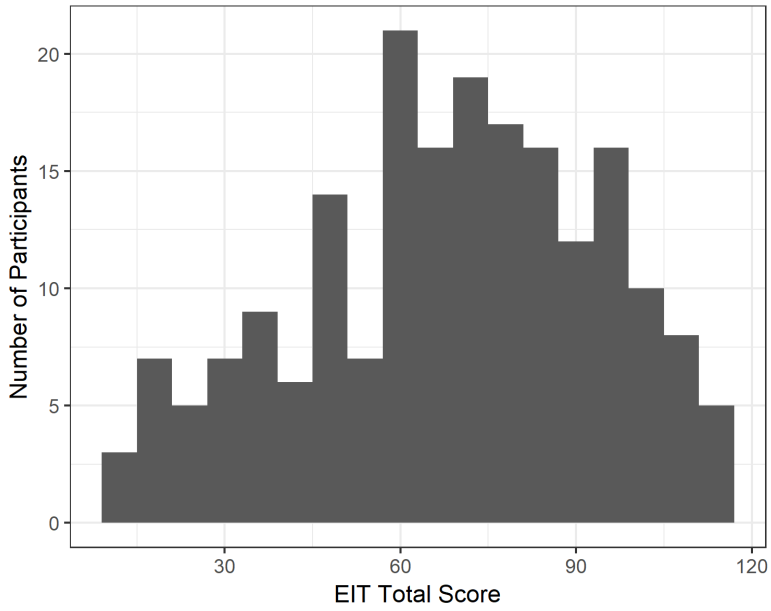


Figure 7.1. Distribution of EIT scores.

The correlation between EIT scores and average production phoneme parcel accuracy was $r = .51$ (95% CI [0.39, 0.60], $p < .001$), and the correlation between EIT scores and average receptive phoneme parcel accuracy was $r = .56$ (95% CI [0.45, 0.65], $p < .001$). The scatterplots in Figure 7.2 visually represent these relationships. As a convenience and reminder for readers, the correlation between average production parcel accuracy and average perception phoneme parcel accuracy originally reported in Chapter 5 was $r = .74$. To further illustrate the relationship between KPD scores and oral proficiency, I divided the 198 learners into quantiles based on their EIT scores and then computed summary statistics for average production and perception phoneme parcel accuracy (Table 7.1). While the mean production and perception phoneme accuracy increases across oral proficiency quantiles as expected, the differences among quantiles are not extremely large. The third and fourth oral proficiency quantiles differ very little in average production phoneme accuracy, with nearly identical means and standard deviations. Further, these two quantiles are not so different from the second quantile in terms of phoneme

production. The progression of perception phoneme accuracy across quantiles is more clear-cut when examining means, yet at the same time there is greater intra-quantile variation in average perception accuracy.

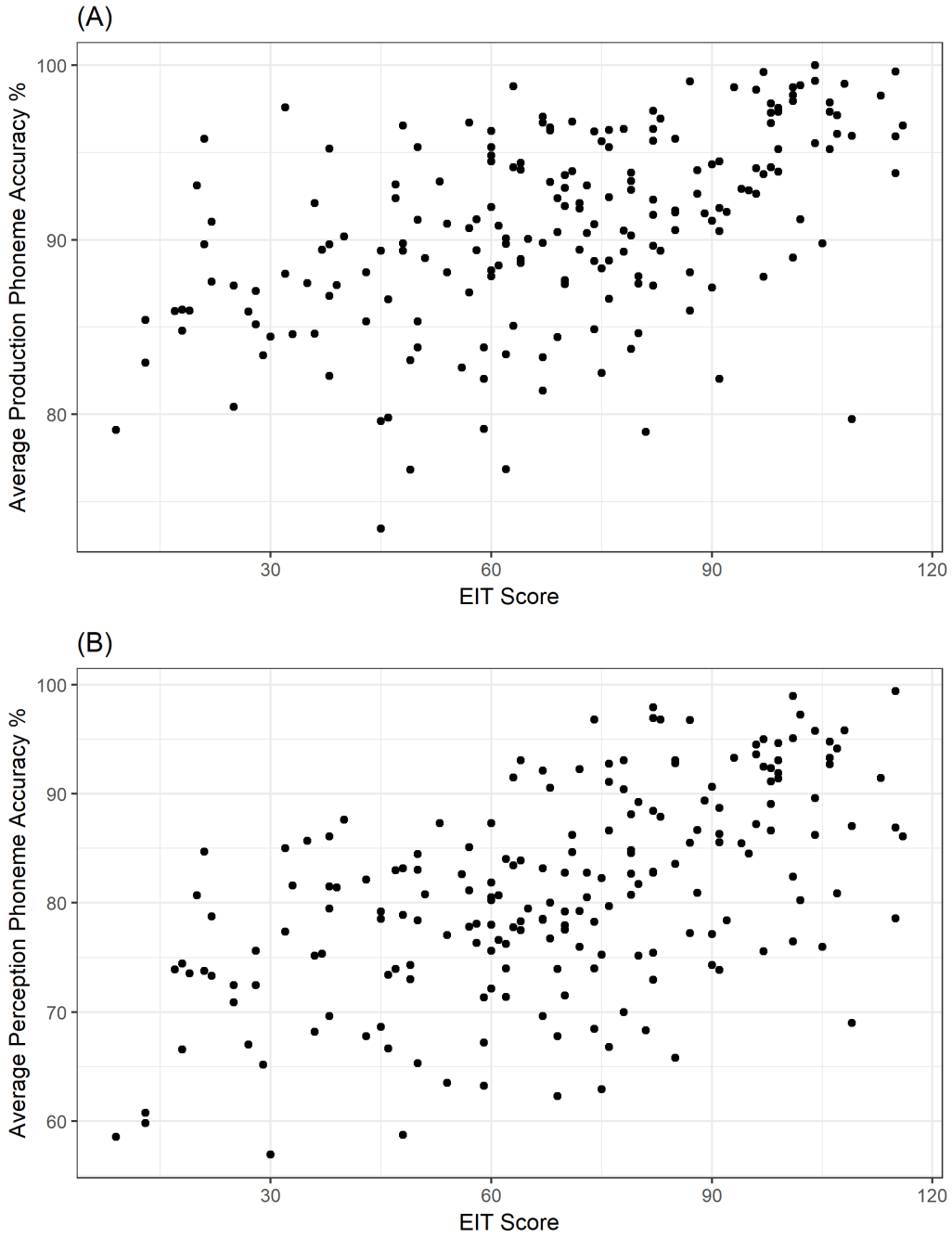


Figure 7.2. Scatterplots of the relationship between EIT scores and (A) average production phoneme accuracy and (B) average perception phoneme accuracy.

Table 7.1

Average Production and Perception Phoneme Parcel Accuracy by Oral Proficiency Quantiles

		Production*				Perception*			
Quantile	EIT Mean (SD)	Mean	SD	Min	Max	Mean	SD	Min	Max
1	30.93 (10.66)	87	5	73	98	74	8	57	88
2	56.60 (5.31)	89	5	77	99	78	7	59	91
3	70.71 (3.71)	91	4	81	97	80	8	62	97
4	84.38 (4.67)	91	4	79	99	84	8	66	98
5	102.90 (6.17)	95	4	80	100	89	7	69	99

Note. *All values are percentages.

In addition to the relationships between oral proficiency and overall phoneme production and perception accuracy, I considered the relationship between oral proficiency and individual phonemes. To this end, I computed correlations for production and perception accuracy and oral proficiency for each phoneme (Table 7.2). The average correlation between EIT scores and production phonemes was .18, with a minimum of -.05 and maximum of .33. For phonemes that were generally easy to produce such as /ɛ, u/ (see Chapter 4), correlations between production accuracy and oral proficiency were small, likely due to attenuation. On the other hand, tense and aspirated consonants had moderate correlations with oral proficiency. For perception phonemes, the average correlation was .28 with a minimum of .08 and maximum of .45. While some of the easier to perceive vowels had smaller correlations, on a whole the correlations between each perception phoneme's accuracy scores and oral proficiency were moderate. In sum, the relationship between phoneme perception and oral proficiency was relatively stronger than the relationship between phoneme production and oral proficiency.

Table 7.2

Correlations between Phoneme Production, Perception, and Oral Proficiency

Phoneme	Production-EIT (*r*)	Perception-EIT (*r*)
ㄱ /k/	0.19	0.27
ㅋ /kʰ/	0.22	0.41
ㄲ /k*/	0.27	0.38
ㄷ /t/	0.17	0.41
ㅌ /tʰ/	0.31	0.28
ㄸ /t*/	0.33	0.44
ㅂ /p/	0.17	0.33
ㅍ /pʰ/	0.29	0.40
ㅃ /p*/	0.31	0.34
ㅈ /ʧ/	0.22	0.44
ㅊ /ʧʰ/	0.29	0.45
ㅉ /ʧ*/	0.26	0.34
ㅅ /s/	0.14	0.18
ㅆ /s*/	0.28	0.09
ㅎ /h/	0.04	0.38
ㄹ /l/	0.15	0.30
ㅁ /m/	0.14	0.16
ㄴ /n/	0.17	0.15
ㅇ /ŋ/	0.22	0.19
ㅏ /ɑ/	0.11	0.08
ㅣ /i/	0.16	0.18
ㅔ /ɛ/	-0.05	0.17
ㅓ /ʌ/	0.26	0.27
ㅗ /o/	0.11	0.21
ㅜ /u/	-0.01	0.12
ㅡ /ɯ/	0.07	0.30
/w/	-0.02	0.30
/j/	0.15	0.35

As a means of visually exploring the relationship between oral proficiency and phoneme accuracy, I plotted phoneme accuracy means in production and perception for each quantile (Figure 7.3). This visualization illustrates how some of the previously discussed correlations vary. The accuracy of phoneme /ɑ/, which had very small correlations with oral proficiency, had

nearly uniform accuracy in perception and production across oral proficiency quantiles. In contrast, the progression in accuracy for /t*/ in both production and perception is visually distinctive, as one might expect given the stronger correlations between accuracy and oral proficiency. For the most part, phonemes which were found to be more difficult according to the measurement analyses in Chapter 4 tended to demonstrate larger correlations between accuracy and oral proficiency and more visible upward progressions across oral proficiency quantiles; low-difficulty phonemes lacked such relationships with oral proficiency. Notable exceptions include /u/ and /s*/ in perception: These phonemes showed little relationship with oral proficiency yet were among the most difficult for learners to perceive accurately, implying that distinct perceptive category formation and/or the ability to discriminate these sounds from similar phonemes eludes even many highly-proficient L2 speakers of Korean.

Relationship between KPD Results and Pronunciation in Spontaneous Speech

Examining the relationship between the KPD results and Independent Speaking production was challenging because roughly 1 minute (or less) of spontaneous speech is not guaranteed to elicit all 28 Korean phonemes, much less multiple instances of each phoneme. The most common phonemes of Korean (e.g., /ɑ, n, k/, Shin et al., 2013) were plentiful in speech samples, but less-common phonemes were minimally present or not present at all. For instance, it is entirely possible to respond to the Independent Speaking prompt without using the phoneme /tɕ*/, which according to Shin et al. (2013) makes up less than 1% of phonemes produced in typical Korean speech (avoiding other phonemes unintentionally is also distinctly possible). S113 did exactly this, which was unfortunate because she had a 0% production accuracy score on the KPD for that phoneme.

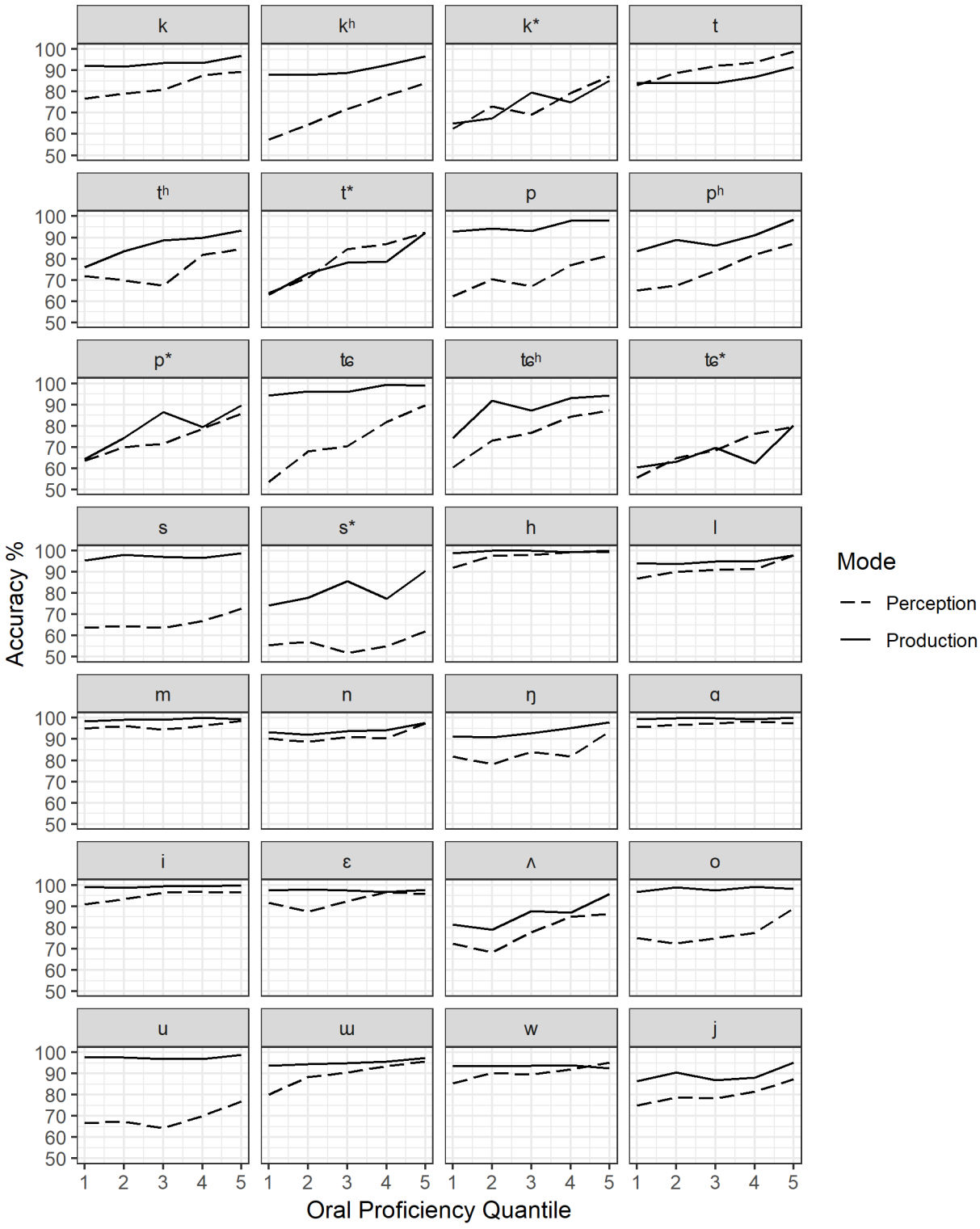


Figure 7.3. Mean production and perception phoneme accuracy across oral proficiency quantiles.

Thus, I focused this analysis on phonemes mispronounced during IS, and on the KPD production scores for those phonemes. For the subset of 21 analyzed IS samples, Table 7.3 summarizes the overlap between errors in learners' spontaneous speech samples and KPD scores.

One observable trend is that learners with lower error rates in spontaneous speech tended to have higher KPD production parcel averages. Learners S005, S016, S105, S111, and S133 all had high average KPD production accuracy (> 90%) and low phonological error rates in speaking (≤ 3%). Meanwhile, learners with < 90% KPD accuracy tended to have higher error rates in speaking (> 5%, e.g., S001, S040, S054).

Table 7.3

Comparison of KPD Results and Independent Speaking Productions

Test Taker	KPD Avg.	IS Tot. Phon.	IS Err. Rate	Phonological Errors & KPD Score
S001	80%	417	6%	/k/: 5 errors, 86% KPD acc., **/tɕ/: 1 errors, 50% KPD acc.** **/tɕʰ/: 1 errors, 50% KPD acc.**, **/tɕ*/: 1 errors, 75% KPD acc.** **/s/: 4 errors, 70% KPD acc.**, /h/: 1 error, 100% KPD acc. **/l/: 2 errors, 75% KPD acc.**, /ɛ/: 2 errors, 89% KPD acc. /ʌ/: 1 error, 80% KPD acc., /o/: 1 error, 100% KPD acc. **/ɯ/: 1 error, 50% KPD acc.**, /j/: 3 errors, 78% KPD acc.
S004	85%	434	3%	/k/: 5 errors, 79% KPD acc., /tʰ/: 1 errors, 100% KPD acc. **/t*/: 1 errors, 50% KPD acc.**, /l/: 3 errors, 100% KPD acc. /n/: 2 errors, 80% KPD acc., /ŋ/: 1 error, 80% KPD acc. /i/: 1 error, 100% KPD acc., **/o/: 1 error, 75% KPD acc.**
S005	94%	149	2%	/n/: 1 error, 100% KPD acc., /ŋ/: 2 errors, 90% KPD acc.
S013	86%	197	7%	/k/: 2 errors, 86% KPD acc., **/t*/: 1 error, 50% KPD acc.** /s/: 3 errors, 90% KPD acc., /l/: 1 error, 92% KPD acc. /ŋ/: 1 error, 100% KPD acc., /o/: 5 errors, 83% KPD acc.
S014	90%	128	5%	**/k*/: 1 error, 50% KPD acc.,** /t/: 1 error, 100% KPD acc. /s/: 3 errors, 100% KPD acc., /o/: 1 error, 92% KPD acc. /j/: 1 error, 89% KPD acc.
S016	98%	337	3%	/k/: 1 errors, 100% KPD acc., /k*/: 1 errors, 100% KPD acc. /p/: 1 errors, 100% KPD acc., /s/: 2 errors, 100% KPD acc. /s*/: 1 errors, 100% KPD acc., /n/: 2 errors, 90% KPD acc. /ɯ/: 1 errors, 100% KPD acc., /j/: 1 errors, 89% KPD acc.

Table 7.3 (cont'd)

Test Taker	KPD Avg.	IS Tot. Phon.	IS Err. Rate	Phonological Errors & KPD Score
S018	93%	304	6%	/k/: 5 errors, 100% KPD acc., /t/: 2 errors, 89% KPD acc. **/t*/: 1 errors, 75% KPD acc., /p/: 1 errors, 71% KPD acc.** /ʧ/: 2 errors, 100% KPD acc., /s/: 4 errors, 100% KPD acc. /l/: 1 errors, 92% KPD acc.
S035	79%	300	5%	**/kʰ/: 1 errors, 67% KPD acc., /tʰ/: 1 errors, 60% KPD acc.** **/pʰ/: 1 errors, 40% KPD acc.,** /s/: 5 errors, 100% KPD acc. /l/: 3 errors, 83% KPD acc., /n/: 1 errors, 60% KPD acc. **/ŋ/: 1 errors, 50% KPD acc.,** /ʌ/: 2 errors, 80% KPD acc.
S040	84%	131	7%	/s/: 6 errors, 100% KPD acc., /n/: 1 error, 80% KPD acc. /ɛ/: 1 error, 100% KPD acc., **/ʌ/: 1 error, 60% KPD acc.**
S048	88%	488	6%	/k/: 1 errors, 86% KPD acc., /kʰ/: 2 errors, 83% KPD acc. **/k*/: 2 errors, 50% KPD acc.,** /t/: 2 errors, 89% KPD acc. **/t*/: 1 errors, 75% KPD acc.,** /p/: 1 errors, 100% KPD acc. /s/: 14 errors, 100% KPD acc., **/l/: 4 errors, 67% KPD acc.** /n/: 1 errors, 90% KPD acc., /j/: 2 errors, 89% KPD acc.
S054	88%	481	8%	/k/: 1 errors, 86% KPD acc., **/k*/: 2 errors, 75% KPD acc.** **/t/: 1 errors, 67% KPD acc., /tʰ/: 1 errors, 60% KPD acc.** **/t*/: 6 errors, 50% KPD acc.,** /ʧ/: 2 errors, 100% KPD acc. /ʧ*/: 1 errors, 100% KPD acc., /s/: 1 errors, 100% KPD acc. /s*/: 1 errors, 100% KPD acc., /h/: 1 errors, 100% KPD acc. /n/: 1 errors, 100% KPD acc., /ŋ/: 11 errors, 90% KPD acc. /ɛ/: 4 errors, 100% KPD acc., /ʌ/: 1 errors, 80% KPD acc. /o/: 1 errors, 100% KPD acc., /u/: 3 errors, 100% KPD acc.
S074	84%	362	5%	**/kʰ/: 3 errors, 50% KPD acc.,** /t/: 4 errors, 78% KPD acc. **/t*/: 3 errors, 75% KPD acc.,** /p/: 1 errors, 100% KPD acc. **/pʰ/: 1 errors, 40% KPD acc.,** /ʧ/: 1 errors, 100% KPD acc. /l/: 1 errors, 100% KPD acc., /m/: 1 errors, 100% KPD acc. /n/: 3 errors, 90% KPD acc.
S088	83%	349	3%	**/kʰ/: 1 error, 33% KPD acc.,** /t/: 1 error, 89% KPD acc. /ʧ/: 1 error, 100% KPD acc., /m/: 1 error, 100% KPD acc. /n/: 5 errors, 80% KPD acc., /o/: 1 error, 100% KPD acc.
S104	85%	346	5%	/s/: 7 errors, 100% KPD acc., **/l/: 7 errors, 67% KPD acc.** /n/: 2 errors, 90% KPD acc., /i/: 1 error, 100% KPD acc.
S105	92%	276	2%	**/t*/: 2 errors, 75% KPD acc.,** /s/: 4 errors, 80% KPD acc.
S111	92%	372	2%	/t/: 2 errors, 78% KPD acc., /t*/: 1 error, 100% KPD acc. /s/: 2 errors, 90% KPD acc., /j/: 1 error, 89% KPD acc.
S113	77%	181	4%	/s/: 5 errors, 100% KPD acc., /n/: 1 error, 100% KPD acc. /ŋ/: 1 error, 100% KPD acc.

Table 7.3 (cont'd)

Test Taker	KPD Avg.	IS Tot. Phon.	IS Err. Rate	Phonological Errors & KPD Score
S121	86%	524	2%	/k/: 1 errors, 79% KPD acc., **/t*/: 1 errors, 50% KPD acc.** /s/: 2 errors, 90% KPD acc., /n/: 2 errors, 100% KPD acc. /ɑ/: 1 errors, 100% KPD acc., /ʌ/: 2 errors, 100% KPD acc. /j/: 2 errors, 89% KPD acc.
S133	94%	345	1%	/k/: 2 errors, 100% KPD acc., **/k*/: 1 error, 75% KPD acc.** /s/: 1 error, 100% KPD acc., **/s*/: 1 error, 71% KPD acc.**
S139	91%	507	5%	/k/: 1 errors, 93% KPD acc., **/k*/: 2 errors, 50% KPD acc. /t/: 2 errors, 67% KPD acc., /t*/: 1 errors, 75% KPD acc.** /ʧ/: 1 errors, 88% KPD acc., /l/: 4 errors, 92% KPD acc. /n/: 3 errors, 100% KPD acc., /ŋ/: 2 errors, 70% KPD acc. /ɑ/: 3 errors, 100% KPD acc., /ɛ/: 1 errors, 100% KPD acc. /ʌ/: 1 errors, 80% KPD acc., /o/: 2 errors, 100% KPD acc. /u/: 1 errors, 100% KPD acc., /w/: 1 errors, 90% KPD acc. /j/: 1 errors, 89% KPD acc.
S156	93%	139	5%	/k/: 2 errors, 93% KPD acc., /ʧ/: 1 error, 100% KPD acc. /s*/: 2 errors, 86% KPD acc., /j/: 2 errors, 100% KPD acc.

Note. KPD Avg. = Average production phoneme accuracy. IS Tot. Phon. = Total phonemes uttered in independent speaking task. IS Err. Rate = Phonological error rate.

Looking at errors in spontaneous speech in greater detail, the rightmost column of Table 7.3 lists all phonemes which were erroneously produced, the number of times they were produced erroneously, and the KPD production parcel accuracy score for that phoneme. I have bolded phonemes where the KPD accuracy score was at or below the diagnostic flag criterion (75%). In many cases, phonemes which would be interpreted as a substantial difficulty according to the KPD diagnostic flag did show up as problematic in spontaneous speech. Consider learner S133. This learner had relatively high production accuracy overall, according to both the KPD and phonological error rate. The errors this learner did produce in speech, though, aligned in part with phonemes identified as difficult on the KPD: /k*, s*/. Similarly, learner S074 had several points of close alignment with KPD results (/k^h, t*, p^h/). Of course, not every phoneme erroneously produced in the speech samples aligned with KPD results. This could be due to the

complex phonological adjustments in connected speech that were not captured by the KPD, poorly-formed phonological representations of words used in the response, different criteria of KPD scoring and phonemic transcription, or any number of other potential factors.

One phoneme deserves some additional attention and explanation: /s/. This phoneme occurred in the speech sample extremely frequently, in large part due to its connection with the prompt: The word for city, 도시 (/tosi/, [t̪oɕi]), which was central to the topic, contains an /s/. With so many productions of /s/, it is to be expected that more errors could occur. Moreover, this /s/ is also realized as a marked allophone in this word context, [ɕ], which only occurs when followed by /i, j/ and requires a substantial change to articulation. This allophonic variant, in addition to being perhaps more challenging for learners than [s], may also have triggered greater sensitivity on the part of the transcribing team. As a result, for many of the 21 learners, /s/ errors in spontaneous speech were not well-reflected by their KPD scores.

Relationship between KPD Results and Self-Assessments

In the following subsections, I first present a summary results of learner self-assessments. Then, I present the results of analyses of absolute differences between KPD scores and SA responses, correlations between KPD and learner self-assessments, and agreement between KPD diagnostic flags and learner SA responses.

Summary of Learner Self-Assessment. On a scale of 1 (Always Difficult) to 7 (Almost Never Difficult), the average self-assessment of Korean phonemes was 5.47 (SD = 1.64, min = 1, max = 7) in production and 5.33 (SD = 1.71, min = 1, max = 7) in perception. Table 7.4 provides descriptive statistics for SA responses at the phoneme level. For each item, learners in the sample used the full range of the SA scale, and with a few exceptions (a rating of 2 for /ɑ/ in production and perception, a rating of 2 for /i/ in perception) there were observations for every scale

category for every phoneme. Learners rated several tense consonants (/k*, t*, tɕ*, s*/) and, surprisingly, the mid-front vowel /ɛ/ as most difficult to produce, whereas the vowels (/a, i/) and consonants /m, h/ were assessed as the easiest sounds to articulate. In perception, the trio of /tɕ, tɕʰ, tɕ*/ were rated as rather difficult, along with several tense consonants (/k*, t*, s*/), both glides (/j, w/) and several vowels (/ɛ, ʌ, o/).

Table 7.4

Learner Self-Assessment Results: Phoneme/Item-Level Descriptive Statistics

	Production				Perception			
Phoneme	Mean	SD	Min	Max	Mean	SD	Min	Max
ㄱ /k/	5.79	1.33	1	7	5.60	1.42	1	7
ㅋ /kʰ/	5.54	1.64	1	7	5.32	1.69	1	7
ㄲ /k*/	4.95	1.74	1	7	4.98	1.76	1	7
ㄷ /t/	5.64	1.42	1	7	5.51	1.51	1	7
ㅌ /tʰ/	5.60	1.50	1	7	5.40	1.62	1	7
ㄸ /t*/	4.99	1.76	1	7	4.93	1.75	1	7
ㅂ /p/	5.66	1.37	1	7	5.51	1.44	1	7
ㅍ /pʰ/	5.45	1.57	1	7	5.22	1.63	1	7
ㅃ /p*/	5.18	1.67	1	7	5.16	1.66	1	7
ㅈ /tɕ/	5.12	1.65	1	7	4.89	1.72	1	7
ㅊ /tɕʰ/	5.16	1.71	1	7	4.84	1.82	1	7
ㅉ /tɕ*/	4.72	1.79	1	7	4.65	1.77	1	7
ㅅ /s/	5.51	1.55	1	7	5.48	1.54	1	7
ㅆ /s*/	4.74	1.74	1	7	4.61	1.77	1	7
ㅎ /h/	6.17	1.12	1	7	6.01	1.27	1	7
ㄹ /l/	5.24	1.89	1	7	5.78	1.46	1	7
ㅁ /m/	6.37	1.07	1	7	6.24	1.23	1	7
ㄴ /n/	5.85	1.60	1	7	5.75	1.69	1	7
ㅇ /ŋ/	5.43	1.78	1	7	5.38	1.75	1	7
ㅏ /ɑ/	6.55	0.90	1	7	6.53	0.92	1	7
ㅣ /i/	6.49	1.03	1	7	6.51	1.02	1	7
ㅐ /ɛ/	4.90	1.96	1	7	4.36	2.03	1	7
ㅓ /ʌ/	5.05	1.65	1	7	4.69	1.82	1	7
ㅗ /o/	5.05	1.70	1	7	4.64	1.84	1	7
ㅜ /u/	5.95	1.42	1	7	5.82	1.54	1	7
ㅡ /ɯ/	5.58	1.69	1	7	5.65	1.64	1	7
/w/	5.24	1.54	1	7	4.94	1.76	1	7
/j/	5.18	1.51	1	7	4.85	1.62	1	7

Note. Higher values = easier.

Phoneme-Level Differences between KPD Results and SA. One way of looking at the relationship between KPD results and learner SA is to compute difference scores. After converting phoneme parcel scores and SA easiness ratings to percentages to facilitate direct comparisons, I calculated difference scores by subtracting SA percentage scores from KPD percentage scores. Across phonemes, the mean difference was 16% (SD = 8%) for production and 9% (SD = 10%) for perception. The results of this analysis are presented in Table 7.5. Positive values indicate learners underestimated a phoneme's easiness (i.e., their accuracy was relatively higher than their perception of easiness) while negative values indicate an overestimation (i.e., their accuracy was relatively lower than their perception of easiness).

For many phonemes in each modality, learners were on average quite accurate. For example, learners showed only trivial gaps (-1%) between their perceptions and accuracy in perceiving /k^h/. However, in almost all cases, standard deviations were considerable, often greater than 20% or 30%. Even more crucially, the range of difference scores was generally large. At the extremes, there were learners who vastly overestimated the easiness of a phoneme (e.g., -100% for /t^h/ in production) or vastly underestimated their own accuracy (e.g., +100% for /l/ in perception).

Surprisingly, learners exhibited considerable differences between KPD scores and SA of the phoneme /ɛ/, especially in perception. This is likely attributable to some confusion introduced by the format of the SA (see Appendix F), which attempted to present the two Korean letters ㅖ and ㅐ as both corresponding to the phoneme /ɛ/ (which is the case in modern Korean, see Shin et al., 2013, Chapter 5). However, more conservative descriptions (and prescriptions) of Korean phonology do not include /ɛ/, instead featuring /e/ (front unrounded mid vowel corresponding to ㅖ) and /æ/ (front unrounded low vowel corresponding to ㅐ). I occasionally received queries

about this item, and I was somewhat puzzled when very advanced speakers deliberated for some time on this item before marking a middling degree of easiness. It appears that many learners were under the impression that the two letters corresponded to different phonemes, and that the self-assessment item was asking how well they could distinguish between the two phonemes.

Table 7.5

Differences between KPD Results and Learner Self-Assessments

	KPD Production – SA Production				KPD Perception – SA production			
Phoneme	Mean	SD	Min	Max	Mean	SD	Min	Max
ㄱ /k/	14%	23%	-21%	100%	6%	27%	-33%	83%
ㅋ /kʰ/	15%	28%	-67%	100%	-1%	30%	-83%	83%
ㄲ /k*/	9%	35%	-83%	100%	8%	31%	-75%	100%
ㄷ /t/	9%	27%	-56%	100%	16%	28%	-50%	100%
ㅌ /tʰ/	10%	28%	-100%	83%	2%	28%	-50%	83%
ㄸ /t*/	10%	34%	-83%	100%	14%	30%	-67%	100%
ㅂ /p/	18%	25%	-29%	100%	-3%	31%	-83%	67%
ㅍ /pʰ/	15%	27%	-80%	100%	5%	30%	-50%	100%
ㅃ /p*/	9%	32%	-83%	100%	5%	30%	-75%	83%
ㅈ /ʧ/	28%	28%	-25%	100%	8%	38%	-100%	83%
ㅊ /ʧʰ/	19%	33%	-100%	100%	12%	31%	-75%	100%
ㅉ /ʧ*/	5%	38%	-100%	100%	8%	32%	-75%	83%
ㅅ /s/	22%	26%	-20%	100%	-8%	30%	-63%	88%
ㅆ /s*/	19%	32%	-43%	100%	-4%	32%	-67%	67%
ㅎ /h/	13%	19%	-33%	100%	14%	22%	-25%	100%
ㄹ /l/	24%	31%	-33%	100%	12%	25%	-50%	100%
ㅁ /m/	10%	18%	-13%	100%	9%	21%	-33%	100%
ㄴ /n/	13%	26%	-40%	100%	12%	28%	-50%	100%
ㅇ /ŋ/	20%	30%	-50%	100%	11%	33%	-75%	100%
ㅏ /ɑ/	7%	15%	-7%	100%	5%	18%	-50%	100%
ㅣ /i/	8%	17%	-13%	100%	3%	20%	-67%	67%
ㅖ /ɛ/	33%	32%	-11%	100%	37%	36%	-33%	100%
ㅓ /ʌ/	19%	31%	-60%	100%	16%	34%	-67%	100%
ㅗ /o/	31%	28%	-8%	100%	17%	33%	-67%	100%
ㅜ /u/	15%	24%	-25%	100%	-11%	35%	-83%	83%
ㅡ /ɯ/	19%	29%	-50%	100%	12%	28%	-67%	100%
/w/	23%	26%	-30%	100%	25%	29%	-25%	100%
/j/	20%	25%	-44%	100%	16%	27%	-40%	83%

At the learner level, considering average differences between KPD and SA results across phonemes in each modality provides insight into a learner's overall level of SA accuracy. Figure 7.4 maps each learners SA accuracy along production and perception dimensions. A highly accurate learner will be near the origin; less-accurate learners will be located farther from the origin. Learners in Quadrant I tended to underestimate the easiness of Korean phonemes in both production and perception, while learners in Quadrant III tended to overestimate. Many learners were, on average, quite accurate—within 10% (inner ring)—for both production and perception. However, most learners exhibited greater average differences along each dimension, generally between 10% and 30% (middle ring). Interestingly, very few learners tended to underestimate the easiness of production while overestimating the easiness of perception. In other words, learners did not perceive production to be easy when they perceived perception (on average) to be difficult.

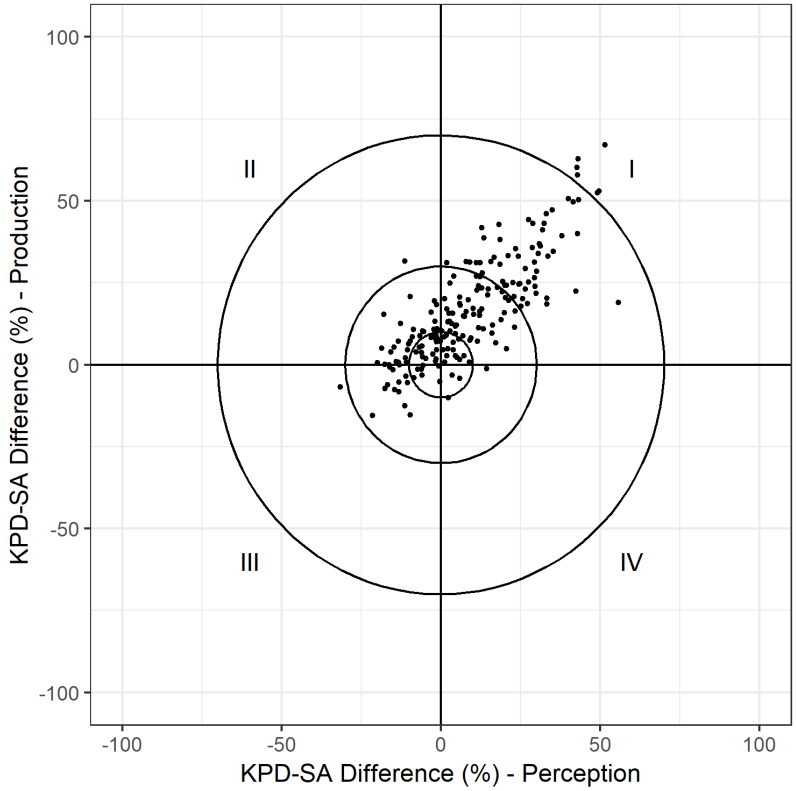


Figure 7.4. Mapping average learner accuracy for production and perception.

Correlations between KPD Results and SA. Another way of examining the relationship between KPD Results and learner SA results is by focusing on the strength of the relationship via correlations. The upper diagonal of Figure 7.5 contains Pearson correlation coefficients (*r*) among KPD accuracy scores averaged across phonemes, SA easiness scores averaged across phonemes, and SA responses for global pronunciation qualities (comprehensibility and accentedness). All correlations were significant at $p < .01$. The figure also features scatterplots for variable pairs (lower diagonal), and the diagonal shows density plots for each variable. The largest correlations were obtained between SA Production and SA Perception ($r = .88$) and KPD Production and KPD Perception ($r = .73$). The global SA measures of comprehensibility and accentedness were moderately correlated and had moderate correlations with averaged KPD Production and Perception scores. Other correlations were smaller; notably averaged SA perception had slightly stronger associations with KPD scores than SA production. The correlations between averaged SA production and averaged KPD scores, even for KPD production, were small.

Looking at finer-grained associations between SA and KPD results, Table 7.6 shows Pearson correlations between KPD results and SA for each phoneme in both modalities; Figure 7.6 shows scatterplots for these relationships. What is perhaps most interesting about these results is the number of small, statistically insignificant correlations. These have arisen generally due to restriction of range effects; for some of the easier phonemes (in terms of both KPD results and SA results), correlations appear to have been attenuated by a lack of variation (i.e., most participants rating the ease of a phoneme such as /i/ at 7/7 and notching very high accuracy scores on the KPD). The strongest phoneme-level correlations were found for objectively difficult phonemes (per KPD results, see Chapter 5) such as /k*, t*, p*/. However, some

relatively strong correlations (though still generally small to moderate in magnitude) were found for some broadly easier phonemes, like /n, ŋ, r/.

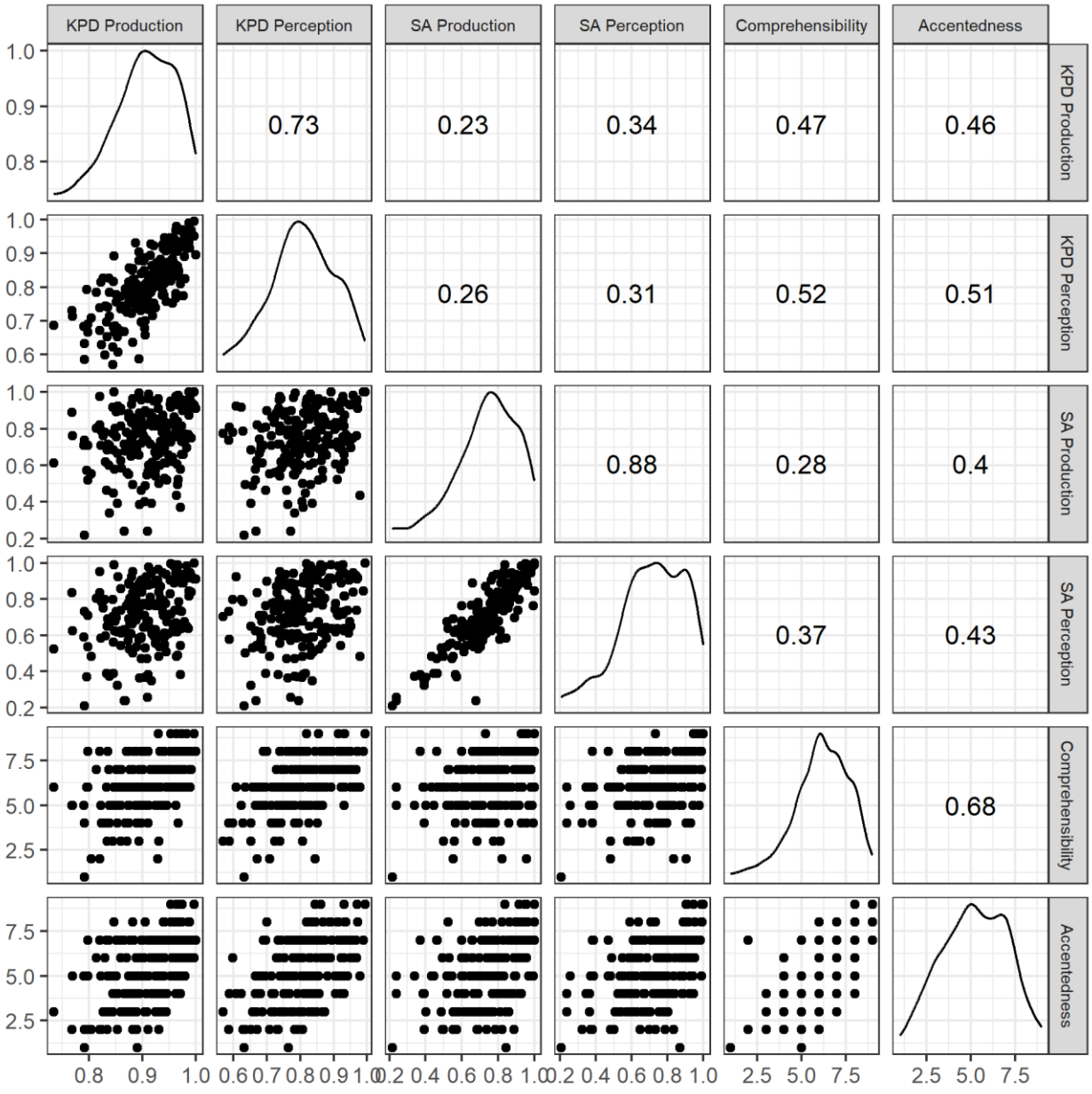


Figure 7.5. Relationships among average KPD scores and SA

Table 7.6

Correlations between KPD Scores and SA for each Phoneme

	KPD & SA Production				KPD & SA Perception			
Phoneme	n	*r*	*p*	95% CI	n	*r*	*p*	95% CI
ㄱ /k/	195	.05	0.463	[-0.09, 0.19]	195	.06	0.378	[-0.08, 0.20]
ㅋ /kʰ/	195	.21	0.004	[0.07, 0.34]	192	.35	0.000	[0.22, 0.47]
ㄲ /k*/	194	.16	0.024	[0.02, 0.29]	193	.25	0.001	[0.11, 0.37]
ㄷ /t/	195	-.02	0.772	[-0.16, 0.12]	194	.02	0.782	[-0.12, 0.16]
ㅌ /tʰ/	196	.27	0.000	[0.13, 0.39]	195	.27	0.000	[0.13, 0.39]
ㄸ /t*/	195	.28	0.000	[0.14, 0.40]	194	.36	0.000	[0.23, 0.48]
ㅂ /p/	195	-.04	0.572	[-0.18, 0.10]	194	.02	0.746	[-0.12, 0.16]
ㅍ /pʰ/	195	.25	0.000	[0.12, 0.38]	194	.25	0.000	[0.11, 0.38]
ㅃ /p*/	196	.35	0.000	[0.22, 0.46]	195	.34	0.000	[0.21, 0.46]
ㅈ /tɕ/	192	.13	0.067	[-0.01, 0.27]	192	.06	0.442	[-0.09, 0.19]
ㅊ /tɕʰ/	193	.08	0.265	[-0.06, 0.22]	194	.29	0.000	[0.15, 0.41]
ㅉ /tɕ*/	194	-.01	0.914	[-0.15, 0.13]	194	.27	0.000	[0.13, 0.39]
ㅅ /s/	196	.01	0.900	[-0.13, 0.15]	195	.02	0.805	[-0.12, 0.16]
ㅆ /s*/	196	.09	0.192	[-0.05, 0.23]	195	.16	0.024	[0.02, 0.29]
ㅎ /h/	193	.01	0.839	[-0.13, 0.15]	193	.07	0.362	[-0.08, 0.20]
ㄹ /l/	194	.19	0.009	[0.05, 0.32]	195	.16	0.027	[0.02, 0.29]
ㅁ /m/	196	.10	0.151	[-0.04, 0.24]	196	.07	0.340	[-0.07, 0.21]
ㄴ /n/	196	.29	0.000	[0.15, 0.41]	195	.24	0.001	[0.10, 0.37]
ㅇ /ŋ/	196	.16	0.021	[0.02, 0.30]	195	.16	0.029	[0.02, 0.29]
ㅏ /ɑ/	196	.00	0.986	[-0.14, 0.14]	196	.10	0.158	[-0.04, 0.24]
ㅣ /i/	194	-.04	0.585	[-0.18, 0.10]	194	.08	0.253	[-0.06, 0.22]
ㅖ /ɛ/	193	.17	0.018	[0.03, 0.30]	193	.06	0.382	[-0.08, 0.2]
ㅓ /ʌ/	195	.13	0.067	[-0.01, 0.27]	196	.18	0.011	[0.04, 0.31]
ㅗ /o/	194	.12	0.108	[-0.03, 0.25]	196	.18	0.010	[0.04, 0.31]
ㅜ /u/	196	.06	0.432	[-0.08, 0.19]	194	.10	0.157	[-0.04, 0.24]
ㅡ /ɯ/	194	.14	0.045	[0.00, 0.28]	194	.35	0.000	[0.22, 0.46]
/w/	194	.08	0.261	[-0.06, 0.22]	195	.23	0.001	[0.10, 0.36]
/j/	196	.20	0.00	[0.06, 0.33]	195	.22	0.002	[0.08, 0.35]

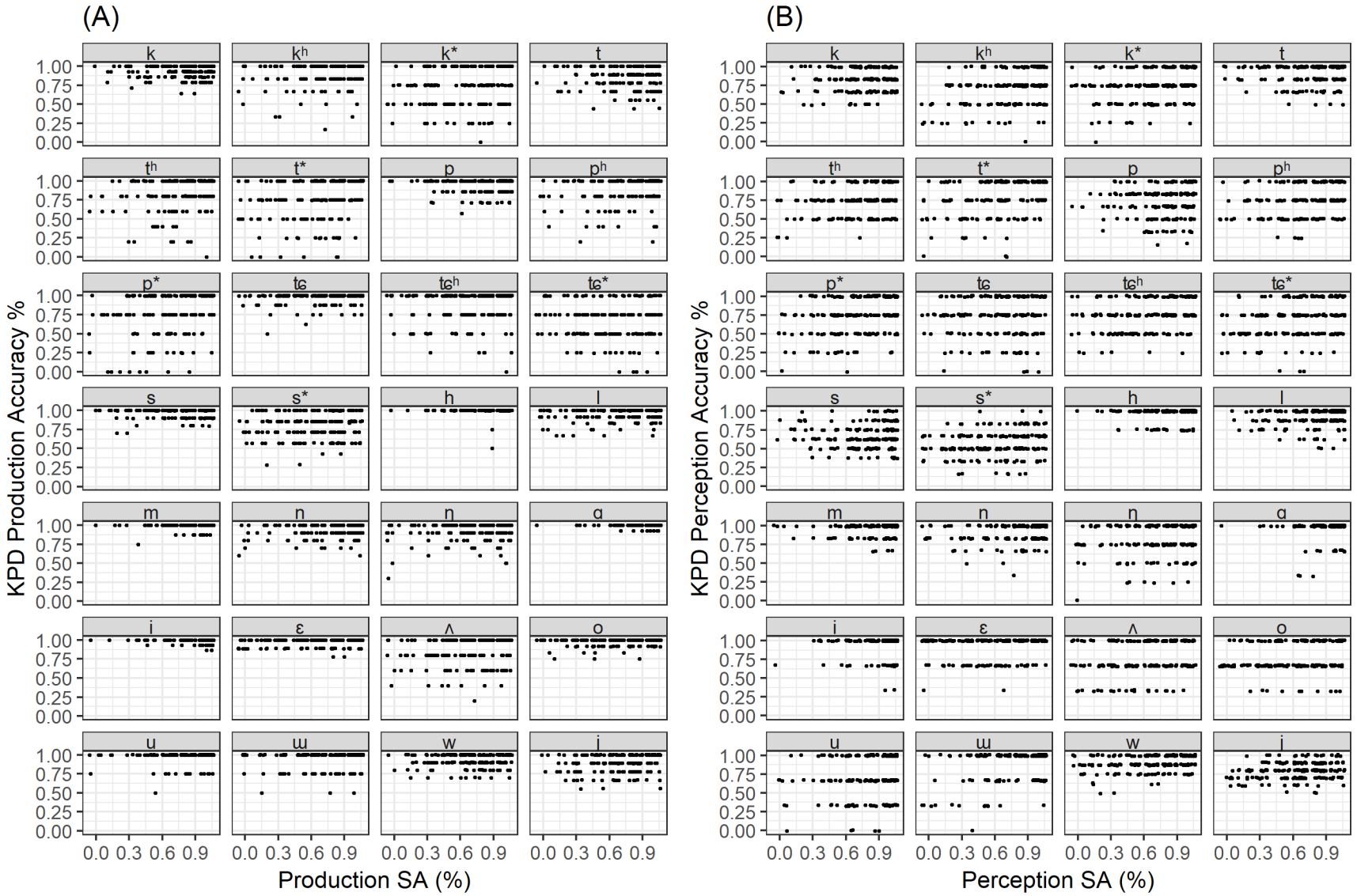


Figure 7.6. Scatterplots of KPD score and SA for each phoneme in (A) production and (B) perception.

Finally, it is worthwhile to consider the strength of association between SA and KPD results at the level of individual learners. This is interpretable as a measure of how well learners were able to discriminate between phonemes along a continuum of difficulty. For production, the average within-learner correlation between KPD and SA results was r = .21 (SD = .23, min = -.32, max = .79) for 194 learners (four learners had no variation in SA responses). For perception, the average within-learner correlation between KPD and SA results was r = .20 (SD = .25, min = -.42, max = .79) for 197 learners. Figure 7.7 illustrates the distribution of learner correlations between KPD and SA results. Learners who discriminated phoneme difficulty well, with positive associations between KPD and SA results for both production and perception, are located in Quadrant I. Learners with overall poor or misguided discrimination of phoneme difficulty are located in Quadrant III. While most learners showed some positive association for both production and perception, some seemed to have misperceptions about their strengths and weaknesses. Other learners could discriminate the difficulty of phonemes in production *or* perception, but not the other (Quadrants II and IV).

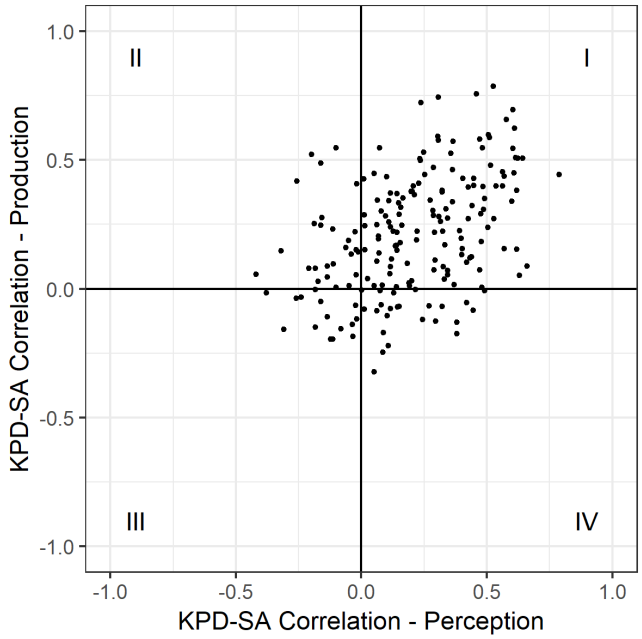


Figure 7.7. Mapping learner discrimination of phoneme difficulty for production and perception.

Agreement between KPD Diagnostic Flags and SA. As a final means of examining the relationship between KPD scores and self-assessments, I turned to the diagnostic flags for especially difficult phonemes. As readers may recall from previous chapters, I set a 75% accuracy threshold for diagnostic flagging of especially difficult phonemes. To compare these diagnostic flags with learner self-assessments, I dichotomized self-assessment scores using the same 75% threshold. In practice, this meant that a learner indicated a phoneme ease of 5 or less (out of 7, 7 = Never Difficult). From the two sets of binary phoneme scores (diagnostic flags and dichotomized self-assessments), I tagged matches between diagnostic flags and learner-recognized critical difficulties. Table 7.7 contains summary statistics for KDP diagnostic flags and learner self-assessment agreement.

Table 7.7

Summary Statistics for KPD Flagged Phonemes and SA Agreement

	# KPD Flagged Phonemes			# KPD-SA Matches			% KPD-SA Agreement		
Mode	M	SD	Range	M	SD	Range	M	SD	Range
Production	3.55	2.30	1 – 10	2.17	2.02	0 – 9	60	39	0 – 100
Perception	8.39	4.67	1 – 23	4.78	4.11	0 – 18	53	33	0 – 100

Due to the difficulty of KPD perception tasks (see Chapter 4, Measurement), learners on average had over twice as many perception phonemes flagged on the KPD compared to production phonemes. Learner recognition of these phonemes as being difficult was close to two-thirds for production and closer to one-half for perception. As might be expected, there was considerable variation among learners in their recognition of phoneme difficulties as revealed by the KPD, including a substantial number of learners who failed to recognize any of their difficulties. Out of 160 learners who had at least one production phoneme flagged according to

KPD results, 33 (21%) had self-assessments that failed to recognize the difficulty of any flagged phonemes. For perception phoneme flags, the number was 31 out of 193 (16%).

Discussion

In this chapter, addressing Research Questions 4, 5, and 6, I presented the results of analyses that compared KPD scores to external measures of oral proficiency, pronunciation in spontaneous speech, and self-assessments of pronunciation ability and phoneme difficulty. Research Question 4 and the analysis of oral proficiency primarily addressed the explanation of KPD scores, with the expectation that more proficient speakers will tend to have more intelligible production and more accurate perception of Korean phonemes. The remaining research questions and analyses primarily addressed the extrapolation of KPD results to pronunciation (and perception) to more general domains of Korean use. Research Question 6 also bears on the utilization of KPD scores, whereby weaknesses in learner self-assessments might be corrected by KPD scores. In what follows, I discuss the results in respect to each research question.

RQ4: To what extent do KPD results show an expected relationship with Korean oral proficiency?

Broadly, KPD results demonstrated relationships with Korean oral proficiency that were in line with expectations, providing support to the explanation inference in the KPD's validity argument. Specifically, medium-sized (Plonsky & Oswald, 2014) correlations were found between oral proficiency and average phoneme accuracy for both perception and production of Korean phonemes. When dividing learners into oral proficiency quantiles, a generally steady upward progression was found for average perception accuracy across phonemes, a pattern that

was somewhat less visible for production phonemes, though it is worth noting that production phoneme averages were higher overall.

For individual phonemes, larger correlations between oral proficiency and accuracy were found for perception phonemes compared to production phonemes, but in both modalities the strongest correlations were generally obtained for more difficult phonemes (see Chapter 4). Progression in phoneme accuracy across oral proficiency quantiles showed a similar trend: Easier phonemes had high average accuracy across quantiles, while more difficult phonemes showed an upward trend from the lower to upper oral proficiency quantiles (with a small number of exceptions related to universally difficult phonemes that tended to elude many of even the most advanced speakers). In sum, the findings here are in alignment with theory and research that suggests phonological competence develops with experience/instruction (Piske et al., 2001) and that some Korean phonemes tend to be more difficult and thus take longer to obtain control, difficulties which can persist into advanced stages of overall proficiency (Lee et al., 2009). In particular, at the group level tense and aspirated consonants tended to be more difficult for low-proficiency learners but became progressively less challenging for more proficient learners.

From the perspective of diagnostic utility, the presence of 'outliers' in these analyses are of great interest: Individuals who have phoneme (specific or averaged) accuracy out of line with the expectations for their overall oral proficiency range. For example, in the fourth and fifth oral proficiency quantiles (i.e., the highest 40% of oral proficiency), there were some learners with average phoneme production accuracy of 79% and 80%, respectively, which would entail the flagging of several phonemes as being difficult. The KPD would be of considerable utility for such learners, who despite having generally high levels of oral proficiency in Korean could nonetheless stand to benefit from targeted pronunciation study, study that their generally high-

proficiency peers might not need. Similarly, lower proficiency learners with excellent segmental pronunciation (e.g., with an average production phoneme accuracy in the 90-99% range) may not need as much segmental pronunciation instruction or individual study as some of their peers; these learners could confidently spend their time on other aspects of learning Korean.

RQ5: To what extent do results reflect difficulties test-takers show in spontaneous, meaning-focused speech?

Based on an exploratory, descriptive analysis of 21 learners, I observed a trend of learners with higher average KPD phoneme accuracy scores tending to produce phonological errors at a lower rate in their contemporaneous speaking. This provides some broad support for the extrapolation of KPD results to more naturalistic, meaning-focused Korean speaking performance. Of course, the speech samples I collected were not extensive and thus this support can only be taken as prospective.

Looking at the learner productions in greater detail, I examined the alignment of errors produced with corresponding KPD phoneme accuracy scores. This examination showed many cases of alignment, where some of each learner's most difficult phonemes according to the KPD appeared as a challenge in naturalistic speech. This alignment is encouraging and provides some support for extrapolating KPD results as a finer grain-size. However, due to limitations in the volume of spontaneous speech collected and other issues (e.g., different error criteria on the KPD and phonemic transcriptions, different abilities/knowledge being tapped), I cannot claim particularly strong support for extrapolation from this analysis. That said, the level of alignment observed between scores derived from the highly-discrete, non-communicative KPD and genuine (though limited) meaning-focused communication is encouraging.

RQ6: To what extent do results reflect self-assessments of pronunciation ability and difficulties?

While learners' self-assessments of global pronunciation abilities, i.e. accentedness and comprehensibility, were strongly related to their overall levels of performance on the KPD, finer-grained self-assessments were less accurate, a finding in line with research on pronunciation self-assessment in L2s such as French (Lappin-Fortin & Rye, 2014) and German (Dlaska & Krekeler, 2008). The challenge faced by learners appears to be in identifying the relative difficulty of Korean phonemes, as evidenced by the generally low average within-learner correlation between KPD parcel score and their self-assessed ratings in each modality. This interpretation is further supported by the alignment between diagnostic flags and learner self-assessments, where learners only recognized substantial difficulty in perceiving or producing between half and two-thirds (respectively) of phonemes that were especially difficult for them according to the KPD. However, it is worth noting that on average, learner absolute accuracy, based on difference scores, was not dismal: On average, learners averaged a 16% difference for production and 9% for perception. Thus, on a whole, I find that KPD scores have a moderate relationship with learner self-assessments, which suggests some meaningful extrapolation of the KPD results to pronunciation and listening in typical Korean use, moderated by learner awareness.

Indeed, I further suggest that the current findings bode well for the utilization inference in the KPD's validity argument: KPD results have the potential to heighten or fill in gaps in learner's self-awareness of *specific* pronunciation difficulties, which in turn could lead to more fruitful instructional decisions, attention to form in typical communication, or both. While my discussion of self-assessments so far has focused on sample means, it is worth pointing out the degree of variation in the sample: Several learners demonstrated major shortcomings in their

ability to accurately self-assess pronunciation difficulties. With such poor awareness of pronunciation (and perception) difficulties, it is unlikely these learners would be able to monitor their own productions, selectively pay additional attention to good exemplars in the input or make productive and efficient decisions when planning additional pronunciation-related study on their own time. While training in self-assessment has been shown to help self-assessment accuracy (Chen, 2008), research on whether it is feasible to train learners at a variety of proficiency levels to better self-assess their pronunciation difficulties is non-existent. It is these learners for which the KPD might be utilized to the greatest, most beneficial extent.

CHAPTER 8: INTERPRETATION AND USE

Moving further up the chain of inferences in the KPD's proposed validity argument, in this chapter I examine the interpretation and use of KPD results by stakeholders. I sought to understand how test-takers (and one teacher) were able to make sense of the KPD score report and interpret results meaningfully. Further, I investigated how test-takers applied information from the KPD score report over a 2- to 4-month period to explore how they were able to use the test results, and to uncover whether or how that use was beneficial. The data I present in this chapter are primarily qualitative, derived from face-to-face semi-structured interviews, but I also provide supporting quantitative information and analyses based on initial KPD results and, for a subset of students, KPD retest results.

Research Questions

For reader convenience, the three research questions I address in this chapter are as follows:

- RQ7: How do (a) teachers and (b) learners understand KPD score reports? To what extent do they learn anything new from KPD score reports?
- RQ8: Do learners report any changes in their self-study routines and/or their attention to phonological form in formal or informal learning situations?
- RQ9: Do learners show improvements in a) overall and/or b) in weak areas after receiving and applying KPD feedback?

My investigation primarily focused on learners' individual interpretation and utilization of results, rather than interpretation and utilization of test results in a classroom context with the guidance of a teacher. Although this is perhaps a shortcoming, I see it as a useful starting point, as individual students are the ground-floor, most immediately impacted stakeholders in any

assessment geared toward learning, and self-regulated learning can lead to desirable pronunciation learning outcomes (Moyer, 2014). Students can benefit from increased awareness of their pronunciation abilities (Kennedy & Trofimovich, 2010; Saito, 2018), and in practice many L2 (Korean or otherwise) classrooms make little time for pronunciation instruction. Additionally, due to diverse learner backgrounds and needs it may be difficult to arrive at suitable whole-class segmental targets (Derwing & Munro, 2014). Such conditions make learners' own efforts to study more autonomously worthy of interest.

In the following sections, I provide methodological details followed by my presentation and discussion of findings.

Methods

A primary description of the interview study procedures was reported in Chapter 4. What follows here are details on the interviewees, supporting KPD score reports, and analytical details.

Interviewees

As mentioned in Chapter 4, I interviewed a total of 22 individuals, including 21 learners and one teacher, who had taught two of the student interviewees in an intensive Korean program class. I refer to all interviewees with pseudonyms. Table 8.1 provides details on the 21 learner interviewees. Among the 21 learners, five were graduate students, four were enrolled in undergraduate programs (with two in English-medium programs), and 12 were enrolled in intensive Korean programs (one of these students was an exchange student also taking English-medium undergraduate courses). The learners represented eight different L1 backgrounds (with nearly half of the group identifying as L1 Mandarin Chinese speakers) and 11 countries of origin (distinguishing Hong Kong as a special region within China). Learners' time spent living in Korea at the time of their initial KPD ranged from approximately 1 month to nearly 11 years.

The teacher, a male native speaker of Korean who I will refer to as Jae-woo, taught Yu-wen and Yuki in different classes over the previous two semesters, and I interviewed him on November 21, 2018.

For test-takers, the first interview generally occurred within a couple of weeks of their field-testing appointment. Fourteen test-takers were available and willing to complete a second interview and take the KPD again. These second interviews took place roughly three months after each participant's first interview (mean = 3.16 months, min = 2.33 months, max = 4.30 months). After each appointment, participants received 10,000 KRW (approximately $10 USD).

Score Reports

The KPD score reports detailed in Chapter 3 were provided to all interviewees during the first interview and revisited as needed in the second interview (see Figure 3.1). However, I made one small change to the score reports given to learners: I selected a threshold of 80% (rather than 75%) for flagging critical phonemes that would appear on the first page. Anticipating a generally higher level of both general Korean proficiency and specific pronunciation ability compared to my pilot sample, I was worried that too many test-takers would receive little in terms of helpful feedback with a 75% diagnostic flag criterion. I chose to err on the side of strictness (e.g., for some phoneme parcels making just one mistake could result in a flag) in order to provide more test-takers, especially those with mostly (but not universally) intelligible pronunciation, with at least some prescriptions for study, as the feedback was advertised as a benefit of participating in the study. In terms of scoring, this did not change anything, and raw accuracy scores on the second page of the reports were unchanged. However, it did introduce some slight changes in interpretations and (potentially) decision-making for some phonemes with scores from 75-79% in comparison with flags reported in previous chapters.

Table 8.1

Interviewees

Pseudonym	Sex	Age	From	Languages[b]	Acad. Status.	EIT[d]	LOR[e]	Interview 1	Interview 2
Graduate Students									
Min	F	23	China	Chinese, Korean, English	KFL (MA)	101	0;1	10/1/2018	1/7/2019
Hoa	F	24	Vietnam	Vietnamese, Korean, English	KFL (MA)	73	0;1	10/1/2018	1/8/2019
Ju-an	F	25	China	Chinese, Korean, English, Japanese	International Trade (MA)	91	1;0	11/5/2018	1/29/2019
Yang	F	30	China	Chinese, Korean	Hospitality (PhD)	87	10;10	11/5/2018	-
Amber	F	23	Hong Kong	Cantonese, English, Chinese, Japanese, Korean	KFL (MA)	96	0;8	11/6/2018	2/1/2019
Undergrad. Students									
Leo	M	26	Russia	Russian, English, Korean	International Studies[c]	38	5;0	10/3/2018	1/15/2019
Xiu Lan	F	23	China	Chinese, Korean, English	KFL (BA)	69	0;8	10/10/2018	-
Sofia	F	23	Belarus	Russian, Belarussian, English, Korean	Business Management	75	2;0	10/10/2018	1/29/2019
Fang	F	21	China	Chinese, Korean, English, Japanese	KFL (BA)	70	1;6	10/11/2018	1/8/2019
Language Students									
Holger	M	28	Germany	German, English, Korean	Level 3	45	0;2	8/29/2018	11/7/2018
Jing	F	23	China	Chinese, Korean, English	Level 4	74	0;11	8/31/2018	1/7/2019
Chia-ling	F	19	Taiwan	Chinese, Korean, English	Level 4	79	1;0	9/5/2018	-

Table 8.1 (cont'd)

Maria	F	23	Mexico	Spanish, English, French, Korean	Level 2, International Studies[c]	18	0;6	10/5/2018	12/17/2018
Noriko	F	29	Japan	Japanese, Korean	Level 5	59	0;6	10/8/2018	1/7/2019
Sakura	F	48	Japan	Japanese, Korean, English	Level 3	67	0;1	11/5/2018	2/13/2019
Yu-wen[a]	F	28	Taiwan	Chinese, English, Korean	Level 5	33	0;11	11/5/2018	-
Aylin	F	19	Kazakhstan	Russian, English, Korean, Kazakh	Level 5	62	0;8	11/6/2018	-
Na	F	23	China	Chinese, Korean, English	Level 5	57	1;0	11/9/2018	2/13/2019
Yuki[a]	F	21	Japan	Japanese, Korean	Level 4	50	0;7	11/14/2018	-
Alice	F	22	France	French, English, Korean	Level 2	47	0;2	11/15/2018	-
Xiu Ying	F	20	China	Chinese, Cantonese, Korean, English	Level 5	72	0;5	11/9/2018	2/14/2019

Note. [a]Jae-woo taught Yuki (Fall 2018) and Yu-wen (Summer 2018). [b]Self-reported, in order of dominance. [c]English-medium degree. [d]Elicited Imitation Test (oral proficiency measure, scale 0-120). [e]Length of Residence at time of initial KPD testing in years;months.

KPD Retesting

For each of the 14 participants who completed a second interview and KPD retest, I calculated their average production and average perception accuracy across all phonemes at initial test and retest, and computed change scores (retest minus initial test). I also examined their production and perception phoneme flags at initial test and retest, tallying the total number of flags. At the group level, I computed descriptive statistics. At the individual level, I focused on production phoneme average accuracy over time, and further analyzed the production flags by examining which flags were lost or gained from initial test to retest. I then interpreted these analyses alongside learners' comments about learning activity and perceptions of change (see following section).

Analysis of Interview Data

The 36 interview sessions took 16 hours (970 minutes) in total. All interviews were transcribed in the originally-used language(s). I used two approaches to transcribing the interview data: manual transcription (i.e., completed by myself or a research assistant) and manually-checked automated transcription (i.e., using automated transcription software such as *Vocalmatic*, www.vocalmatic.com, followed by manual correcting). Interviews remained in the original languages in my subsequent analysis of the data.

My approach to analyzing the interview data was primarily qualitative content analysis, "a method for systematically describing the meaning of qualitative data … by assigning successive parts of the material to the categories of a coding frame" (Schrier, 2014, p. 170). More specifically, I took a deductive approach to content analysis (Elo & Kyngäs, 2007; Schreier, 2014), utilizing pre-established categories based on my validity argument-driven research questions and initial review of the interview data. I created a spreadsheet with one row

for each interviewee and columns for relevant categories that served as a *matrix display* (Miles, Huberman, & Saldaña, 2014) or *coding frame* (Schreier, 2014) for analysis of the interview data. This frame layout facilitated cross-case comparisons, allowing me to see similarities and differences across the pool of participants. The categories included *understanding of results, alignment of results with own assessments, typical pronunciation learning/teaching, potential learning activity, actual learning activity,* and *changes in pronunciation*; the latter two categories only applied to learners who completed a second interview. For each topic, I made notes, in English, on what each interviewee said on the topic and compiled illustrative interview excerpts.

Findings

In this section, I present my main findings related to stakeholders' utilization of their KPD results. I start by discussing the learners' understanding and potential for applying KPD results in their continued Korean learning efforts. Then, I turn to the comments of the teacher of two of those learners to consider a more expert perspective on understanding of test results and more conventional classroom-based application of results. Finally, I focus on the second interview and KPD retest data from 14 learners to explore the actual utilization of KPD results and the results' impacts on pronunciation learning. Throughout the findings, I present quotations and excerpts from interviews. If the original comments were in Korean, I provide my English translation followed by the original Korean in parentheses; my translations prioritize meaning and do not attempt to reflect any form-related infelicities present in the original Korean. Occasionally, I lead with original Korean to emphasize linguistic choices made by learners in their comments. I represented Korean letter/sound names with IPA symbol equivalents (e.g., ㄱ, spoken as '기역' /ki.jʌk/, = /k/). For English comments, I provide interviewee's original words

without correcting any lexical or grammatical infelicities. Where I felt it was necessary, I inserted bracketed contextual information or corrections into comments.

Learner Understanding of Results and Potential Application

The following findings are based on cross-case analysis of learner comments, primarily based on their initial reactions to receiving their KPD score reports (see Figure 3.1 for an example) in the first interview.

Interpretation. At a basic level, learners understood that the KPD results provided information on their pronunciation strengths and weaknesses, and learners tended to focus more on the latter. All learners recognized that the phonemes highlighted on the first page were their weaknesses; in Korean they often used terms like "약점" (*weak point,* Xiu Ying) or "문제점" (*problem (point),* Ju-an and Na). Learners latched on to accuracy scores and example words from the second page of the score report, especially for phonemes with low scores. Furthermore, many learners readily thought of these sounds as targets for study and improvement: "Now I know what I should work [on]" (Maria), "After learning what my mistakes are, I can fix them" (저의 실수를 아는 후에 그 실수를 고칠 수 있어요, Aylin).

Across learners, I commonly found general agreement with the information provided by the KPD. Comments indicating broad acceptance of results were common: "After seeing this it really seems right" (이것을 봤을 때 진짜 맞는 것 같아요, Fang), "What I thought was difficult all came out with scores like that" (제가 어렵다고 생각했던 것은 다 그런 점수도 나왔으니까, Yuki), "Ah, as I expected, I am right about the pronunciations I think are difficult" (아 역시, 제가 어렵다고 생각하고 있는 발음 다 맞아요, Noriko). I believe this broad acceptance may be related to learners' epistemic orientation to the KPD results. Namely, learners appeared to regard the results as valid, due to objectivity or externality, and able to fill in gaps in

their own self-knowledge. I offer support for this interpretation through the following illustrative excerpts:

> Evaluating based on my own impressions isn't objective. A Korean saying she's lacking with this or she needs to practice that more are more effective results… In my opinion, I'm not sure about my own evaluation of myself. It's just based on my own thoughts. (자기 생각대로 평가하면 좀 객관 되지 않은 거예요. 한국 사람이나 이거 이 친구는 이거 부족하구나 이 친구는 이거 더 연습해야 되구나 그거는 더 효과적인 결과가 … 제 생각으로 제가 평가한 거니까 좀 모르겠어요… 그냥 자기 생각으로 한 거니까, Hoa)

> But actually it's the first time [getting the KPD results], like when someone like, tell me about like my pronunciation. Yeah. That is why I really wanted to know that because like how our teachers, they never like did it. (Sofia)

> Now I know the problem before I didn't know the problem and just like, okay, they [people I talk to] don't understand [me], maybe because – Like now I kind of know what kind of problems do I have. (Leo)

However, learner interpretation of the KPD results was not without limitations. A commonly occurring obstacle to understanding results appeared to be a lack of familiarity with linguistic vocabulary. On the first page of the KPD results, supplemental information on difficult articulatory features and contexts was provided. Few learners knew terms such as 경음 (*tense)*, 격음 (*aspirated*), or 파찰음 (*affricate*) that were used to label features or terms such as 종성 (*final consonant*) used to label contexts. Only some of the more advanced learners, particularly but not exclusively those pursuing degrees in Korean as a second/foreign language, were immediately able to understand what these terms meant, such as Amber and Yu-wen. Interestingly, some learners with more informal/self-directed Korean learning histories had difficulty talking about Korean sounds, having never formally learned the names of Korean letters, though they clearly had adequate enough sound-symbol correspondence knowledge to interpret the scores (instead of using letter names for consonant names she did not know well, learners such as Sakura and Chia-ling simply constructed a short syllable of /<consonant> + ɯ/

to refer to phonemes; this would be like saying “wuh” to represent the letter *w* in English). Also related to linguistic deficiencies, lower proficiency learners such as Alice (who was enrolled in a Level 2 Korean language class) demonstrated some difficulty understanding the explanatory prose on the KPD report.

Occasionally, information in the KPD score reports was difficult for students to reconcile. Rarely, their score on a sound or feature, in production or perception, was so different from their self-appraisal of pronunciation abilities that they voiced some disagreement or disbelief. Several learners were surprised about their low perception scores, either generally or in reference to specific sounds. However, three learners, Jing, Chia-ling, and Sakura, all expressed difficulty in accepting that their listening (perception) was worse than their pronunciation. In each case, I elaborated on the different scoring criteria for the production and perception sections. I also commented on how narrowly *listening* was being operationalized on the KPD.

Other learners voiced more specific disagreement with the information provided by KPD results. For example, Aylin believed her production accuracy scores for /o/ (100%) and /ʌ/ (20%) were a reversed representation of her actual pronunciation of those two vowels. Similarly, Fang had trouble accepting that she had difficulty pronouncing tensed consonants (KPD production parcel scores: /k*/ = 50%, /s*/ = 57%, /t*/ = 75%, /ʧ*/ = 75%), genuinely believing that she had little trouble producing them. In one case of disagreement with a KPD score, a learner referred to an external assessment of her pronunciation by a teacher: Xiu Ying said " But my teacher said that in my last presentation my /l/ sound wasn’t clear" (근데 선생님이 제가 지난번 발표할 때 ㄹ [/l/] 소리가 잘 안 나와 가지고 그렇게 말했어요). Na indicated some disagreement with her low score on /k*/. Her first reaction was that she could produce that sound with little difficulty. In fact she demonstrated that to me in the interview, producing example words from her score

report, 토끼 (/t^{h}o.k*i/) and 꿀 (/k*ul/), intelligibly and accurately (to my ears, at least). As we discussed this further, she appeared to come to a realization, or offer a concession, that her knowledge of articulation might not always match up with her accuracy in production: "Although I know how to pronounce it, I might not [always] be that accurate" (하지만 어떻게 발음하는지 알고 있는데 그렇게 정확하지 않아요).

New Information. Rare disagreements aside, learners' broad acceptance of KPD results led to many discrepancies with their prior self-appraisals to be considered as new information to process and incorporate. Almost all interviewees expressed surprise at—but not rejection of—some piece of information contained in their score report. In some cases, there was no surprise that a given sound was difficult, but learners were nonetheless surprised at the degree of difficulty it presented. For example, Aylin readily agreed that several tensed sounds were difficult for her to produce but expressed some shock at scores of 0% for /t*, p*, ʧ*/. Sakura was similarly surprised by her perception score of 0% for /u/ but was grateful to now be aware of how acute that difficulty was. In both the first and second interview, Hoa was appreciative to learn that she had difficulty distinguishing between /p/ and /p^{h}/ in her production. Other learners, such as Holger, were surprised by the overall number of pronunciation difficulties identified by the KPD. In the first interview, Holger commented that he did not think his pronunciation was a big obstacle to being understood, referring instead to vocabulary and grammar as being bigger challenges. In my personal experience, though, I had considerable difficulty understanding Holger's pronunciation, and clearly the Korean teacher who scored the KPD often found his articulations ambiguous. Learners often viewed these surprises as targets and/or motivation for improvement. Hoa had an especially even-keeled yet highly motivated reaction to the surprises in her results:

> I always think I have to keep up my efforts. It's not a matter of feeling bad or feeling good [about the results]. The only thought that came to mind was "Wow, I really have to practice and study more." (항상 노력해야 된다고 생각해요. 기분 나쁜 거 아니고 그냥 좋은 거도 아니에요. 그냥 더 많이 연습하고 공부해야 됐구나 라는 생각만 들었어요.)

Sometimes the surprises were pleasant, providing learners with a boost in confidence or an opportunity to reappraise their abilities. Consider the following excerpt from my interview with Xiu Ying, who had voiced some disagreement about the KPD's assessment of her /l/ pronunciation:

> Dan: If you look at the back side, there are more detailed results. So /k/-
> 그 뒤쪽에서 보면 더 자세한 결과 나와요. 그래서 ㄱ은-
> Xiu Ying: [gasp] Really?!
> 진짜요?!
> Dan: Yes, 100%- It came out as 100% accuracy
> 네 100%- 100% 정확도 나왔어요, 이거
> Xiu Ying: When I was learning [Korean] it was what I thought was the most difficult…
> 네가 배울 때 이거 제일 어려운다고 생각했는데

Here, Xiu Ying's reaction seems to be more reflective of pleasant surprise, perhaps in reference to overcoming her initial struggles articulating /k/ without fulling realizing it herself. At a more general level, Na commented that:

> Pronunciation accuracy came out higher than my expectations. Maybe because of that accent of mine, my pronunciation confidence isn't so high and for the first time seeing pronunciation scores coming out on the higher end has me feeling pretty good.
> (발음 정화도 예상보다 더 높은 편이 나왔어요. 아마 그 억양 때문에 자기 발음 자신감 그렇게 높지 않아 가지고 처음으로 이런 좀 높은 편인 점수 나와 가지고 좀 기분이 좋아요.)

In Na's case, the pleasantly surprising results may provide correction for her perhaps undeservedly low confidence in her pronunciation abilities. Noriko, another learner with low confidence in her pronunciation ability and many genuine difficulties, was very glad to see high scores for phonemes such as /tɕ, tɕ^{h}/. To me, these comments highlighted how not just

weaknesses could be potentially informative or otherwise useful to learners. Clearly, to some extent at least, learners were interested in their underappreciated strengths as well.

Potential Application. Before considering whether and how learners thought they might apply their KPD results, it is worth briefly reviewing what learners said about their typical pronunciation learning activity. First, several participants were not enrolled in any formal Korean language courses at the time of field testing and the interview(s). These learners generally reported not having any current pronunciation learning activity outside of daily-life Korean interaction (e.g., in academic or social settings) and consuming Korean media such as television dramas and pop music. Second, some of the graduate and undergraduate students were enrolled in degree programs for Korean as a second/foreign language; these programs train students to be translators, interpreters, or Korean language teachers. As such, some students were taking courses on Korean phonology, teaching Korean pronunciation (which included material on Korean phonology), or both. Some of these learners reported recording their pronunciations in connection with course assignments and receiving feedback from their instructor. Third, the learners who were taking classes in intensive Korean language programs generally reported minimal attention to pronunciation in their courses. When I asked them about their typical pronunciation learning activity in their Korean classes, they most commonly referenced general speaking activities with their classmates, instructors incidentally addressing major pronunciation mistakes during read-aloud activities, and occasional choral repetitions, mostly of single words (i.e., commonly-used controlled pronunciation activities: Baker, 2014; Celce-Murcia et al., 2010). Outside of class, learners mostly mentioned watching dramas and perhaps trying to shadow lines, if they did anything at all. This information will be useful for interpreting their comments on what they might (and later, did) do after receiving their KPD results.

Turning to what learners said about potential subsequent pronunciation learning activity, which was framed as "study" (공부) and "practice" (연습) in interview questions, the majority of learners (n = 19) said they wanted to study or practice their pronunciation. Some learners, such as Hoa, commented on how the KPD results will help narrow down study targets: "Before coming here [to the interview], I always felt I had to study more. Everything. Now that I've come here, I understand which areas I should focus on more" (여기에 오기 전에도 항상 항상 더 연습해야 된다고 생각해요. 모든 거 다요. 오늘 와서 어떤 부분에 더 집중해야 된다는 것을 알게 되었어요). When it came to specific approaches or techniques for study and practice, learners came up with several ideas: using a textbook (Leo, Noriko), speaking Korean more with friends and getting feedback (Min, Ju An, Xiu Lan, Maria, Aylin, Na), reading aloud and/or self-recording (Hoa), watching dramas (Chia-ling), and asking a teacher or tutor for help (Maria, Noriko).

However, one common finding was a lack of knowledge about how to study pronunciation. Although learners were perhaps put on the spot to come up with something during the interview, many outright confessed that they did not know what to do that would help their pronunciation (Amber, Sofia, Fang, Holger, Jing, Yu-wen, Alice). Another reoccurring comment was that pronunciation practice was something they could not do on their own (Holger, Maria, Sakura, Alice), as they saw no way of getting feedback on whether they were pronouncing clearly or not. When I gave learners an opportunity for their own comments or questions at the end of the interview, several learners asked me for advice or additional ideas for studying pronunciation (Amber, Sofia, Fang, Jing, Noriko). In my responses, I mentioned activities such as shadowing, recording one's own pronunciation and comparing to a model, using a textbook to focus on difficult sounds, using a program like Praat (for Amber specifically, who was familiar

with the program and had a strong base in Korean phonetics and phonology), and doing listening practice such as those in pronunciation textbooks or dictation.

A Teacher's Perspective

Interviewing Jae-woo added a valuable perspective to the understanding and potential utilization of KPD results. Jae-woo taught Yuki in Level 4 of an intensive Korean program during the Fall 2018 semester and taught Yu-wen in Level 4 during the Summer 2018 semester. Before the interview, I obtained permission from Yuki and Yu-wen to share their information with Jae-woo. At the time of the interview, Jae-woo's semester with Yuki had recently finished and it had been approximately three months since he had taught Yu-wen. During the interview, I asked Jae-woo if he would like me to play a sample of Yu-wen's speech (the Independent Speaking task) to jog his memory. Jae-woo said that he could remember, but that hearing the speech sample would help with his memory accuracy, so I played Yu-wen's file before asking Jae-woo to reflect on her pronunciation. To facilitate comparisons between learner, teacher, and KPD perspectives on pronunciation difficulties, I have summarized and compiled the information in Table 8.2. The self-assessment column contains phonemes that the students indicated were especially difficult to produce on their paper self-assessment. The teacher observations are based on Jae-woo's interview comments, and the KPD results are based on diagnostic flags for production phonemes and supplemental information (with a < 80% criterion) from the first page of the score reports.

Table 8.2

Multiple Perspectives on Pronunciation Difficulties

	Self-Assessment*	Teacher's Observations	KPD Results
Yuki	**/k***, **d***, **t**, p, **tɕ***, s*, n, ŋ, l, **ʌ**, **o**, u, ɯ, w, j/	*Phonemes:* /**o, ʌ/**, typical Japanese L1 influences on other phonemes *Contexts:* Syllable coda, consonant clusters *Other:* Unexpected pitch-accent	*Phonemes:* / **tɕ^{h}**, k^{h}, p^{h}, p*, **k***, **t***, **tɕ***, **t/** *Features:* Aspirated, Tense, Affricate *Contexts:* Initial Consonant
Yu-wen	/k*, **tɕ**, **tɕ***, **s***, ŋ, **l**, w, **j/**	*Phonemes:* /**l**, u, o/, broad L1 Chinese interference *Contexts:* **Syllable coda** (esp. /l/) *Other:* lack of facial expression, lack of gesture, muted physical articulation of speech sounds	*Phonemes:* /**s***, t, p*, ʌ, **l**, ŋ, **tɕ**, **tɕ***, **j/** *Features:* tense, fricative, sonorants *Contexts:* **final consonant**

Note. Bolded elements indicate agreement among two or more sources. *Both learners had a median and mode of 4 (out of 7) on their self-assessments; phonemes shown were rated at 3 or below in production.

Interpretation. As an experienced teacher with strong knowledge of Korean phonology, Jae-woo considered the KPD results more critically than the learners. At first glance, Jae-woo was not quite sure how to interpret the information on the KPD, remarking that:

> At first, not knowing how the mechanism worked for scoring these two [Yuki and Yu-wen], the results were a little vague to me- when I looked at it, everything said "pronunciation is difficult" and it seemed like anything that was difficult for foreigners was included. (일단 그 둘이 어떤 그 메커니즘으로 만들어졌는지 제가 정확히 모르기 때문에 이 결과에 대해서도 역시 조금 막연하다 막연한 부분이 있는데 볼 때는 다 발음하기 어렵다는 외국인들이 발음하기 어려운 발음들이 다 대부분 포함이 되어 있는 거 같아요.)

However, as we talked more about the results and Jae-woo asked several detailed questions about how scores were calculated, standards for scoring production and perception, example words in the first page explanations and on the second page example column, etc., and became more familiar with the structure of the test, he seemed to move past initial skepticism and "got a

feel for what it was about" (어떤 걸 얘기한는구나를 느낄 수 있었어요). Like some of the learners had commented, he saw the KPD information as filling in gaps in what he was able to observe or perceive:

> The biggest reason [the results are useful] is that even though I know the students and what their pronunciation difficulties are, like we've talked about here, there are limits on content and limits on pronunciations that I hear. Right? Like during reading class time I can hear students and when we share a new text it would be great if I could judge how students pronounce sounds that are in that text, but the fact is the classroom environment isn't like that. Considering the education is centered on must-teach grammar and sentence patterns, what I didn't know about the students' pronunciation, even though it appeared in the [KPD] results, works out to about 50%. (가장 큰 이유는 제가 학생들을 알고 있고 그 사람들의 발음이 뭐가 문제다라고 여기서 이야기 있긴 하지만 제한된 내용과 제한된 발음을 들을 뿐이에요 그죠 제가 뭐 읽기 시간을 통해서 들을 수도 있고 새로운 텍스트를 나눠주고 텍스트 안에 있는 여러가지 소리 듣고 판단하면 좋겠지만 사실 수업 환경이 그렇지 못 하잖아요. 가르쳐야 하는 문법 문형 중심의 교육이다 보니까 학생들이 발음하는 것들은 결과에서도 그대로 나타났지만 저는 그 학생들이 가지고 있는 발음 문제점에 한 50% 정도 밖에 모르고 있었던 셈이죠.)

Jae-woo's comments here mirror what Lado (1961) wrote about the limits of a teacher's observations in identifying a full range of specific learner difficulties. Further, what is in theory possible to accomplish in the classroom will not always happen, and language education which prioritizes other aspects of linguistic competence will impose additional limits on what even a knowledgeable and conscientious teacher can achieve through observation of students.

New Information, Gaps, and Incongruencies. The interview with Jae-woo provided a unique opportunity to triangulate the self-assessments of learners and the KPD results. Here, I highlight new information introduced by the KPD, gaps in all three assessments, and incongruencies among the three sources. It is worth pointing out that although I consider the KPD results to be generally reliable and reflective of pronunciation and perception abilities (see Chapters 3-6), gaps and incongruencies among assessments here should not, by default, be settled by what the KPD results say, as the KPD has measurement error, an arbitrary diagnostic

flag criterion (though seemingly appropriate, see Chapter 5), and other limitations. It is quite possible that Yuki, Yu-wen, and Jae-woo made accurate observations that the KPD distorted or failed to detect.

First, based on the information in Table 8.2, it was clear to me that both learner self-assessments and the KPD, both of which featured items for every Korean segment, resulted in much more detailed information related to individual phonemes. Yuki's self-assessments of production difficulties had moderate alignment with KPD results, with several phonemes showing up as difficulties on both, and Yu-wen's self-assessment was remarkably well-aligned with her KPD results. Although Jae-woo did identify a few specific phonemes that were troublesome for both Yuki and Yu-wen, he broadly characterized their segmental difficulties as L1-driven: Yuki had "the errors that Japanese speakers have when pronouncing Korean" (일본어 화자가 한국어를 발음할 때 나타나는 오류들이 그대로 있는 편인데) and Yu-wen had "all the difficulties that generally appear for Chinese speakers when they learn Korean" (그 중국어를 사용자들이 한국어 배울 때 나타나는 문제점들이 전반적으로 들어가 있어). In this sense, there was little specific overlap that I was able to observe between Jae-woo's segmental observations and either students' self-assessments or KPD results. There was some congruency between his assessment of Yu-wen's difficulties related to pronunciation contexts and the KPD, as Jae-woo's observation of her difficulties with syllable coda pronunciation (particularly for /l/) aligned with the KPD supplemental results about final consonants. However, he viewed Yuki as having a similar problem with codas and particularly with consonant clusters (which can be found in sequences such as CVC.CV in Korean), which he attributed to Japanese having highly-restricted codas, an observation not reflected in KPD results. Curiously to me, Jae-woo made no specific comments about generally difficult articulatory features (e.g., tense,

aspirated) for either learner. Any observations he might have had were possibly subsumed under his comments related to L1 influence (e.g., Japanese phonology lacks a tense feature for consonants).

While the KPD clearly provided more details related to segments, what I found most interesting about Jae-woo's assessments of Yuki's and Yu-wen's pronunciation were aspects not covered by the KPD. For Yuki, Jae-woo talked at some length about her use of pitch accent and gave examples of how her pitch accent differed markedly from standard Korean. He went on to attribute this to her specific Osaka variety of Japanese. For Yu-wen, Jae-woo made comments about her muted oral articulation ("When pronouncing, she doesn't try to open her mouth much." 발음할 때 입을 크게 벌리고 노력하지 않은 편이에요.), which he attributed to an introverted personality. In our interview, Yu-wen revealed what she described as a complete lack of confidence in her pronunciation; Jae-woo appeared to be cognizant of this. Jae-woo saw Yu-wen's muted style of communication extending to supporting strategies, noting that Yu-wen did not utilize much facial expression or gesture when she spoke. He thought such strategies could help an interlocutor cope with her sometimes unintelligible pronunciation (at one point, Jae-woo commented that Yu-wen was only about 70% as intelligible as Yuki, who despite occasional errors was generally intelligible in communication). Thus, I found Jae-woo's observations, while not as fine-grained at the phoneme level, to contribute to a more well-rounded understanding of Yuki's and Yu-wen's pronunciation challenges.

There was also information not provided by the KPD that Jae-woo would have liked to know, perhaps to further support his limited opportunities to observe student pronunciation in detail: Phoneme-level information on difficult pronunciation contexts, e.g., whether a learner's pronunciation difficulties with /l/ were related to syllable codas (as he had observed with Yu-

wen). Jae-woo reiterated this several times throughout the interview. Such comments relate to the issue of grain-size in diagnostic assessment, and clearly, Jae-woo hoped for even finer details.

Potential Application. As with the previously discussed learner findings on potential application of KPD results, it is first crucial to consider Jae-woo's comments on typical pronunciation teaching as well as his beliefs related to pronunciation teaching. With respect to the latter, Jae-woo placed the greatest importance on helping learners be able to communicate with Koreans. While the framing of communicating with Koreans might be seen as narrow or prioritizing nativelikeness, Jae-woo's original comment in Korean ("한국 사람들과 의사소통이 **가능한** 수준", *level at which communication with Koreans* ***is possible,*** emphasis mine) is something I interpreted more or less oriented toward intelligibility in line with Levis (2005). Jae-woo also stated his awareness of the importance of pronunciation in language education academics and research but felt that level of importance has not really entered Korean teaching practice. Jae-woo described his typical pronunciation teaching as follows:

> For example, when teaching lower levels where there is more focus on form, that part has some exclusive pronunciation practice and [for example] after the instructor reads [a word] the students repeat. Or, when presenting a sentence to highlight syntax, the instructor reads aloud and the students follow along and then instruction can be given to students based on [pronunciation] errors that arise. (예를 들어서 초급 같은 경우에는 형태 좀 더 초점을 맞춰서 교육을 하고 있기 때문에 그 과정에서 여는 거 같은 발음은 연습할 뿐이고 교사가 읽은 후에 학생들이 따라 읽고 또는 통사적으로 문장을 교사가 읽으면 또 문장을 따라 읽고 거기서 생기는 오류들을 학생들에게 지도하는 편입니다.)

This conventional, choral repetition-based classroom pedagogy is in line with what many of the Korean language program students reported in interviews and supports Jae-woo's view that the importance of pronunciation is not generally treated adequately in Korean language education. He went on to ascribe this mismatch to curricular demands and lacks in pedagogical materials,

which puts teachers in a difficult situation when it comes to devoting more time to pronunciation. In sum, while Jae-woo appeared well-versed in Korean phonology (and at least some learner L1 phonologies) and believes pronunciation to be important, his teaching practice was constrained by the status quo.

Despite some of his initial skepticism and critical interpretation of the KPD results for Yuki and Yu-wen, Jae-woo was positive about the potential for both learners and teachers to apply them:

> Through diagnostic results like these learners can know what kind of difficulties they have and if instructors could incorporate these in class it seems like it could make for a really effective class. (이런 진단 결과를 통해서 학습자들이 어떤 발음 상의 문제점이 있는지를 알고 교사가 수업에 들어갈 수 있다면 훨씬 효과적인 수업이 될 수 있을 것 같아요)

This quote indicated to me that he sees some value in learner awareness as well as potential for teacher-driven application. Although he acknowledged that Yuki and Yu-wen differed in their pronunciation weaknesses and that students from the same L1 could have different profiles, he felt that "90%" of learners from the same L1 background would have the same pronunciation difficulties, barring any extensive time in a target-language environment or extensive self-study. He imagined separate pronunciation classes for students of different L1 backgrounds, an idea I found concordant with his previous description and attribution of learner pronunciation difficulties along lines of L1 interference. In these classes, he would use repetition activities, but also add self-listening, perhaps keying into the perception information in KPD score reports. He also thought it would be helpful to correct students' place of articulation ("조음 위치를 교정하는 거" *place of articulation correction*), which I took to mean providing explicit articulatory instruction (which is well-supported in the pronunciation instruction literature, Derwing & Munro, 2015; Derwing et al., 1998; Lee et al., 2014). While I was hoping for

comments more specific to Yuki and Yu-wen's difficulties, I did find it interesting that many of the teaching ideas brought up by Jae-woo were not part of what he is typically able to do in his classroom.

Learner Utilization and Impact

Fourteen of the 21 learners were available to complete a second interview, which was focused on their application of KPD results and pronunciation learning activity. During the second interview, I also had them retake the KPD. The second-interview data provided a better, more concrete understanding of how Korean learners might apply the information from their KPD score reports compared to their speculative comments from the first interview. The quantitative KPD retest data, though small in scale, shed light on the link between pronunciation learning activity led and measurable pronunciation development. To a limited extent, examining the KPD test-retest data alongside learning activity also allowed me to consider measurement stability and sensitivity. For the sake of coherence and conciseness, I focus my reporting of findings primarily on the production of Korean phonemes rather than on supplementary information (features, contexts), with some consideration of perception at a broad level.

In what follows, I first consider quantitative data describing the differences in phoneme perception and production from initial test to retest, followed by an analysis of the interview data to connect learner activity and perceptions with retest scores.

Changes in Production and Perception. Overall, the 14 learners made modest improvements to their production and perception of Korean phonemes over the 2 to 4 months between initial KPD and retest (Table 8.3). On average, learners became 1% more accurate in their average phoneme production and 2% more accurate in phoneme perception. It is worth pointing out that phoneme perception averages were lower to begin with, making the somewhat

larger gains unsurprising; there was greater variability in phoneme perception accuracy. In terms of diagnostic flags, learners were able to ameliorate less than one net phoneme flag on average.

Table 8.3

Group-Level Summary of Changes in KPD Production and Perception Scores

	Production		Perception	
	mean	SD	mean	SD
Initial Parcel Average	88%	5%	79%	9%
Retest Parcel Average	89%	5%	81%	9%
Change	1%	4%	2%	5%
Initial Flag Count	6.43	3.27	12.79	4.44
Retest Flag Count	6.14	2.44	12.21	5.34
Change in Flag Count	-0.29	1.94	-0.57	2.87

Note. Based on 14 learners who completed the KPD a second time. Diagnostic flags based on < 80% accuracy criterion.

Throughout this subsection, as well as the subsection on learner utilization and impact, readers will find it helpful to refer to Table 8.4. Table 8.4 summarizes the differences in KPD production scores from initial to retest, as well as the learners' descriptions of their pronunciation learning activities. Learners are listed in the table in descending order according to the magnitude of improvement to their average phoneme accuracy from initial testing to retesting. After each learner's name, the table contains information on average phoneme production accuracy. This is followed by information on diagnostic flags at initial KPD and retest. The last column of the diagnostic flag part of the table, labeled *Description*, uses a - sign to note which phoneme flags did not appear again on the retest results and a + sign to note which flags newly appeared on the retest.

Some learners made impressive accuracy gains and showed largely expected patterns in phoneme flag reduction. For example, Maria improved her production accuracy by 7% and removed one phoneme flag without adding any new flags. Similarly, Noriko's average

production accuracy improved by 5% and she was able to remove 6 phoneme flags (though added two new ones at retest). For Maria, a student with limited Korean experience to begin with (EIT score of 18/120 and was enrolled in Level 2 in an intensive Korean program), the magnitude of improvement in just over two months is not surprising (especially considering her specific learning activity, discussed later). Noriko, however, was in a Level 5 course at the time of her initial test and had a mid-range EIT score yet was able to make noticeable improvements. Ju-an's results are interesting—a gain of 4% accuracy and net loss of two diagnostic flags—because she had high production accuracy to start with (92%) and considerable Korean experience (EIT 91/120, enrolled in a Korean-language master's degree program, and one year of residence in South Korea).

Not all learners' KPD results showed signs of progress. In the middle of the pack in Table 8.4 lies Amber, a multilingual student from Hong Kong with high levels of Korean experience and a very high average phoneme production at her initial KPD: 94%. Amber's average accuracy showed virtually no change at retest, and she only shuffled two phoneme flags, with a net loss of zero flags. Several other learners showed small decreases in average production phoneme accuracy and unclear patterns in diagnostic flags. At the extreme, Leo, an English-medium program undergraduate with moderate Korean proficiency but extensive in-country experience (EIT 38/120, 5 years residence in South Korea), saw on his retest a decrease in average phoneme accuracy of 9% and the addition of four diagnostic flags.

Table 8.4

Individual Summaries of Changes in KPD Production Scores and Learning Activity

	Avg. Phoneme Acc.			Diagnostic Flags (n)				
Name	Initial	Retest	Change	Initial	Retest	Change	Description	Summary of Learning Activity
Maria	86%	93%	7%	6	5	-1	- /t/	Thought a lot about results, esp. tense consonants. Paid extra attention to her teacher's pronunciation of difficult sounds. Started to visualize written form of words to aid in remembering to articulate tense sounds. Began exaggerating tenseness. Asked her teacher for feedback on her pronunciation.
Noriko	79%	84%	5%	13	9	-4	- /ŋ, t*, n, p, ɯ, j/ + /s*, ʌ/	Met with a tutor once a week for month to work on pronunciation (did not show KPD scores to tutor). Did typical class activities such as read aloud and presentations. Watched Korean TV, studied for TOPIK listening.
Holger	80%	83%	4%	13	11	-2	- /t*, p*, s, j/ + /u, ɛ/	Thought about results frequently. Practiced reading sentences aloud, asked language exchange partner to correct mispronunciations when reading news articles aloud.
Ju-an	92%	95%	4%	5	3	-2	- /t, k*, s*, u/ + /t*, tɕ^{h}/	Memorized short list of weaknesses and tried to keep them in mind while interacting, trying to pronounce those sounds more clearly.
Jing	85%	87%	2%	6	7	1	- /t, t*, k, ʌ, o/ + /s*, b*, k^{h}, k*, l, u/	Did not think about results or practice much outside of incidental Korean use at work or with boyfriend. Reported paying more attention to syllable coda sounds.
Fang	88%	90%	2%	6	5	-1	- /s*, t*, ʌ/ + /t, u/	Paid more attention to difficult sounds in daily use. Took a Korean pronunciation class, but it had little practice opportunity. Worked on /l/ pronunciation by learning a popular song. Paid attention to expressions her Korean coworkers used with customers.

Table 8.4 (cont'd)

Xiu Ying	92%	95%	2%	5	4	-1	- /k^h/	Did not think much about results or practice on her own. Teacher in translation/interpretation course corrected imprecise pronunciation. Used Korean informally with friends. Paid more attention to some difficult targets in general use.
Amber	94%	94%	0%	5	5	0	- /p*, u/ + /k^h, t*/	Shared her results with Korean friends in phonology/pronunciation course. Little focused practice and did not think about results too often.
Hoa	93%	92%	-1%	5	5	0	- /p, t*, tɕ^h/ + /s*, p*, k*/	Used Google voice-to-text technology to practice, esp. /p, p^h/. Used a proverb and expressions books to find meaningful language to practice pronouncing. Practiced TOPIK listening; did some shadowing of passage extracts.
Na	91%	90%	-1%	5	7	2	- /t*/ + /tɕ^h, k, y/	Did some practice of difficult sounds; individual words and sentences with feedback from Korean friend. Tried to learn more phonological processes. Little feedback from teacher in regular Korean class but did get some advice about syllable codas.
Min	98%	96%	-2%	1	2	1	- /l/ + /j, tɕ*/	Tried to pay more attention to her /l/ pronunciation in daily life. Took a Korean phonology/pronunciation class during fall semester, which included practice opportunities and self-recording homework.
Sofia	88%	86%	-3%	7	7	0	- /t^h, j/ + /p^h, ʌ/	Did not think much about results. Could not take a Korean class in current semester. No specific pronunciation practice. Used Korean in daily life and watched Korean dramas.
Sakura	83%	81%	-3%	9	8	-1	- /tɕ*, ɯ, w, y/ + /t, n, ʌ/	Did not practice or study much. Asked a Korean friend for confirmation of her difficulties with a few sounds. In-class pronunciation feedback focused mostly on phonological processes. Bought a pronunciation textbook but did not use it.
Leo	90%	81%	-9%	4	8	4	+ /t^h, p^h, p*, k/	Did little specific pronunciation practice. Spoke Korean in social settings, watched Korean YouTube.

Application of KPD Results. Follow-up interviews with learners illuminated the quantitative test-retest data and showed a range of ways that learners applied what they learned from their KPD results. Learners with some of the largest improvements to their KPD results reported engaging in sustained and focused pronunciation learning activity. In my view, Noriko's learning activity demonstrated the greatest investments: Noriko hired a tutor specifically to work on her pronunciation. She could only afford this for one month, and found the experience of getting intensive pronunciation feedback a little "scary" ("무서워요"), but ultimately found it helpful. Noriko described some of what she did with the tutor as reading aloud and getting evaluation and corrections from the tutor; she further remarked that "I couldn't distinguish things like this [on my own]" (이런 거는 제가 구별이 할 수 없었어요) (this read-aloud with feedback activity is similar to the tandem exercise described in Horgues & Scheuer, 2014). Beyond this specific pronunciation learning activity, Noriko reported engaging in general speaking and listening practice in class, and extra listening practice outside of class for pleasure (watching Korean dramas) and test preparation (TOPIK listening section).

Maria, who showed the greatest overall improvement from initial test to retest, also revealed a considerable degree of engagement with her KPD results and commitment to pronunciation learning activity. Similar to Noriko's seeking of external help, Maria reported showing her KPD report to her teacher and asking for additional feedback on her pronunciation, which she received periodically after class: "So after class I would get the feedback, and she would say 'No your pronunciation is not good, you are still doing this wrong.' She mentioned about the 드[/tɯ/], 트[/t^hɯ/]." Interestingly aligned with this comment, /t/ was the one phoneme flag that Maria was able to clear on her KPD retest. On her own, she reported paying more attention to how her Korean teacher produced difficult sounds, and she also made efforts to

pronounce tense consonants more exaggeratedly. Holger and Hoa were two other learners whose learning activity stood out. Holger, who made some noticeable improvements on his KPD retest, reported regular sentence read-aloud practice and working on pronunciation with a language exchange partner. Hoa, who had rather high production accuracy to begin with yet did not make many overall gains, took a more tech-infused approach: She used Google's automated speech recognition (ASR) service to work on her pronunciation (see McCrocklin, 2019, for a classroom-based application of ASR for pronunciation instruction), with a special focus on her /p-p^{h}/ contrast (encouragingly, she did manage to lose her /p/ diagnostic flag at retest). Finally, Fang reported paying attention to difficult sounds in daily use, and more specifically to work on /l, she practiced the popular Korean children's song "Baby Shark" ("상어 가족", Pinkfong, 2016), which features a nonlinguistic refrain of /t*u.lu.lu.t*u.lu/ ("뚜루루뚜루").

While some learners did not engage much with specific, focused pronunciation learning activities, they did describe how the KPD results led to awareness-raising and low-level, continuous application of results in daily language use. Ju-an, who I previously noted had impressive gains with respect to her initial high production accuracy, reported memorizing her major difficulties and then reflecting on them whenever an interlocutor had difficulty understanding something she said. In addition, she generally tried to be more conscious of her articulation of difficult sounds. Xiu Ying and Jing, who both posted modest improvements on their retest results, did not engage in much focused pronunciation learning activity but did report paying more attention to difficult sounds in their daily language use.

Last, some learners neither engaged in much focused pronunciation learning activities nor tried to maintain awareness of difficult sounds in daily use, though many reported engaging in general listening and speaking practice. Unsurprisingly, many of these learners showed little or

no evidence of improvement in their KPD retest score: Leo, Sakura, Sofia, and Amber all followed this pattern. Despite their lack of pronunciation learning activity, they did show at least some initial engagement with results. Leo, Sakura, and Amber all reported talking with Korean friends about the results. Amber even went so far as to say words/syllables and ask her classmates what consonants they heard her say, specifically focused on the tense consonants that were diagnostically flagged on her score report. Amber found agreement between her classmates' uncertainty of her tense consonant production and her KPD results. Ultimately, other demands prevented further engagement with results. For example, Sakura reported actually buying a Korean pronunciation book but was unable to free up enough time to study it, and Sofia and Leo were both kept busy by their undergraduate coursework and part-time jobs.

Perceptions of Change. As might be expected given the varying levels of learning activity and varying levels of initial pronunciation accuracy, learners varied in the perceptions of change from initial test to retest. Six learners stated that they noticed some kind of improvement to their pronunciation (Min, Ju-An, Fang, Maria, Xiu Ying, and Jing). Three of these six could describe their improvements in detail, though they said it was not easy to judge for themselves: Maria (improvements to /p*, tɕ, tɕ*/), Ju-an (improvements to /t, t*/ and /ʌ, u, o/), and Min (consonant relinking, a phonological process that is not directly assessed by the KPD). While Maria did not clear her diagnostic flags for /p*, tɕ*/, she did make substantial improvements, going from 0% to 75% accuracy for /p*/. Ju-an did clear her diagnostic flags for /t/ (going from 78% to 89% accuracy) and /u/ (going from 75% to 100% accuracy), though she did regress in accuracy on /t*/ (dropping from 100% to 75%).

Two other learners, Amber and Noriko, spoke of a lack of development in relatively certain terms. Amber felt confident that she did not make any substantial improvements,

especially in relation to the tense consonants that still eluded her. In her case, while she did show improvement in one tense consonant (/p*/), a different one (/t*/) was newly flagged on retest, and her overall production phoneme average showed virtually no difference at retest. Noriko, despite making considerable gains in her KPD scores, did not perceive much improvement and felt that her difficulties persisted. To some extent, she was not wrong: Even at retest, she had a total of 9 diagnostic flags on production phonemes and still had room for general improvement (84% average production phoneme accuracy). With some limits, it appeared to me, perception of improvement (or lack thereof) was possible and reasonably accurate for some learners, in some cases even at the phoneme or feature level.

The remaining learners expressed uncertainty when it came to noticing changes in their pronunciation. Some learners reported positive or negative impressions of their progress but qualified them immediately before or after by saying that they were not sure (Hoa, Sofia, Na) or could not tell (Leo, Holger, Sakura). Despite Hoa's uncertainty and overall limited development, she did nonetheless appear to improve in one phoneme that she had practiced, /p/. Holger, who could not tell on his own whether he had progressed, posted relatively strong improvements on his KPD retest. Where these less-certain learners had difficulty judging their own gains, they sometimes turned to the assessments of others: Sofia reported customers at the restaurant she worked at understood her better compared to a few months prior while Hoa described comments from Korean friends about reduced Vietnamese-like intonation in her speech and that Google's ASR still indicated she had some difficulty with /p/ and /p^h/.

Discussion

In this chapter, I drew on interview data to report on the interpretation and utilization of KPD results by key stakeholders: Korean learners and a Korean teacher. Furthermore, I brought

in KPD retest data to shed light on the connection between utilization of KPD results and subsequent learning, a key consideration in diagnostic language assessment and in turn an important piece of evidence for the utilization inference in the KPD's validity argument. In this discussion section, I reflect on the findings in respect to my primary research questions and then offer additional considerations arising from analysis of the data.

RQ7: How do (a) Teachers and (b) learners understand KPD score reports? To what extent do they learn anything new from KPD score reports?

The teacher I interviewed, Jae-woo, came to understand KPD score reports as a source of information on segmental pronunciation issues that filled in gaps in his own observations. At a more basic level, he had no trouble understanding the content of the score report, though he needed more explanation of the test structure and scoring procedures in order to develop a better sense of how to interpret the information contained in the score report. This points to a need for documentation to be made available to test users. While I have created documentation of the KPD design and task/item specifications, I did not provide these to Jae-woo, nor have I developed more succinct, stakeholder-friendly documentation that would undoubtedly aid in appropriate score interpretation.

Learners, who had all taken the KPD themselves and at least had a first-hand understanding of the KPD design, tended to view the KPD as an external, more objective assessment of their pronunciation weaknesses and strengths. As Chapter 7 illustrated, fine-grained self-assessment of segmental production and perception abilities was not easy for learners to do accurately, and learner uncertainty about their own strengths and weaknesses was a topic that came up during interviews as well. In addition to filling in gaps in their knowledge, the KPD results also helped learners confirm or reject what they had (uncertainly) thought about

their own abilities. On the other hand, several learners had difficulty reconciling their lower perception scores, which were presented on the same scale as the production scores. Although it may be the case that some learners do have substantial perception difficulties, much of the disparity could be attributed to differences in scoring standards across the two modalities, as discussed in previous chapters. Thus, additional explanation or re-scaling of perception scores may help learners more appropriately interpret their scores.

Although learners appeared to immediately understand the information on the first page of the score report as feedback on weaknesses and intuitively grasped the meaning of the percentages for each phoneme on the second page, learners' understanding of their score reports was not absent of stumbling blocks. Learners of all levels of overall Korean proficiency were unfamiliar with some linguistic terminology (e.g., 경음 *tense*), and learners with lower levels of proficiency showed some difficulties comprehending the orienting prose at the top and bottom of the first page. Furthermore, learners frequently asked whether the example words given for phonemes on the second page of the report were from the production or perception section of the KPD (interestingly, I had separated the example words by modality in an earlier version of the score report). Thus, improvements to the score report addressing these stumbling blocks could improve learner interpretation and perhaps utilization of KPD results.

Key to the utilization of diagnostic instruments is that the diagnostic feedback provides information which the users could not have easily obtained otherwise; after all, it would make little sense to go through the process of administering a diagnostic test if teacher and learner observations could provide the same benefits. The quantitative comparison of learner self-assessments and KPD results in Chapter 7 suggested that learners had limited awareness of their strengths and weaknesses in production and perception of segmentals, and interview data

explored in this chapter provided further support for learners becoming aware of new information through KPD score reports. This was not just students becoming aware of weaknesses, but also becoming aware of (and/or more confident in) their strengths. The teacher perspective also supported the idea that KPD results could provide additional information for test users. Jae-woo commented that his observations of students, whom he had taught for a full semester, amounted to perhaps half of the picture.

RQ8: Do learners report any changes in their self-study routines and/or their attention to phonological form in formal or informal learning situations?

Some learners reported concrete changes in their learning activity in response to the KPD results, but not all. Among those who engaged in focused pronunciation learning activities, several learners discussed activities that are well-supported in research or long-standing pedagogical practice: shadowing (Foote & McDonough, 2017), read aloud with ASR feedback (McCrocklin, 2019) or partner feedback (Horgues & Scheuer, 2014), practice through songs (Graham, 2001; Richards, 1969), and seeking feedback during (or after) meaning-focused interaction (Saito & Lyster, 2012). In the initial interview, some learners reported a lack of knowledge related to studying or practicing pronunciation effectively. This suggests that the utilization of the KPD could be enhanced by providing learner-friendly information on pronunciation learning activities and/or the delivery of results by a teacher who can provide more specific guidance in this area.

Learners also reported that the KPD results guided them to effortfully raise their awareness of difficult phonemes, or to pay more attention to how they pronounce difficult phonemes in typical language-use situations, or both. This combination of awareness and deliberate attention could lead to increased levels of incidental focus-on-form which learners

may otherwise miss out on during typical meaning-focused Korean use (Kennedy & Trofimovich, 2010; Saito, 2018; Schmidt, 1990, 1993).

A minority of learners who completed the second interview and KPD retest reported doing little if anything with the KPD results. While they had perfectly understandable reasons for not committing to additional pronunciation learning activities (e.g., limited time), the more important takeaway is that diagnostic assessment alone cannot be considered an instructional intervention; learning activity naturally depends on learner and/or teacher efforts. As Alderson et al. (2014) emphasized, test users are at the heart of diagnostic assessment, and the use of a diagnostic instrument is just one phase of a larger process. Nonetheless, the KPD was able to be fruitfully applied by learners, which is promising on its own, especially for learners not currently engaged in formal Korean instruction, and bodes well should the test be used by a knowledgeable teacher/diagnostician (such as Jae-woo) within a classroom context.

RQ9: Do learners show improvements in a) overall and/or b) in weak areas after receiving and applying KPD feedback?

With many qualifications, I believe the answer to this research question is "yes." The learners who appeared to take focused, sustained measures to address their Korean pronunciation after receiving their initial KPD results made clear gains. Maria, Noriko, and Holger all took substantive action, guided by their KPD results, to improve their pronunciation, including self-study, paying closer attention to difficult sounds in their input and output, and seeking help from others (teacher, tutor, language exchange partners). It is worth noting that those learners who made the most impressive gains were those with some of the most initial production difficulties and only moderate amounts of Korean language experience, though not exclusively: Ju-an made impressive gains despite initially high levels of production accuracy and Korean exposure.

Although to directly relate this KPD-motivated and guided activity to their visible improvements at retest is difficult without a control group, I believe it is reasonable to conclude that beneficial outcomes of post-diagnostic learning activity are certainly possible.

The learners who did relatively little with their results and subsequently showed little improvement, either in global production accuracy or accuracy of specific problem phonemes, which provides some counter-factual support for this conclusion. Namely, when KPD results are not meaningfully applied, learners are not likely to experience improvements to their segmental pronunciation abilities over the course of 2 to 4 months (in absence of other directed pronunciation learning activity). This counter-factual situation and outcome is intuitive (i.e., what improvements would be expected when no effort is made?) and also makes sense on a more theoretical level given that learners' pronunciation development often plateaus, showing little to no change over extended periods of time, after a phase of rapid L2 phonological development that starts with initial exposure to the language (i.e., the Window of Maximal Opportunity, Derwing & Munro, 2015). Of course, the data on learning gains presented in this chapter is extremely small in scale, and larger-scale quantitative investigations would provide stronger support for the beneficial consequences of using the KPD to guide pronunciation learning activity. Further, the findings related to post-diagnostic learning activity and pronunciation development again underscores the conceptualization of diagnosis as a process that must feed into instruction (Alderson et al., 2014; Lee, 2015).

Additional Considerations

The test-retest results in this chapter provide additional glimpses into potential support for three other inferences in the KPD's validity argument: Generalization, Explanation, and Extrapolation. While the data presented in this chapter have a variety of limitations, including

depth of interview questioning, learner (and interviewer, in some cases) language proficiency for interviews, and small number of test-retest participants, I find the implications for these inferences too interesting to not consider.

Generalization inferences in validity arguments broadly pertain to the consistency of results across observations of test takers with the same, presumably static (or temporarily stable) level of ability in attributes of interest. Analyses related to the generalization inference should attempt to account for variation in scores across different forms of a test (e.g., test equating), across different human raters (e.g., inter-rater reliability or agreement), and across different points in time (e.g., test-retest reliability). However, most commonly, generalization is investigated via estimation of internal consistency (e.g., Cronbach's alpha), which theoretically aligns with the average of all possible split-half reliability estimates (Crocker & Algina, 1986). This is actually among the weakest forms of evidence in support of the generalizability of test scores, as it only offers conclusions based on one administration at one point in time. With the KPD test-retest data, it was possible to (at least) consider the test-retest consistency of some individuals who took the test at two different points in time without engaging in behaviors that would produce a substantive change in their ability. Amber, Sofia, Sakura, and Leo appeared to do the least amount of pronunciation study, and thus can be assumed to have reasonably similar ability levels at initial test and retest, though Sakura's lower proficiency and limited exposure could have led to more development than the others. Three of these learners showed very little difference in KPD scores across two observations in time, providing some support for generalization of scores. Leo, however, had noticeably lower production scores on his second KPD, a direction of change that would not follow most predictions for L2 phonological development for a learner with five years of residence in the target language environment. While

intra-scorer variability may partially explain the discrepancy in Leo's scores, intra-speaker variation may play as large of a role, or perhaps an even greater one. Recent work by Smith, Johnson, and Hayes-Harb (2019) on L2 intra-speaker variability in vowel production, one of very few such papers on L2 speaker variability, found that while L2 speakers in their study did not exhibit larger variation in vowel production than L1 speakers; the L2 speakers' variations were outside of L1 norms approximately 50% of the time. In the context of the KPD, although NS-like productions are not required, substantial deviations could nonetheless lead to unintelligible production of target phonemes and have a negative impact on scores. The phenomenon of intra-speaker variability could also (at least partially) explain subscore variability for someone like Amber, who despite maintaining virtually the same production phoneme average across two points in time varied slightly in individual phoneme accuracy scores.

Explanation inferences draw on a wide range of support, from documenting test-taker response processes to investigating relationships with external measures informed by theory and substantive empirical research (Chapelle et al., 2010; Kane, 2013). At the core of classical perspectives on validity, the key consideration is whether variations in scores produce (or reflect) variation in the ability measured (Borsboom et al., 2004). One of the most rigorous ways of investigating this relationship is by testing individuals at multiple points in time, before and after interventions or experience that theoretically should produce a change in the individual's underlying ability. Once again, the KPD test-retest data provides an interesting perspective on this aspect of score meaning. As I have already discussed at some length, learners who engaged in substantial learning activity over a period of 2 to 4 months, especially those with lower initial phoneme production abilities, showed changes in their KPD scores in the expected direction,

while those who did less or started with higher abilities (or both) showed comparatively smaller change or little change at all. While the sample size and specificity of learning activity and intervening exposure to the language is insufficient for rigorously testing this relationship and evaluating the magnitude of change, it was nonetheless promising to observe a chain of test score → learning activity → ability development → higher score.

Finally, interviews with learners and a teacher provided additional evidence pertaining to the extrapolation of KPD scores to other domains, in this case, learners' daily life, social interactions, and classroom language use. To some extent Jae-woo's comments about Yuki and Yu-wen's pronunciation offered some support for a connection between KPD scores and classroom language use, though this was limited, in part due to Jae-woo's lack of specificity in his observations of phoneme-level difficulties experienced by the two students. Like the self-assessments analyzed in Chapter 7, learner interview comments related to their (dis)agreement with KPD scores provided support for the notion that KPD scores reflect non-test performance reasonably well. Learner anecdotes of pronunciation and/or hearing difficulties were especially illuminating and persuasive due to their specificity. For example, Sofia could specifically and vividly recall how customers at her part-time job misunderstood her rendering of 꿀 (*honey*, /k*ul/) likely due to the tense stop /k*/ in the onset of the word. Leo shared an amusing anecdote about his difficulty pronouncing two of his Korean friends' minimal-pair names (differentiated only by an initial /ʨ-ʨʰ/ aspiration contrast). Some participants endeavored to find their own evidence to support the extrapolation of test results, such as when Amber tasked her Korean classmates to identify whether she was making a tense or non-tense sound, and when Hoa checked her /p, pʰ/ pronunciation against Google's ASR. The specificity and vividness of this qualitative data provides substantial support for the KPD's extrapolation inference.

CHAPTER 9: SUMMARY OF FINDINGS AND EVALUATION OF THE VALIDITY ARGUMENT

In Chapter 2, I outlined a proposed validity argument for the interpretation and use of KPD scores. In this argument, I sketched out what kind of information could be used to support each inference in the argument which allows test users to go from test observations to real-world, beneficial use of the KPD. For some of the earlier inferences in the KPD's validity argument, my already completed design and initial piloting efforts provided support (Chapter 3). However, for most inferences, I identified gaps in necessary support, which led to the formation of research questions and the collection of data to answer them. I have reported on these findings over the last several chapters (Chapters 5 through 8). In this chapter, I return to the validity argument and summarize the evidence gathered and interpret it in respect to specific inferences necessary to support KPD score interpretation and use. Following this synthesis, I critically evaluate the strength of the argument and consider gaps and weaknesses to be addressed in the future.

Summarizing the KPD Validity Argument

In this section, I return to the proposed validity argument and synthesize all extant support for each inference. In what follows, I provide a formal, detailed rendition of the validity argument, following Chapelle et al. (2008, 2010): I articulate the warrants that explicate each inference, and for each warrant I articulate key assumptions, and for each assumption I review the extant support for each assumption.

Operationalization Inference

Warrant: Observations of learners' Korean segmental production and perception reveal underlying strengths and weaknesses in phonological knowledge and processing that are important to communication and the development of intelligible pronunciation.

Assumptions and Support:

(1) Items represent the inventory of Korean phonemes.

(2) Test tasks are designed with reference to theories of L2 phonological learning.

(3) Test tasks are sufficiently delimited to lower-level subprocesses of speech production and perception.

With regards to the first assumption, the KPD suitably delimits the target domain to Korean segmental phonology, with tasks that exclude most suprasegmental aspects of pronunciation as well as opaque phonological processes found in spontaneous connected speech. Related to the second assumption, the KPD features both production and perception tasks, as informed by theories of speech learning (Flege, 1995) and empirical findings on the link between perception and production (Sakai & Moorman, 2018). Addressing the third assumption, KPD task designs limit the non-phonological resources necessary for learners to respond, requiring only knowledge of basic sound-script correspondences and commonly-taught, high-frequency vocabulary. In this way, language production and perception on the KPD is relegated to lower-level subprocesses (Field, 2011, 2013), minimizing construct-irrelevant variance from higher-level subprocesses unrelated to segmental pronunciation knowledge and ability. Reviews of literature on these areas are found in Chapter 2, and Chapter 3 and Appendices A and B contain detailed information on the development and specification of the test and its component tasks.

Evaluation Inference

Warrant: Observations of phoneme production and perception on the KPD are evaluated to yield scores that are (a) instructionally-relevant, (b) indicative of strengths and weaknesses, and (c) in line with the ultimate goal of intelligible oral communication.

Assumptions and Support:

(1) Task responses are scored based on appropriate criteria.

(2) Measurement characteristics of the KPD differentiate learners by overall ability in perception and production, while phoneme parcel subscores provide appropriate diagnostic information.

Pertaining to the first assumption, evaluation of KPD responses is appropriate, well-defined, and verified. Evaluation of production task responses is based on a clear criterion and heuristics that draw on research and best practices in L2 pronunciation pedagogy (i.e., Levis' Intelligibility Principle, 2005; see Chapter 2). Evaluation of the production tasks are aided by an easy-to-use scoring sheet, and training materials were created to orient new scorers to the scoring criteria. The perception task responses are evaluated based on accurate keys which were verified by a NS linguistic informant. Furthermore, two rounds of piloting provided additional verification of item keys (Chapter 3), as did the almost universally maximal scores of NS test-takers (Chapter 5)

Regarding the second assumption, measurement analyses found that a measurement model based on phoneme parcels had statistical characteristics that were highly similar to models based on individual items (Chapter 5). At the same time, the phoneme parcels align with the intended use of the KPD, i.e., to provide information on phoneme-level strengths and weaknesses to guide instruction. In both CTT and Rasch analyses, some phoneme parcels had very low difficulty and/or discrimination. However, from a diagnostic perspective, this is fine: The major concern is the capability to detect low performance on phonemes, even if they tend to be easy for most learners. Moreover, Rasch analyses of phoneme parcels showed that parcel information was greatest at lower score levels, supporting this aim.

Generalization Inference

Warrant: Observed KPD scores estimate learners' abilities with stability and are similar across scorers.

Assumptions and Support:

(1) Items are sufficient in number and quality to yield stable estimates of overall production and perception abilities and individual phoneme ability.

(2) KPD production section scores are stable across scorers.

Relevant to the first assumption, KPD overall production and perception scores based on phoneme parcels are internally consistent and have adequate precision (Chapter 5). Internal consistency estimates for both individual item and item parcel scoring models were suitable for low-stakes assessment and provide positive evidence for the (lower-bound) of the KPD's precision of measurement (Crocker & Algina, 1986). Additionally, there is limited support for the generalization of KPD results across test occasions (Chapter 8). In Chapter 8, several learners took an initial KPD and then took the test again approximately three to four months later, without having engaged in deliberate pronunciation learning activities. These learners saw little to no change in their overall KPD scores, as would be expected for individuals whose underlying abilities had not changed.

Inter-scorer agreement was also found to be high. At the individual item level and at the parcel-based diagnostic flag level, different scorers on average had nearly perfect levels of agreement. Phoneme parcel scores across raters varied widely from phoneme to phoneme, and many low estimates of interrater reliability were obtained. However, this was found to be due to a preponderance of very high scores with near-universal agreement for some phonemes. In sum, different scorers introduce little variability to KPD scores.

Explanation Inference

Warrant: KPD scores are reflective of learners' underlying phoneme knowledge and processing ability.

Assumptions and Support:

(1) Response processes align with theoretical expectations.

(2) KPD test tasks relate to one another in accordance with theory.

(3) Hierarchies of phoneme parcel difficulty in production and perception align with expectations.

(4) KPD Scores reflect distinct learner profiles.

(5) Phoneme production and perception scores relate to overall oral proficiency to an expected degree.

(6) Changes in phoneme production and perception scores reflect changes in ability due to learning and experience.

Regarding the first assumption, observations and test-taker interviews during piloting indicated that response processes aligned with expectations (Chapter 5). Regarding the second assumption, the internal structure of the KPD generally aligned with expectations (Chapter 5). Overall production and perception section scores correlated, and the pattern of correlations among overall task scores generally aligned with expectations.

Pertaining to the third assumption, the hierarchy of item difficulties aligned with expectations and empirical findings (Chapter 5). One consonant that was easier than expected based on research findings was /l/. However, the ease of /l/ could be reasonably attributed to the KPD's scoring criteria and lack of similar sounds which might increase the likelihood of learner articulations being ambiguous.

For the fourth assumption, KPD scores increased moderately alongside overall oral language proficiency, as expected (Chapter 7). Regarding the fifth assumption, KPD parcel scores indicated several identifiable general profiles for difficulties in phoneme production and perception that were not simply determined by L1 influence or Korean proficiency (Chapter 6). Lastly, related to the sixth assumption, small-scale exploratory analyses of test-retest data suggest that KPD results appear to reflect changes in underlying pronunciation abilities of learners (Chapter 8).

Extrapolation Inference

Warrant: The knowledge and abilities measured by the KPD are relevant to learner performance in general Korean oral communication.

Assumptions and Support:

(1) Strengths and weaknesses in phoneme production and perception are related to pronunciation in general Korean language use.

As the KPD by design isolates aspects of L2 phonology and of language processing, an idealized one-to-one correspondence between KPD results and meaningful, spontaneous Korean language use could not (and should not) be expected. This delimitation in mind, the alignment between KPD results, learner self-assessments of production and perception abilities, and learner errors in spontaneous, meaning-focused speaking provided mostly positive support for this assumption (Chapter 8). Additionally, alignment between a teacher's observations of two students and several learner anecdotes of production and perception difficulties provides additional support for the extrapolation of KPD results to strengths and weaknesses in general Korean use (Chapter 7).

Utilization Inference

Warrant: KPD phoneme scores and diagnostic flags are interpretable and useful to learners and teachers for planning pronunciation learning activity and raising awareness of difficulties.

Assumptions and Support:

(1) KPD feedback is interpretable by learners and teachers.

(2) KPD feedback can support instructional decisions.

Regarding the first assumption, key stakeholders, learners and a teacher, were able to appropriately and beneficially utilize KPD results (Chapter 8). Learners and a teacher were able to easily understand key information on KPD score reports related to phoneme strengths and weaknesses in each modality. However, learners struggled to interpret some supplemental information contained in the score report.

Regarding the second assumption, several learners were found to engage in substantial pronunciation learning activity, ranging from exercises such as shadowing to deliberate awareness raising and attention to target phonemes in daily language use (Chapter 8). However, some learners had few ideas on how to apply the KPD results, and others engaged in little to no self-directed pronunciation learning activity after obtaining KPD results. Nonetheless, many learners' self-assessments were found to contain misconceptions of their segmental strengths and weaknesses, highlighting the potential for KPD results to correct learner understandings and more usefully focus learners' awareness and learning efforts (Chapter 7).

Test Usefulness & Impact Inference

Warrant: Appropriate application of KPD scores by learners and teachers leads to beneficial outcomes through the development of more intelligible segmental pronunciation and accurate perception.

Assumptions and Support:

(1) Application of KPD results contributes to pronunciation development.

Learners have the potential to fruitfully apply KPD results on their own through engagement in a variety of pronunciation learning activities (Chapter 8). For learners who sustain pronunciation learning activity to a sufficient degree, KPD retest results suggested meaningful improvement to pronunciation abilities. This suggests that utilizing the KPD can have beneficial consequences on pronunciation learning, fulfilling the primary purpose of a diagnostic assessment.

Evaluation of the KPD Validity Argument

Before proceeding to my evaluation of the KPD's validity argument, I must concede that I have become rather personally invested in the development and use of the KPD, and that a neutral party with an etic perspective would be the ideal evaluator. Nonetheless, dissertations require solo authorship, and so I have endeavored to be as objective (and self-critical) as possible. I hope that I do not fall too short of that goal.

Overall, the KPD's validity argument is well-supported, but that support is thinner toward the end of the chain of its constituent inferences. Support for the operationalization inference draws on well-researched findings from L2 speech learning and psycholinguistics in tandem with Harding et al.'s (2015) cutting-edge ideas on the design of diagnostic language assessment instruments. While Field (2014) pointed out that phonemes are not static, easily delimited entities in the minds of language uses and instead draw on a network of numerous variations due to contexts and speakers, phonemes as an abstraction of phonological knowledge serve as a useful heuristic that is interpretable by stakeholders. Furthermore, the KPD features several instances of each Korean phoneme, all in different phonological contexts, which at least partially

addresses this potential weakness in operationalization. In sum, the inference that the KPD suitably operationalizes the target construct in alignment with desired measurement outcomes and uses is well supported.

Support for the evaluation inference can also be regarded as strong. The KPD production scoring criteria draw on Levis' (2005) Intelligibility Principle, which prioritizes effective communication and recognizes typical limits on (adult) L2 acquisition. The KPD's scoring guide, training materials, and scoring sheet facilitate consistent scoring. The degree to which KPD production scoring actually reflects intelligibility in naturalistic language use may be questionable, but the surprising findings for /l/ parcel difficulty suggest that the scoring of KPD responses genuinely did not require native-like articulation, at the very least. For perception tasks, multiple rounds of piloting, consultation with linguistically-informed NSs, and NS KPD score data all contributed to the verification of answer keys, leaving very little room to question support for this inference. In sum, the inference that the evaluation of test-taker responses is appropriate is strongly supported.

The generalization inference is well-supported. Conventional CTT and IRT estimates of reliability for the desired phoneme-parcel measurement model are adequate, especially when considering the KPD's relatively low assessment stakes. This lends a considerable amount of support to the generalization inference. The test-retest reliability data is promising, but ultimately too small to provide substantial support. More research on test-retest reliability would be desirable. The evidence pertaining to inter-scorer agreement also adds strong support for this inference and is especially valuable considering that the KPD was designed to be scored locally by individual Korean teachers or tutors.

A wide range of evidence exists to support the explanation inference, but some of these sources are not without limitations. The use of cluster analysis to examine learner profiles was a useful way to examine broad differences in test-taker profiles, but ultimately it is unclear whether those clusters are stable and generalizable. It is hard to connect these clusters to any theory underlying L2 phoneme production and perception, but it does seem safe to interpret the clusters with minimal difficulties as having reached (or nearly reached) a desirable, high-level of phoneme control with limited need for additional instruction. The data pertaining to expected changes in KPD scores before and after a period of substantial pronunciation learning was exploratory and small in scale; clearly, more rigorous investigation of intra-individual changes in ability would be useful for supporting this inference. All in all, however, the evidence accumulated so far strongly suggests that KPD scores reflect learners' productive and perceptive abilities related to segmental phonemes.

Support for the extrapolation inference is perhaps the most challenging to interpret. Learner self-assessments provided some evidence of the extrapolation of KPD results to Korean use more generally, but this relationship had to be viewed as attenuated by limits in learner self-assessment accuracy. Similarly, alignment between KPD production phoneme results and phonological errors in spontaneous, meaningful speech was necessarily attenuated by the numerous influences on real-time speech that were intentionally excluded in the KPD's design. Felicitously, teacher observations of two learners and learner anecdotes from interviews augmented the support for this inference. Ultimately, however, extrapolation is one of the weaker links in the KPD validity argument.

Evidence collected to support the utilization inference was crucial with respect to the instructionally-relevant, diagnostic purpose of the KPD. Learner and teacher interpretation of

phoneme-level strengths and weaknesses was strongly supported, though I identified some necessary changes to score reports. While some learners were able to come up with and ultimately arrange for or execute quality pronunciation instruction/learning activity, other learners were less capable of applying their results to learning. Thus, while support for the utilization inference is promising and indicative of potential, it could be strengthened by clearer ties between KPD results and learning activity, whether that be in the form of linked resources for self-directed learning or in the form of a teacher/tutor who can provide structured pronunciation instruction.

The usefulness and impact inference may be the weakest link in the KPD's validity argument. Promisingly, a small number of learners were shown to have made non-trivial improvements to their pronunciation after receiving their initial KPD feedback. A handful of other learners made smaller gains, but these gains are arguably less certain due to precision limits of the KPD. The weakness in support for this inference comes not so much from the quality of the evidence, but the quantity: The accumulated evidence is based on a small number of learners. Going forward, it would be helpful to collect larger scale test-retest data. It would also be helpful to do so in a context where there was more structure or guidance leading students to suitable classroom-based instruction or other learning activities, similar to my recommendations for bolstering evidence for the utilization inference.

In sum, I judge there to be a substantial and reliable connection between learner segmental pronunciation abilities and KPD scores, which leads to appropriate interpretations of scores relevant to making decisions about learners' strengths and weaknesses. The KPD also has the potential to be fruitfully utilized, and the potential to make a positive impact on pronunciation learning. The weaknesses in the validity argument point to the necessity of

additional validation research, which I have already alluded to. While this dissertation presents a broad collection of evidence to support the KPD score interpretation and use, the evidence is not all-encompassing. In my evaluation, further investigating support for the utilization and impact inferences is most critical for future validation research. Specifically, examining the use of the KPD in a classroom setting, preferably in several classrooms taught by several different teachers, would be a valuable source of evidence that could further illuminate the degree to which KPD results can be beneficially applied.

Conclusion

As discussed above, despite some shortcomings, the evidence I collected in this dissertation largely supports the interpretation and use of KPD scores for the diagnosis of L2 Korean segmental pronunciation. Concurrently, the four aims I outlined in Chapter 2 have largely been achieved: The test has been developed (Aim 1), field tested to facilitate interpretation of results (Aim 2), examined in relationship to spontaneous speech and oral proficiency (Aim 3), and studied in terms of how teachers and learners understand test results (Aim 4). In the next chapter, I situate the KPD project in the larger L2 pronunciation and DLA literatures.

CHAPTER 10: DISCUSSION & CONCLUSION

In Chapters 5 through 8, I presented the results of validation research and discussed findings in respect to specific research questions. In these discussions, I situated my findings within the L2 pronunciation literature and the broader literature of DLA. In Chapter 9, I summarized and interpreted these findings in respect to the KPD's validity argument, which culminated in an evaluation of the validity of KPD score use and interpretation.

In this chapter, I offer broader considerations for diagnosing second language pronunciation and for creating and using diagnostic language assessments. I do this by first situating my research findings within the broader literature of second language pronunciation theory, research, and pedagogy, and the broader literature of DLA theory, research, and use. I then conclude the dissertation with some parting thoughts on the role of diagnostic assessment and language assessment professionals in the landscape of language learning practice and research.

Discussion on Diagnosing Second Language Pronunciation

Moving beyond the scope of the KPD, I now discuss the broader goal of diagnosing second language pronunciation by considering prior research and how the results of this research project fit within it. I start by situating my research results within the field of second language pronunciation, and then by discussing what I see as key questions for diagnosing L2 pronunciation and sharing my tentative answers based on my dissertation work. Then, I explore ways to expand and further develop the tools and practice of L2 pronunciation diagnosis. Next, I connect important areas of research that need to be combined to develop an interface between pronunciation instruction and diagnosis. Last, I present several implications for DLA theory and practice based on my findings related to the KPD.

Situating the KPD in L2 Pronunciation and DLA

The results of this dissertation give support to the idea that teachers and learners can benefit from detailed, individualized information when it comes to making informed and confident instructional decisions about teaching and learning pronunciation. To my knowledge, the KPD stands as the only stand-alone pronunciation assessment tool that (a) diagnoses learner phoneme-level strengths and weaknesses in pronunciation (cf. holistic approaches of Isaacs et al., 2018 and others), (b) integrates both production and perception (Flege, 1995; Sakai & Moorman, 2018), (c) explicitly promotes intelligibility-based evaluation of pronunciation (Levis, 2005), (d) does not rely exclusively on read-aloud tasks (Levis & Barriuso, 2012; Munro, 2008; Saito & Plonsky, in press), (e) is relatively easy to administer and score, (f) has been shown to positively inform pronunciation learning (Lee, 2015), and (g) has been rigorously evaluated using an argument-based validity framework (Kane, 2013; Chapelle et al., 2008, 2010). While other dedicated, instructionally-relevant pronunciation assessments may share some of these features (e.g., Dlaska & Krekeler, 2008; Kim, 2006; Lappin-Fortin & Rye, 2014, Tsurutani, 2008), they do not possess all of them. Beyond the KPD and Kim (2006), there would seem to be few, if any, other detailed, instructionally-relevant assessments of L2 Korean pronunciation (Lee, 2017b). At the same time, the KPD may be the first diagnostic tool for a productive language skill to successfully incorporate Alderson et al. (2014) and Harding et al.'s (2015) recommendations for DLA instruments to be designed based heavily on language learning theory, to have discrete tasks focused on lower-level aspects of language processing, and to provide feedback that is directly relatable to subsequent instruction (see also Lee, 2015). Viewed this way, I believe the KPD fills gaps in pronunciation assessment and DLA and represents a new direction for the interface between pronunciation teaching, learning, and assessment.

A key feature of the KPD is its capacity to raise learner awareness, which in turn promotes learner attention to phonological forms in their regular language use and during specific (classroom or self-directed) learning activities. This is very likely to be beneficial to pronunciation learning based on a body of research on the relationship between phonological awareness and pronunciation outcomes (Kennedy & Trofimovich, 2010; Moyer, 2014; Saito, 2018; Venkatagiri & Levis, 2007). Importantly, through KPD results, self-assessments, and interview data, I was able to observe how learner misperceptions could be addressed through the provision of diagnostic feedback. There is a potential link here to learners' perception skills: Those who cannot hear their own difficulties (or strengths) are more likely to have lacks or errors in awareness, both of which would hinder attention-focusing on critical pronunciation targets. Due to the challenges and constraints of classroom teaching, even phonologically-knowledgeable and experienced teachers may not always be able to help students fill in or correct gaps in their awareness of pronunciation difficulties (see Chapter 8), which further highlights the utility of the KPD.

The KPD stands out as a pronunciation assessment tool that incorporates and promotes the well-supported link between the perception, production, and learning of L2 speech sounds (Flege, 1995; Möttönen & Watkins, 2009; Nora, Renvall, Kim, Service, & Salmelin, 2015; Sakai & Moorman, 2018). Aside from expectations of relationships perception and production abilities generally being met in the results of this dissertation, what is perhaps more encouraging from a learning perspective is the attention learners gave to perception when interpreting and applying their KPD results. Learners reacted strongly to low perception scores for individual phonemes and reported specific learning activities related to perception, such as devoting more attention to target sounds in their input (e.g., from their teacher) and deliberately using audio models (e.g.,

songs) when practicing their pronunciation. Language tests are known to have washback effects on teaching and learning (Messick, 1996): Tests influence the *what* or *how* of language teaching and learning in classrooms or learner practices. As shown in this dissertation, pronunciation assessments that incorporate learning principles into their design and feedback have the potential to positively influence learners' awareness and application of these principles in their self-directed learning activities, a positive form of washback. Furthermore, specific to the perception-production link, there seems little reason not to incorporate it in the design of pronunciation assessments: Perception activities are widely recommended in pronunciation instruction (Celce-Murcia et al., 2010; Derwing & Munro, 2015; Thomson, 2011), perception items are quick to administer and easy to score, and promoting perception practice would likely have only beneficial side-effects on L2 listening abilities (Field, 2013; see also Vandergrift & Goh, 2012; Yeldham & Gruba, 2014).

Rigorous examination of the KPD was facilitated by the application of argument-based validity. Language testing specialists have been grappling with ways to ensure that tests, including diagnostic tests, that they make are reliable and valid. While test reliability is rather easy to evaluate psychometrically (and psychometric test properties are mostly undebatable), validity is anything but easy, as the concept itself and ways to investigate it in relation to a test and its scores is debated: Is it a relatively straightforward relationship between test scores and what is being measured (e.g., Borsboom et al, 2004) or a more sprawling concept that extends to stakeholder use of test scores and consequences of that use (Messick, 1989)? For DLA at least, with its strong emphasis on usefulness in subsequent instruction, I believe Messick's broader view necessarily prevails. Diagnostic tests, like other tests concerned with the interpretation and use of test scores, can be placed and investigated within an argument-based framework to

examine the validity of score uses (Bachman & Palmer, 2010; Chapelle, Cotos, & Lee, 2015; Chapelle, Enright, & Jamieson, 2008, 2010; Kane, 2013), which I did in this dissertation. Although argument-based validity theorists in educational assessment (e.g., Kane, 2013) and language assessment (Bachman & Palmer, 2010; Chapelle et al., 2008, 2010) differ somewhat in their specifications of validity arguments, the general structure and approach I took involved a series of progressive inferences that led from test-taker responses to the use of test results by a range of stakeholders. As demanded by the validity argument I constructed, I collected a wide range of relevant evidence (learner background data, self-assessment, oral proficiency measure, spontaneous speech samples, interviews, and KPD retests) that would bear on the critical evaluation of each inference. This work provided the backbone of the dissertation, and opened doors to future research questions that must be investigated in more detail.

Important Questions and Tentative Answers

After analyzing the KPD phoneme parcel data, it was clear that some phonemes presented no substantial difficulty for virtually any learner. This raises the question of whether all phonemes need to be assessed when diagnosing segmental pronunciation. Trimming has the obvious benefit of freeing up resources to either collect more information about other phonemes or include more aspects of pronunciation in diagnosis. However, at the outset of this project, I did not wish to make any assumptions about what might be possible in terms of learner pronunciation weaknesses. Now, with data in hand, I feel it is appropriate to consider possible delimitations. For L2 Korean specifically, it appeared that the phonemes /ɑ, i, ɛ, h, m/ were universally not problematic in either production or perception. Given the wide range of L1 backgrounds and levels of language experience in the sample, it appears that these sounds might be non-issues for virtually any learner. In some cases, deciding whether a phoneme is non-

problematic is a little less clear cut. /u/, for example, was very easy in production at the group level, yet was (a) flagged as a substantial difficulty for a small number of learners and (b) presented a considerable challenge in terms of perception. I would argue that /u/ should be kept.

For other languages, I recommend at least collecting pilot data on all phonemes with learners from a range of backgrounds and then see what might be appropriate to trim. For English, where it might be desirable to diagnose based on a delimited set of phonemes deemed crucial for lingua franca communication (Jenkins, 2002), I feel the need to point out that just because certain phoneme *contrasts* might be unimportant (e.g., English /θ-ð/), intelligibility issues related to constituent members of the contrast cannot be ruled out (e.g., imagine a learner whose /θ/ articulation is closer to [s] in words like *think* or *thin*). In cases such as this, explication in scoring criteria might best handle the delimitation rather than removal of certain phonemes from test specifications.

Similarly, pedagogical arguments have been made to prioritize segments and contrasts with high functional load (Kang & Moran, 2014; Munro, Derwing, & Thomson, 2015). This can lead to some sensible recommendations, such as not devoting too much time to the low-FL English /θ-ð/ distinction, but in other cases, application of FL to teaching and assessment is less intuitive. For example, in Korean, the highest FL contrast for vowels is /i-ɛ/ and the highest FL vowel is /i/ (Oh et al., 2015). However, on the KPD, /i/ was one of the easiest phonemes for learners to produce and perceive (see Chapter 5), and I suggested that it could be excluded from a revision of the KPD. In instructional settings, it would seem most learners would have a phoneme extremely close to /i/ in their linguistic repertoires that could be immediately drawn on in Korean (Flege, 1995; see Chapter 7, where learners across the range of L2 Korean oral proficiency had high production and perception accuracy), suggesting limited benefit of a

pedagogical focus on this high FL phoneme. On the other hand, for consonants, the contrast with highest FL is /l-n/, with /n/ as first- and /l/ as the third-highest FL consonants. While /n/ and /l/ were not in general difficult for learners to produce and perceive (>90% accuracy on average), several individual learners did have considerable difficulty with these two sounds, suggesting that assessment inclusion is justified and instructional attention would be worthwhile for learners who need it. FL could conceivably be applied in the construction of diagnostic assessment items. In KPD perception tasks, I primarily designed stimuli (Task 3 – Pronunciation Judgment) and distractors (Task 4 – Nonword Identification) on the basis of articulatory and acoustic similarity. This did yield items featuring the high-FL /i-ɛ/ and /l-n/ contrasts. However, other high-ranking FL contrasts like /i-o/ or /k*-t/ did not appear like they would serve as useful distractors due to considerable articulatory differences – items based on such contrasts would likely be extremely easy for learners, resulting in overestimation in the robustness of their perception of target phonemes. In sum, while FL may have some useful pedagogical applications, such as eliminating low-importance phonemes from pronunciation syllabi, it is less clear how FL might be applied more broadly in pedagogy and the specification of pronunciation diagnostic tests.

In designing KPD tasks, I devoted considerable attention to tapping into lower-level speech production and perception processes (Field, 2011, 2013). However, in these models, the role of speededness or automaticity is not emphasized, presumably because naturalistic language processing is generally assumed to be occurring at the speed of real-time communication. On the other hand, the pronunciation instruction literature has largely favored outcome measurement tasks that are discrete, controlled, and mostly unspeeded (Lee, Jang, & Plonsky, 2015; Saito & Plonsky, in press). Saito and Plonsky's (in press) recently proposed framework for measuring L2 phonological knowledge in instructional settings distinguishes between pronunciation in

controlled, relatively unspeeded contexts, tapping into declarative phonological knowledge, and pronunciation in spontaneous, speeded production, tapping into the degree of proceduralization and automatization of phonological knowledge. How do the conceptualizations of declarative and procedural phonological knowledge impact diagnostic assessment? How might this distinction be applicable to the perception of phonological features? I believe the answers to these questions have substantial implications for construct representation as well as hold instructional implications in line with the skill acquisition theory (DeKeyser, 2017) that Saito and Plonsky draw on. I also see a tension between diagnostic assessment and Saito and Plonsky's framework, as DLA experts have begun to recommend discrete, controlled tasks (Harding et al., 2015, who did deal with speed in terms of listening stimuli rate of speech or articulation but not response time), which according to Saito and Plonsky falls short of fully understanding learners' pronunciation abilities. Alderson (2005) anticipated aspects of this tension, noting that "speeded diagnostic tests" could be more useful for assessing proceduralized (or "implicit") knowledge but cautioned that speeded tests may have limited diagnostic utility: "Knowing that somebody reads slowly does not tell us why that person reads slowly, which is the essence of diagnosis" (pp. 260-261).

I deliberately designed the KPD to include multiple tasks for measuring production and perception knowledge, but in my original specifications I did not explicitly account for declarative and procedural knowledge. Retrospectively, I believe that Task 1 – Picture Naming would loosely fit Saito and Plonsky's definition of a *spontaneous production* task due to its primary focus on meaning (test-takers must first search for a word that matches the meaning of the picture) and relative degree of speededness (the whole word must be uttered at once), reflecting to some extent learners' proceduralized pronunciation ability. Task 2 – Nonword

Reading, due to its greater structure and form-focus (i.e., provision of phonemes that learners must produce; no concern for meaning), likely qualifies as a *controlled knowledge* task, which more likely reflects consolidated declarative pronunciation knowledge. Saito and Plonsky (in press) do not apply their framework to receptive phonology, but I think such an extension is possible and logical, in line with work on receptive morphosyntactic knowledge (e.g., Suzuki, 2017; Suzuki & DeKeyser, 2017). I would hazard to say that both KPD receptive tasks are controlled knowledge tasks, as the lack of time constraint on the task responses allows learners to deliberately call on their declarative representations of phonemes and compare them to the stimulus for as long as they hold the stimulus in their phonological short-term memory. While I did not report separate scores according to phoneme accuracy in each type of task, I see no reason why it would not be both possible and useful to do so. I recommend that future development of the KPD and other pronunciation diagnostics consider Saito and Plonsky's declarative/procedural distinction, including both (alleviating Alderson's 2005 concerns) could yield information highly pertinent to instructional planning, such as whether to focus on explicit phonetic instruction or simply to practice in more communicative contexts.

For many good reasons, L1 influence holds a prominent place in the study of L2 pronunciation. L1 is widely thought to influence interlanguage phonology (Flege, 1995) and psycholinguistics research widely agrees that the phonemes of L1s and other languages remain active during L2 speech perception (Imai et al, 2005; Weber & Cutler, 2004). On a more practical level, learner L1-specific pedagogical recommendations are abundant (e.g., Avery & Ehrlich, 1992; Derwing & Munro, 2015; Kwon, 2017). In the DLA literature, too, recommendations for L1-based tailoring of instruments can be found (Harding et al., 2015). Should learner L1 play a prominent role in the design of pronunciation diagnostics? I argue that

they should not. In the cluster analysis in Chapter 7, although many learners who had pronunciation difficulties shared them with many of their same-L1 peers, I also found that learners from the same L1 background could have distinct differences in terms of phoneme difficulties. For example, there were English and Japanese learners who struggled to produce tense consonants, which a simple contrastive analysis would predict. However, there were also small numbers of English and Japanese learners who had greater difficulties with aspirated consonants – a phenomenon that would not be so readily predicted by a contrastive analysis, given that English and Japanese have aspiration. Contrastive analysis or other L1-influence approaches risk over-simplifying cross-linguistic differences (e.g., in the case of English and Japanese, aspiration and voicing have featural overlap which could lead to less-predictable feature transfer or phoneme assimilations in L2 Korean), and do not necessarily account for how *individuals* react to unfamiliar new features of the L2 (or L3+, as it may be) phonology. As I have already mentioned, there may be suitable grounds for excluding a small number of segments from a diagnostic across the board, but outside of that I see little compelling reason to further reduce the domain of segments based on learner L1s. Doing so could lead to overlooking difficulties of less typical learners within an L1, and would also limit the applicability of the diagnostic itself, resulting in potentially less mileage out of test development efforts.

Room for Expansion

At the outset of this project, I knew I would have to delimit the scope of the diagnostic instrument, as it would have been beyond my means to develop and validate a truly comprehensive diagnostic of second language Korean pronunciation (if such a diagnostic is even possible). While I feel that my initial focus on segmentals is justified on both practical and theoretical grounds, I remain cognizant of the value of collecting diagnostic information related

to other important aspects of pronunciation. Comments from learners and the teacher I interviewed provided additional reminders of this. Some learners felt that intonation was where they struggled most, that they had difficulties related to (certain phonemes in combination with) syllable structure, or that they needed to work on their ability to apply phonological processes in connected speech. The teacher I interviewed, Jae-woo, also brought up intonation issues with one of his students, and pointed out another student's weaknesses in facial expression and gesture which are known to be important parts of a speaker's repertoire capable of enhancing interlocutor understanding (e.g., Hardison, 2018; Sueyoshi & Hardison, 2005).

It is thus more appropriate to view the KPD and present study as one piece of the diagnostic puzzle. As Alderson et al. (2014) pointed out, diagnosis is a process that begins with a diagnostician and benefits from multiple sources of evidence. Imagine a more robust and holistic classroom-based diagnosis context, with the teacher featured in this dissertation, Jae-woo (the Korean teacher featured in Chapter 9), as a diagnostician whose informal observations of Yu-wen's (one of the learners from Chapter 9) pronunciation difficulties initiate the process of diagnosis. As Jae-woo noticed some segmental difficulties in Yu-wen's production, he might ask her to complete a self-assessment and administer the KPD. Jae-woo might also utilize other diagnostic tools or observations to provide Yu-wen on feedback related to her gesturing and expression while speaking. Afterwards, Jae-woo could apply the diagnostic information by recommending self-study material or providing some additional homework assignments. Yu-wen, as an active participant in the diagnostic process, might share her lack of confidence in pronunciation with Jae-woo, who in turn might be able to counsel her on the affective challenges involved with second language pronunciation.

Such a view on diagnosing L2 pronunciation points to many possibilities for diagnostic instrument development and formulation of principles and procedures for teacher-driven (or teacher-guided) diagnosis. I believe the KPD provides a useful starting point for diagnosing segmental pronunciation (though certainly improvements are possible), and broadly speaking, the KPD represents a set of test specifications that (a) worked as intended and (b) could be easily adapted to other target languages. More original work is needed in the development and validation of practical, reliable, and sufficiently detailed diagnostic tools for suprasegmental aspects of L2 pronunciation and pronunciation supports such as gesture and communication strategies. Training materials and procedures for teachers to diagnose learner pronunciation are other promising avenues for further developing pronunciation diagnosis.

Towards an Interface between Pronunciation Instruction and Diagnostic Assessment

While this dissertation can primarily be viewed as test development and validation project, it is a *diagnostic* test development and validation project in which links to instruction and pedagogy are important. Although I did not examine traditional classroom-based pronunciation instruction (e.g., Isbell et al., 2019) or the use of a structured training program (e.g., Thomson, 2011), Chapter 8 touched on issues related to pronunciation teacher cognition as well as out-of-class and autonomous pronunciation learning activity.

Pronunciation teacher cognition deals with the "knowledge, beliefs, thoughts, attitudes, and perceptions" of teaching pronunciation (Burri, Baker, & Chen, 2017, p. 110, see also Baker, 2014), and is an under-researched area in general. My interview with Jae-woo, an in-service teacher with considerable knowledge of phonology and experience teaching students from a variety of L1s, had an orientation to teaching pronunciation that was largely driven by learner L1. He also saw considerable constraints on his classroom pronunciation teaching practices,

resulting in limited use of teacher-centered read-alouds and repetitions as his primary means of addressing pronunciation in the classroom (Baker, 2014; Foote et al., 2013). When asked how he would apply the KPD's diagnostic feedback, he described separate classes that focused on explicit instruction tailored to addressing difficulties related to L1-influence. What is interesting here, to me, is how pronunciation teacher cognition and teaching practices might interface with teacher competence in diagnostic assessment of pronunciation issues (Edelenbos & Kubanek-German, 2004). Jae-woo mostly described his two students' pronunciation difficulties broadly in terms of L1 interference and did not offer much detail in terms of specific phonemes each learner experienced. By Jae-woo's own admission, he missed "50%" of the picture (in his defense, I would like to point out he was able to comment insightfully on some other non-segmental aspects of their pronunciation). I wonder: Did Jae-woo's strong orientation to L1 influence in L2 pronunciation constrain his diagnosis of his students' specific segmental difficulties? Research examining pronunciation teacher cognition alongside diagnostic assessment practice would be informative and potentially provide guidance to teacher training. Although knowledge of learner L1 phonology and common transfer-related influence is almost certainly useful for pronunciation teaching, it may be time to consider a shift to training teachers how to observe individual difficulties without relying solely on L1-based assumptions (e.g., Avery & Ehrlich, 1992; Kwon, 2017).

Findings related to actual learner application of KPD results were just as interesting as the teacher's orientation to diagnosing pronunciation difficulties. The learning activities and strategies applied by learners included:

- Shadowing
- Reading aloud

- Song-based practice
- (Seeking) feedback in meaning-focused interaction
- Heightened attention to target sounds in input and own output

Most of these strategies and learning activities find at least some support in research (Derwing, Munro, & Wiebe, 1998; Foote & McDonough, 2017; Horgues & Scheuer, 2014; Loewen & Isbell, 2017; Moyer, 2014; Saito & Lyster, 2012) or commonly-used and recommended pedagogical practices (Baker, 2014; Celce-Murcia et al., 2010). However, much of the support for these practices is based on more formal instructional contexts, such as teacher-led classrooms or well-structured computer-assisted pronunciation training tools. Moyer's (2014) review of exceptional L2 pronunciation outcomes emphasized the role of learner autonomy and engagement in activities and strategy use to promote L2 phonological learning. This raises several questions: How effective are these practices in non-classroom/autonomous learning contexts? Which activities might be best suited for learners to effectively pursue on their own time and initiative? Which activities can learners be trained to do on their own without investing a great deal of time? Empirical research addressing these questions would contribute to the less-understood facet of self-directed/autonomous pronunciation instruction and at the same time address the DLA-instruction interface: If learner feedback from a test like the KPD could be integrated with recommendations for specific, effective self-directed learning activities, learner pronunciation development could be positively impacted.

Finally, I believe this dissertation should motivate L2 pronunciation researchers to consider more broadly the role of assessment in promoting student pronunciation learning. Previous low-stakes, instructionally-relevant pronunciation assessment efforts, such as Lappin-Fortin and Rye (2014), have shown the potential of self-assessments and pre-post achievement

tests for raising learner awareness and tracking learning outcomes. The KPD has shown how diagnostic pronunciation assessment can shape and promote learner attention, pronunciation learning strategy use, and out-of-class pronunciation learning activity. More research on the interface of pronunciation instruction and diagnostic assessment could lead to more concrete recommendations for practitioners and more fruitful self-directed learning for students.

Implications for Diagnostic Language Assessment

I now turn to implications for diagnostic language assessment. Through the course of setting a purpose and scope of diagnosis, developing the KPD, constructing a validity argument, and seeking evidence to support the validity of KPD score interpretation and use, I have arrived at several key considerations for DLA pertaining to grain size, measurement models, validity, and DLA instrument design.

As discussed in Chapter 2, grain size is a key consideration in DLA. Finer grain-size in the design and subsequently test scores can lead to more concrete, instructionally-relevant information to be utilized by stakeholders. At the same time, finer grain size has an inverse relationship with practicality, requiring more and more tasks, items, or other observations in order to isolate smaller bits of linguistic knowledge and competence. In other tests labeled as diagnostic, learners are given more general feedback, such as feedback on how well the understand main ideas or details. To make a crude analogy to pronunciation diagnosis, that would be like giving a learner just two scores for *segmental* and *suprasegmental accuracy*, which perhaps might be a useful starting point, but such score categories are of too large a grain size to provide concrete guidance for instruction. With the KPD, I feel that I have struck an effective balance: A substantial subcomponent of pronunciation (and in turn, of speaking ability) is diagnosed at the level of individual phonemes, which are discrete and relatively fine bits of

linguistic knowledge, over the course of about 15 minutes and in a format which only takes a teacher roughly 5 additional minutes to score (caveat: compiling results may take extra time without the specialized software I used to administer the test). The teacher I interviewed in Chapter 9, Jae-woo, nonetheless expressed desire for an even finer grain size: He wanted to know about allophone-level difficulties (i.e., performance in different syllable positions/phonological environments) *within* each phoneme. I can certainly see the pedagogical application of such information and did consider allophonic distribution of phonemes in the design of the KPD (Chapter 4). However, concerns related to reliability/information sufficiency and practicality (test length and complexity of score reports) led me to avoid fully pursuing that level of detail in the KPD's design and reporting of results. Was that the right call? I leave that question open to readers. Nonetheless, it appears that relatively fine grain size in diagnostic instruments has benefits, as shown by the KPD's utility in helping learners narrow down their list of study targets and sounds to pay special attention to in their general Korean use.

In an interview featured on Glenn Fulcher's *Language Testing Bytes* podcast, Eunice Jang expressed hope for methodological diversification in DLA, including "psychometrically less constrained diagnostic modelling, such as latent class or profile analysis, clustering methods, or subscoring approaches" (Fulcher, 2015). I agree with Jang, as doing anything else would likely limit DLA to retrofitting existing proficiency tests to provide more detailed score reports (Jang, 2009) due to sample size and resource constraints associated with test development and administration that is not large-scale and high-stakes. While it would have been wonderful to collect data from 1,980 (or 19,800) L2 Korean learners and construct more sophisticated psychometric models (e.g., cognitive diagnostic models based on combinations of phonological features and syllable/phonological contexts), I must wonder what practical use that would have

yielded. At the most immediate level of interpreting diagnostic test scores, it appears to me that easily interpretable subscores linked to concrete learning targets are key for learners and teachers. Such subscores are useful for delimiting study targets and for promoting awareness of a manageable, tailored list of targets when using the language. While I concede that the identification of a range of stable diagnostic profiles through very large datasets could be useful for tracking students into predetermined instructional modules, doing so would be fruitless without following through on the development of such modules.

This dissertation and other work on diagnostic instruments, such as those developed for L2 writing, have utilized argument-based validity to set an agenda for validation research that gathered necessary evidence to support the use of a test. Tasks in L2 writing diagnostic tests tend to more closely resemble authentic writing tasks instead of discretized tasks that target subcomponents of writing ability, with diagnostic information coming from thorough, detailed analysis of written products (Chapelle et al., 2015). As such, validity arguments for such writing diagnostics have been able to support extrapolation inferences with rather straightforward links between test tasks and real-world writing tasks. In contrast, I have encountered a challenge in establishing an appropriate connection between a highly-discrete diagnostic test and real-world, meaningful language use. Extrapolation support for a test like the KPD based on task features or parameters, e.g., pointing out that it elicits words and phonemes used in real-world Korean use, would seem to be trite. This leaves alignment between test task responses with authentic language use performance as a suitable source of support. The unreasonableness of expecting (composite) scores derived from highly discretized diagnostic assessment task responses to closely reflect learner language behavior in spontaneous and holistic language use begs the question: How much support for extrapolation is needed? I do not have a clear answer to that

question. Presumably, a comprehensive set of discrete diagnostic scores pertaining to a communication skill or subskill (say, diagnostic information coming from both segmentally- and suprasegmentally-focused pronunciation diagnostics) should be able to explain a large portion of learner performance in authentic language use. However, I suspect that any one piece of the diagnostic picture can at most explain a proportionally small piece of the larger picture in authentic language use, and even this relationship may be difficult to isolate due to the confluence of linguistic, cognitive, and situational factors that bear on typical language use. I see sorting through this issue as an important area of work for DLA, especially in the approach to DLA espoused by Alderson and colleagues (Alderson, 2005; Alderson et al., 2015; Harding et al., 2015).

Finally, with respect to developing instruments for DLA, I believe this dissertation serves as proof of concept for Harding et al.'s (2015) recommendations for diagnostic instrument design. Harding et al. focused on diagnosing L2 reading and listening skills, and I was able to apply many of their recommendations for diagnosing listening subskills (i.e., phoneme perception). I was also able to adapt their recommendations to an aspect of speaking ability (i.e., pronunciation). The resulting product, a new test built from the ground up on the basis of learning theory and models of linguistic processing, was capable of providing detailed linguistic feedback that could be appropriately interpreted and applied to support learning. Without discounting other purposes and practices in L2 assessment sometimes referred to as diagnostic, such as retrofitting existing proficiency tests to provide enhanced score reports to learners (Jang, 2009; Jang et al., 2015) or to identify students requiring additional language support broadly writ (Knoch & Elder, 2016), I see the Harding et al. model of diagnostic instrument design, exemplified by the KPD, as the way forward for DLA.

Final Thoughts

The KPD and this dissertation, for all their limitations (and there are many), represent a considerable undertaking: I developed a brand-new test with four distinct tasks and a total of over 350 items through multiple rounds of piloting, field tested it with nearly 200 learners, and investigated the validity of its interpretation and use through the collection and analysis of diverse types of evidence. I hope that the fruits of these efforts extend beyond the pages of this dissertation, as the final product appears to be useful in promoting the pronunciation development of L2 Korean learners. At the very least, I believe this research benefited the learners who took the KPD and received score reports, many of whom expressed appreciation for and sincere interest in their results and the research itself. To the end of reaping more value from the KPD development and validation efforts, I plan to produce and release a free, publicly available version of the KPD with user-friendly documentation and score-calculation tools for Korean teachers and/or tutors to use.

More broadly, I believe the kind of test development and validation efforts showcased in this dissertation raise important points for how language testers can contribute to learning-oriented (Turner & Purpura, 2015), instructionally-relevant (Pellegrino et al., 2016), low-stakes assessments. The first point is that rigorous test development and validation can be worth it for low-stakes assessments. While many low-stakes assessment practices and tools are justifiably simple and economical, low-stakes does not necessarily have to be synonymous with low-value: A high-quality diagnostic assessment tool, with interpretations supported by rigorous validation research, can provide teachers and learners with relatively easily-obtained, detailed information mostly unavailable from observations and other informal assessments.

The second point is that these sorts of test development and validation efforts are not reasonable to expect from classroom teachers, who often assume responsibility for the creation of many, if not most, classroom assessments. Rather, I see the development of instruments like the KPD as a prime opportunity for language testing professionals and researchers to contribute to learning-oriented, instructionally-relevant, low-stakes classroom assessment practices: Build a useful tool and put it in the hands of teachers and learners (for free if possible, or at least for cheap). I believe such efforts are an opportunity for language testers to do more good in our work by creating something that is directly and concretely useful to language learning, above and beyond producing knowledge and developing theory through academic research that may (or may not) be relevant to low-stakes classroom assessments.

My final point is that in both the design and validation phases of this instructionally-relevant assessment project, I had to give considerable attention to SLA theory and pronunciation instruction concerns. This is not common in development efforts for many kinds of language tests, such as proficiency tests which are more concerned with norm-referenced construct definitions and domain of use descriptions over developmental trajectories and applicability of results to teaching and learning practices. Here, with DLA and instructionally-relevant language assessments more broadly, I see an opportunity for greater interaction and collaboration among language assessment professionals, SLA and especially ISLA researchers, and language pedagogy experts. I hope that this dissertation inspires more of those connections to be made.

APPENDICES

APPENDIX A

KPD Table of Specifications

Table A1
KPD Table of Specifications

Target		**Production – Part 1**		**Perception – Part 2**		
Consonants	Context	1. Picture naming	2. Nonword Reading	3. Pronunciation Judgement	4. Identification	# of items
ㅂ p (lax bilabial stop)	initial	2 불, 버스	1 보	1 비	1 바	13
	medial	1 나비	1 우부	1 일본	1 오보	
	final	1 집	1 압	1 입	1 웁	
ㅃ p^* (tense bilabial stop)	initial	2 빨간색, 빵	1 빠	1 뿌리	1 삐	7
	medial		1 오뽀	1 아빠	1 우뿌	
	final					
ㅍ p^h (aspir. bilabial stop)	initial	2 피아노, 포도	1 포	1 팔	1 푸	9
	medial	1 컴퓨터	1 우푸	1 소파	1 이피	
	final					
ㄷ t (lax alveolar stop)	initial	2 돈, 돼지	1 도	1 다리	1 디	14
	medial	1 포도	1 이디	1 바다	1 아다	
	final	1 초콜릿	1 안	1 옷	1 욷	
ㄸ t^* (tense alveolar stop)	initial	2 딸기, 땅콩	1 따	2 똥, 딸	1 뚜	8
	medial		1 우뚜		1 오또	
	final					
ㅌ t^h (aspir. alveolar stop)	initial	2 토끼, 택시	1 토	1 탈	1 티	9
	medial	1 컴퓨터	1 아타	1 외투	1 우투	
	final					
ㄱ k (lax velar stop)	initial	2 귀, 그림	1 기	1 개	1 구	14
	medial	2 시계, 발간색	1 우구	1 미국	1 오고	
	final	1 발간색	1 옥	1 책	1 익	
ㄲ k^* (tense velar stop)	initial	1 꽃	1 까	1 꿀	1 꼬	8
	medial	1 토끼	1 이끼	1 어깨	1 우꾸	
	final					
ㅋ k^h (aspir. velar stop)	initial	1 코	1 키	2 칼, 카메라	1 쿠	8
	medial	1 땅콩	1 오코		1 아카	
	final					
ㅈ tɕ (lax alv-pal. affric.)	initial	1 집	1 자	1 짐	1 지	12
	medial	5 돼지, 여자, 의자, 아저씨, 화장실	1 이지	1 사진	1 오조	
	final					

Table A1 (cont'd)

ㅉ tɕ* (tense alv-pal. affric.)	initial		1 쭈		1 쪼	8
	medial	2 맥주, 왼쪽	1 오쪼	2 잡지, 오른쪽	1 아짜	
	final					
ㅊ tɕʰ (aspir. alv-pal. affric.)	initial	2 침대, 초콜릿	1 치	1 차	1 추	9
	medial		1 우추	1 기차	1 이치	
	final					
ㅅ s (lax alv-pal. fricative)	initial	1 새	1 수	1 소	1 사	18
	medial	3 버스, 빠간색, 원숭이	1 아사	1 세상	1 우수	
	before i/j	2 시계, 화장실	2 샤,셔	2 음식, 도시	2 시, 쇼	
ㅆ s* (tense alv-pal. fricative)	initial	1 쓰레기	1 쏘	2 쌀, 싸움	1 쑤	13
	medial	1 학생	1 우쑤		1 오쏘	
	before i	2 아저씨, 택시	1 씨	1 접시	1 씨	
ㅎ h (lax glottal fricative)	initial	2 학생, 화장실	1 하	1 호랑이	1 히	8
	medial		1 오호	1 아홉	1 우후	
	final					
ㅁ m (bilabial nasal)	initial	1 맥주	1 미	1 목	1 모	13
	medial	2 레몬, 침대	1 오모	1 나무	1 우무	
	final	1 그림	1 움	1 사람	1 임	
ㄴ n (alveolar nasal)	initial	1 나비	1 노	1 노래하다	1 니	13
	medial	2 피아노, 빨간색	1 이니	1 하나	1 아나	
	final	1 돈	1 안	1 문	1 온	
ㅇ ŋ (velar nasal)	initial					13
	medial	2 화장실, 땅콩	1 옹오	1 창문	1 앙아	
	final	5 빵, 용, 땅콩, 학생, 왕	1 잉	1 가방	1 웅	
ㄹ l (alveolar liquid)	initial	1 레몬	1 루	1 라디오	1 라	19
	medial	3 쓰레기, 그림, 빨간색	1 이리	1 사랑	1 오로	
	final	2 불, 화장실	1 알	1 별	1 울	
	geminate	1 초콜릿	1 울루	1 콜라	1 일리	
Vowels						
ㅣ i (high front unrounded)		10 집, 지아노, 화장실, 돼지, 쓰레기, 아저씨, 나비, 택시, 토끼, 원숭이	1 이	2 아기, 시장	1 이	14
ㅐ,ㅔ ɛ (mid front unrounded)		8 레몬, 쓰레기, 택시, 맥주, 빨간색, 새, 시계, 침대	1 에	2 백, 뱀	1 에	12

Table A1 (cont'd)

ㅡ ɯ (high back unrounded)		3 그림, 쓰레기, 버스	1 으	2 하늘, 음악	1 으	7
ㅓ ʌ (mid back unrounded)		3 버스, 컴퓨터 (x2)	1 어	2 커피, 머리	1 어	7
ㅏ ɑ (low back unrounded)		9 아저씨, 빵, 피아노, 빨간색(x2), 학생, 나비, 의자, 여자	1 아	2 산, 강	1 아	13
ㅜ u (high back rounded)		3 불, 원숭이, 맥주	1 우	2 눈, 둘	1 우	7
ㅗ o (mid back rounded)		11 꽃, 돈, 포도(x2), 레몬, 초콜릿(x2), 토끼, 왼쪽, 피아노, 코	1 오	2 손, 호주	1 오	15
Glides*						
j_ /_j		4 여자, 의자, 용, 컴퓨터	6 애, 의, 여, 야, 유, 요	6 고양이, 교수님, 의사, 연필, 우유, 예술가	6 애, 의, 여, 야, 유, 요	22
w_		6 왼쪽, 돼지, 귀, 왕, 화장실, 원숭이	4 위, 왜, 워, 와	4 교회, 원, 가위, 화	4 위, 왜, 워, 와	18
	Total	128	63	72	63	326

*Glides combine with monopthongs to form 10 diphthongs: ㅖ/ㅒ (jɛ), ㅢ(ɯj), ㅕ (jʌ), ㅑ (jɑ), ㅠ(ju), ㅛ(jo),ㅞ/ㅙ(wɛ), ㅝ(wʌ), ㅟ(wi), ㅘ(wɑ)

APPENDIX B

KPD Item Specifications

I. Specification Title: Picture Naming

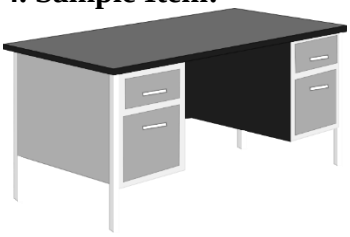
1. General Description: Learners should be able to recall phonological representations of words and articulate them accurately.

2. Prompt Attributes: A picture and English text will be displayed. Prompts are words selected due to 1) high frequency 2) word class (nouns are more image-able) and 3) length (preference for shorter words- less potential distraction). Pictures should clearly elicit the target word. Images should thus use color, indicators such as arrow or circles, and possibly even text (but not for the target word) to ensure that expected responses are given.

3. Response Attributes:

Responses scored by human judgment. Judge should be trained in Korean phonology and native/near-native proficiency. Responses will be scored for accuracy of all phonemes.

4. Sample Item:



<expected response is "책상">

<4 seconds are given to respond>

5. Supplemental Information: Phoneme inventory includes 19 consonants, 7 vowels, and 2 glides (which may combine with vowels to form 10 diphthongs), following Shin, Kiaer, and Cha (2013). Refer to Table of Specifications for environments that must be tested for each phoneme. Refer to *A Frequency Dictionary of Korean* (Lee, Jang, & Seo, 2017) for word frequency information; words among top 1,500 most frequent are preferred but any word in the top 5,000 and/or determined to be commonly introduced in instructional settings is permissible.

6. Directions and Practice Item(s)

Directions: In this section, will name pictures. First, you will see a picture and a sentence with a blank. Then, you will speak the word for the picture and the blank.

Practice Item A:

You should have said "책상".

II. Specification Title: Nonword Reading

1. General Description: Learners should be able to articulate sounds of Korean.

2. Prompt Attributes: A nonword target of one to two syllables (V, CV, VC, or CVC) will be displayed. For consonant items, the Korean vowels /i/ /a/ /o/ and /u/ will be used to provide context, as these are common in many languages and likely to present little challenge. For glides and some allophones, more complicated syllables may be used, but these should not be unnecessarily complex. Items targeting vowels will utilize 1 syllable targets with a single vowel.

3. Response Attributes:

Responses scored by human judgment. Judge should be trained in Korean phonology and native/near-native proficiency. Responses will be scored for accuracy of the target phoneme only.

4. Sample Item:

"다" is displayed.

<the test taker has two seconds to respond>

5. Supplemental Information: Phoneme inventory includes 19 consonants, 7 vowels, and 2 glides (which may combine with vowels to form 10 diphthongs), following Shin, Kiaer, and Cha (2013). Refer to Table of Specifications for environments that must be tested for each phoneme.

6. Directions and Practice Item(s)

Directions: In this section, you will see Korean letters and read them out loud.

Practice Item A:

"다" is displayed.

<the test taker has two seconds to respond>

III. Specification Title: Pronunciation Judgment

1. General Description: Korean users must be able to decode speech sounds and match sounds to phonological representations of words, drawing on contextual clues. Examinees will judge the quality of phonemes in spoken Korean.

2. Prompt Attributes: A single, ideally short, word will be played once and a picture of the word will be displayed. An English caption of the picture will be provided. The picture and caption will be displayed prior to hearing the word (1 second if computer administered) The spoken word may be correct (i.e., standard pronunciation) or incorrect. Incorrect words will have one (and only one) phoneme intentionally mispronounced. Prompts are words selected due to 1) high frequency 2) word class (nouns are more image-able) and 3) length (preference for shorter words- less potential distraction).

The prompt will be recorded by a native speaker of Standard Korean with normal duration and neutral pitch/intonation.

For each prompt, the question "Is this right?" will be displayed.

3. Response Attributes: The response involves selecting "Yes" or "No" within 2 seconds after the prompt is done playing. "Yes" and "No" will be displayed side-by-side. The response will be indicated by key press or circling; in the former case a reaction time may also be recorded.

4. Sample Item:

<a speaker recites "착상" //, a mispronunciation of the second phoneme>

Is this right?

Yes No*

5. Supplemental Information: Phoneme inventory includes 19 consonants, 7 vowels, and 2 glides (which may combine with vowels to form 10 diphthongs), following Shin, Kiaer, and Cha (2013). Refer to Table of Specifications for environments that must be tested for each phoneme.

Refer to *A Frequency Dictionary of Korean* (Lee, Jang, & Seo, 2017) for word frequency information; words among top 1,500 most frequent are preferred.

6. Directions and Practice Item(s)

Directions: In this section, you judge the accuracy of Korean word pronunciations. First, you will see a picture that represents a word. Then, you will hear a sound for that word. Last, you will decide whether the word was pronounced correctly.

Practice Item A:

<a speaker recites "착상" //, a mispronunciation of the second phoneme>

Is this right?

Yes No*

The correct answer was "No". The pronunciation you heard was "착상" instead of "책상".

IV. Specification Title: Nonword Identification

1. General Description: Korean users must be able to decode speech sounds. Examinees will identify individual phonemes in spoken Korean.

2. Prompt Attributes: A one- or two-syllable nonword (when possible) will be played once. The target for each item is one Korean phoneme embedded in a *V*, *CV*, *VC*, or *VCV* nonword carrier. The different sequences allow for the phoneme to be tested in a variety of phonetic environments that span its allophonic distribution. For consonant items, the Korean vowels /i/ /a/ /o/ and /u/ will be used to provide context, as these are common in many languages and likely to present little challenge. For vowel items, /n/ /m/ /k/ /p/ /t/ are candidates for context, as they are also extremely common across world languages. These are also among the most commonly occurring phonemes in spoken Korean.

The prompt will be recorded by a native speaker of Standard Korean with normal duration and neutral pitch/intonation.

For each prompt, the question "Which sound did you hear?" will be displayed.

3. Response Attributes: The response will involve selecting one of two text options. The selection will be made by pressing a key or circling with a pen/pencil; in the former case a reaction time may also be recorded. The response may be made as soon as the audio is played, with a limit of 2 seconds per item.

The two text options will be identical except for one difference: the target phoneme. The key will match the prompt, and the distractor will be another phoneme with similar articulation/acoustical properties. For example, if the key is a tensed consonant, the distractor would be a lax or aspirated counterpart (e.g., ㄲ and ㅋ).

The two options will be displayed side-by-side, and the location of the key should be random.

4. Sample Item:

<a speaker recites the nonword "다" /dɑ/>

What sound did you hear?

a. 다* </dɑ/> b. 마 </mɑ/>

<the test taker has two seconds to respond>

5. Supplemental Information: Phoneme inventory includes 19 consonants, 7 vowels, and 2 glides (which may combine with vowels to form 10 diphthongs), following Shin, Kiaer, and Cha (2013). Refer to Table of Specifications for environments that must be tested for each phoneme. Refer to Shin et al. (2013) and Choo & O'Grady (2003) for suitable key-distractor contrasts.

6. Directions and Practice Item(s)

Directions: In this section, you will identify Korean sounds. First, you will hear a sound. Then, you will select the *Hangeul* letters that match the sound.

Practice Item A:

<a speaker recites the nonword "다" /tɑ/>

What sound did you hear?

a. 다* </tɑ/> b. 마 </mɑ/>

The correct answer was "다".

APPENDIX C

KPD Production Task Scoring Sheet

Task 1 – Picture Naming / 1 – 그림 말하기

#	단어	표적 소리들
1	빵	ㅃ ㅏ ㅇ
2	피아노	ㅍ ㅣ ㅏ ㄴ ㅗ
3	원숭이	ㅝ ㄴ ㅅ ㅜ ㅇ ㅣ
4	나비	ㄴ ㅏ ㅂ ㅣ
5	토끼	ㅌ ㅗ ㄲ ㅣ
6	여자	ㅕ ㅈ ㅏ
7	돼지	ㄷ ㅙ ㅈ ㅣ
8	아저씨	ㅏ ㅈ ㅓ ㅆ ㅣ
9	집	ㅈ ㅣ ㅂ
10	새	ㅅ ㅐ
11	택시	ㅌ ㅐ ㄱ ㅆ ㅣ
12	코	ㅋ ㅗ

#	단어	표적 소리들
13	귀	ㄱ ㅟ
14	땅콩	ㄸ ㅏ ㅇ ㅋ ㅗ ㅇ
15	컴퓨터	ㅋ ㅓ ㅁ ㅍ ㅠ ㅌ ㅓ
16	포도	ㅍ ㅗ ㄷ ㅗ
17	돈	ㄷ ㅗ ㄴ
18	화장실	ㅎ ㅘ ㅈ ㅏ ㅇ ㅅ ㅣ ㄹ
19	시계	ㅅ ㅣ ㄱ ㅖ
20	학생	ㅎ ㅏ ㄱ ㅆ ㅐ ㅇ
21	딸기	ㄸ ㅏ ㄹ ㄱ ㅣ
22	맥주	ㅁ ㅐ ㄱ ㅉ ㅜ
23	의자	ㅢ ㅈ ㅏ
24	그림	ㄱ ㅡ ㄹ ㅣ ㅁ

#	단어	표적 소리들
25	용	ㅛ ㅇ
26	침대	ㅊ ㅣ ㅁ ㄷ ㅐ
27	쓰레기	ㅆ ㅡ ㄹ ㅖ ㄱ ㅣ
28	왕	ㅘ ㅇ
29	라면	ㄹ ㅏ ㅁ ㅕ ㄴ
30	왼쪽	ㅚ ㄴ ㅉ ㅗ ㄱ
31	불	ㅂ ㅜ ㄹ
32	초콜릿	ㅊ ㅗ ㅋ ㅗ [ㄹ ㄹ] ㅣ ㄷ
33	빨간색	ㅃ ㅏ ㄹ ㄱ ㅏ ㄴ ㅅ ㅐ ㄱ
34	꽃	ㄲ ㅗ ㄷ
35	버스	ㅂ ㅓ ㅅ ㅡ

#	글	표적 소리
1	아사	ㅅ
2	쏘	ㅆ
3	어	ㅓ
4	이끼	ㄲ
5	쭈	ㅉ
6	왜	ㅙ
7	우구	ㄱ
8	도	ㄷ
9	키	ㅋ
10	오뽀	ㅃ
11	자	ㅈ
12	와	ㅘ
13	으	ㅡ
14	여	ㅕ
15	포	ㅍ
16	이리	ㄹ

#	글	표적 소리
17	얘	ㅒ
18	이지	ㅈ
19	유	ㅠ
20	기	ㄱ
21	오호	ㅎ
22	워	ㅝ
23	압	ㅂ
24	셔	ㅅ
25	하	ㅎ
26	잉	ㅇ
27	아타	ㅌ
28	이	ㅣ
29	우부	ㅂ
30	옥	ㄱ
31	에	ㅔ
32	토	ㅌ

#	글	표적 소리
33	오코	ㅋ
34	오모	ㅁ
35	의	ㅢ
36	옴	ㅁ
37	우쑤	ㅆ
38	위	ㅟ
39	알	ㄹ
40	이니	ㄴ
41	씨	ㅆ
42	옹오	ㅇ
43	빠	ㅃ
44	우푸	ㅍ
45	앋	ㄷ
46	오	ㅗ
47	루	ㄹ
48	우뚜	ㄸ

#	글	표적 소리
49	노	ㄴ
50	미	ㅁ
51	보	ㅂ
52	따	ㄸ
53	이디	ㄷ
54	우	ㅜ
55	샤	ㅅ
56	우추	ㅊ
57	아	ㅏ
58	오쪼	ㅉ
59	안	ㄴ
60	까	ㄲ
61	치	ㅊ
62	울루	[ㄹㄹ]
63	수	ㅅ

APPENDIX D

Scoring Guidelines for KPD Production Tasks

Scoring Guidelines for KPD Production Tasks

Supplies: Scoring Sheet, Pen or Pencil, Headphones

How to Score: Write the student's name or ID number on the Scoring Sheet. With the Scoring Sheet in front of you, listen to the student's audio file for Task 1 – Picture Naming and Task 2 – Nonword Reading. Judge each target sound as correct (easily identifiable) or incorrect (uncertain or unclear). Mark incorrect target sounds by crossing them out on the scoring sheet.

Scoring Criteria:

This test is designed to identify pronunciation weaknesses. However, pronunciation of target sounds does not have to be "perfect" or exactly native-like. Instead, target sounds should be clearly and easily recognizable, without ambiguity.

You should mark a target sound as incorrect if…

- It could be understood as a different Korean sound
- It is not 100% clear to you as the target sound
- You hesitate or have to really think whether you heard the target sound
- If the sound sounds starkly out of place
- The sound does not sound at all like a Korean sound

Note: Sometimes, a student will self-correct, or the test administrator will prompt them to try a different word. In this case, judge the student's final production.

한국어 발음 진단 검사 (KPD) 조음 채점법

필수품: 채점지, 연필이나 볼펜, 이어폰/헤드폰

채점법: 채점지에 학습자의 이름 또는 번호를 적는다. 채점지를 앞에 두고 학습자의 오디오 파일을 듣는다. 각 표적 소리를 정답(쉽게 구분함) 또는 오답(불확실, 분명하지 않음)으로 평가한다. 오답일 경우, 채점지의 표적소리에 줄을 그어 표시한다.

채점 기준:
이 시험은 학습자 발음의 취약점을 알아보기 위한 것이다. 하지만, 표적소리의 조음은 완벽하거나 원어민의 조음과 똑같지 않아도 된다. 그 대신에 표적소리가 애매한 것 없이 분명하고 쉽게 구분할 수 있어야 한다.

다음과 같은 경우의 조음은 오답으로 평가해야 한다:

- 표적소리가 아닌 다른 한국어 소리로 알아들을 수 있다
- 100% 표적소리가 전혀 아니다
- 조음을 들은 후에 망설이게 되고 표적소리 인지를 고민하게 된다
- 주어진 단어 환경에서 조음이 자연스럽지 않다
- 한국어의 소리가 전혀 아닌 것 같다

특이 사항: 가끔 학습자가 조음을 혼자서 수정한 후 다시 말하거나 반복해서 말할 때가 있다. 또는 시험 감독자가 다른 단어를 말하도록 유도하기도 한다. 이런 경우에는 학습자의 마지막 조음을 평가한다.

APPENDIX E

Language Background Questionnaire

배경 설문 – Background Information

1. 기본 정보 – Basic Information

성/Last Name: (영문/English)		**명/First Name: (영문/English)**		**날짜 Date:**	
국적 Home Country:		**출생년도 / Year of Birth: (예: 1986)**		**성별**	□남/Male □녀/Female

2. **지금 대학교나 어학당/학원**에 다닙니까? Are you currently attending a university or language school?

□ 아니요 / No □ 예 / Yes 학교 이름/School Name:

_____

3. **연락처**가 무엇입니까? What is your contact information? (결과를 받고 싶을 경우/If you want to see your results)

이메일:_____

카카오 ID:_____

3. 지금 **몇 급 수업**을 듣습니까? What level of Korean class are you taking now?

□ 1 급 □ 5 급
□ 2 급 □ 6 급
□ 3 급 □ 지금 대학교나 대학원 수업을 듣는다
□ 4 급 □ 다른 경우/Other:

_____

4. 이번에 **언제** 한국에 들어왔습니까? (년/월/일)/When did you arrive in Korea? (YYYY/MM/DD): _____

5. **최종 학력**을 표시하세요: (가장 최근의 교육 수준을 표시하세요.)

Please check your highest education level:

□ 고등학교 시작 / Less than high school
□ 고등학교 졸업 / High school graduate
□ 직업교육훈련 / Vocational training
□ 대학교 시작 / Some college
□ 학사 (대학교 졸업) / 3-4 year degree (B.A., B.S., etc.)
□ 대학원 시작 / Some graduate school
□ 석사 졸업/ Master's degree
□ 박사 등 / Ph.D./M.D./J.D.
□ 다른 학위/Other:

_____

언어 배경 – Language Background

1. 아는 언어 중에서 잘하는 언어를 순서대로 다 쓰세요:
Please list all the languages you know in order of **dominance** (i.e., strength):

1 위:	**2 위:**	**3 위:**	**4 위:**	**5 위:**
기타/Others:				

2. 위의 **언어**들을 요즘 얼마나 많이 **사용하는지 퍼센트**(%)로 표시하세요 (총 퍼센트(%)가 100 이어야 합니다):
Please list what percentage of the time you are currently and on average exposed to each language you listed above. (Your percentages should add up to 100):

언어/language:					
사용%:					

3. **한국어**는 내가 ____번째로 배운 언어입니다. / Korean is my ____th language.

□ 1 □ 2 □ 3 □ 4 □ 5 (또는 5 이상/ 5 or later)

4. 아래에 있는 것을 했을 때 **몇 살**이었습니까? How old were you when you...

한국어 배우기를 시작했을 때/began learning Korean?	**처음 한국어로 이야기할 수 있을 때/became conversational in Korean?**	**한국어로 읽기 시작했을 때/began reading in Korean?**	**처음 한국어를 유창하게 할 수 있을 때/became fluent in reading Korean?**

5. 다음을 경험한 기간이 **몇 년 동안 / 몇 개월 동안**이었는지 쓰세요.
Please list the total number of years and months you spent in each Korean language environment:

	몇 년 / Years	몇 개월 / Months
한국에서 살기 **Living in South Korea**		
한국어를 말하는 가족과 살기 **Living with a family that speaks Korean**		
한국안에 있는 학교나 학원에서 한국어를 공부하기 **Studying Korean in a school in South Korea**		
다른 나라에 있는 학교나 학원에서 한국어를 공부하기 **Studying Korean in a school in another country**		
한국 여행하기 **Vacation in South Korea**		

6. 당신의 한국어 말하기, 듣기, 쓰기, 읽기 **능력**을 0 부터 10 중에서 표시하세요:
On a scale from 0 to 10, please select your level of proficiency in speaking, listening, writing, and reading Korean:

0	1	2	3	4	5	6	7	8	9	10
못함 none	매우 낮음 very low	낮음 low	보통 fair	거의 충분 slightly less than adequate	충분 adequate	충분보다 조금 높음 slightly more than adequate	잘함 good	매우 잘함 very good	훌륭함 excellent	완벽 perfect

능력/Your Skills (숫자를 쓰세요/write a number):

말하기/Speaking:		**듣기/Listening:**		**쓰기/Writing:**		**읽기/Reading:**	

7. **한국어 능력 시험**을 본 적이 있습니까? 시험 이름, 날짜와 점수를 쓰세요.
Have you ever taken a Korean proficiency test? Please write the test name, date, and score.

시험 이름/Test Name	날짜(년/월) / Date (Year/Month)	점수 / Score

8. 아래 있는 것들이 당신의 **한국어 배우기에 얼마나 기여했는지** 0 부터 10 중에서 표시하세요:
On a scale from 0 to 10, please select how much the following factors contributed to your Korean learning:

0	1	2	3	4	5	6	7	8	9	10
전혀 도움이 안됨/not at all	아주 조금 도움됨 minimally				보통 moderately					가장 많이 도움됨/most importantly

친구와 소통/Interacting with friends:		**혼자서 공부/Self-study:**	
가족과 소통/Interacting with family:		**TV 나 영화 보기/Watching TV or movies:**	
읽기/Reading:		**음악 듣기/Listening to music:**	

9. 현재 아래의 상황에서 **얼마나 한국어를 사용하는지** 표시하세요:
On a scale from 0 to 10, select how much you are currently exposed to Korean in the following contexts:

0	1	2	3	4	5	6	7	8	9	10
전혀 사용 안함 none	거의 사용 안함 almost never				50%정도 half the time					항상 always

예: 읽을 때 50%정도 한국어로 읽어요 (즉 남은 50%는 내 모국어로, 아니면 다른 언어로 읽어요).
Example: When I am reading, it is in Korean about half the time (5). (e.g., the other half is English).

친구와 소통/Interacting with friends:		**혼자서 공부/Self-study:**	

가족과 소통/Interacting with family:		**TV 나 영화 보기/Watching TV or movies:**	
읽기/Reading:		**음악 듣기/Listening to music:**	

10. 아래에 있는 것들이 한국어 배우기에 얼마나 **동기**를 주는지 0 부터 10 중에서 표시하세요:
On a scale from 0 to 10, please select your how much each of the following motivate you to learn Korean:

0	1	2	3	4	5	6	7	8	9	10
전혀 도움이 안됨/not at all	아주 조금 도움됨 minimally				보통 moderately					가장 많이 도움됨/most importantly

취직 Getting a job:		**돈 더 벌기/Earning more money:**	
대학교 입학, 다른 교육 Going to university or other training:		**가족과 친구에게 감명주기 Impressing friends and family:**	
한국어 하는 가족/Korean-speaking family:		**한국어 하는 부부나 애인/Korean-speaking spouse or romantic partner:**	
한국인 친구 사귀기/Friendship with Koreans:		**한국문화/Korean culture:**	

APPENDIX F

Pronunciation Self-Assessment

발음 자기 평가 – **Pronunciation Self-Assessment**

1 부: 발음 전체 인상/Part 1: Your Overall Impressions

당신의 전반적인 한국어 발음 능력을 생각해 보세요. 무엇을 말하는지 (단어, 문법)가 아니라 당신이 어떻게 말하는지 (소리, 억양/인토네이션, 율동/리듬)에 집중하세요. 또한, 다른 사람이 당신의 한국어 말하기 어떻게 반응하는지를 생각해 보세요.

Think about your general pronunciation ability in Korean. Focus on how you speak (sounds, intonation, and rhythm) rather than what you say (vocabulary, grammar). Also, think about how others react to your Korean speaking.

1. **이해 난이도**: 다른 사람들이 당신의 말을 얼마나 쉽게 이해합니까?

Comprehensibility: How easily are you understood by others? Your Korean speaking is...

아주 이해하기 어려움								아주 이해하기 쉬움
1	2	3	4	5	6	7	8	9
Extremely hard to understand								Extremely easy to understand

2. **(외국) 억양**: 당신이 한국어를 말할 때 한국사람처럼 들립니까?

Accent: Do you sound like a South Korean when you speak Korean? You have...

외국 억양이 매우 심함								외국 억양이 거의 없음
1	2	3	4	5	6	7	8	9
A very strong foreign accent								Very little accent

3. **만족감**: 당신의 한국어 발음에 얼마나 만족합니까?

Satisfaction: To what extent are you happy with the way you pronounce Korean? You are...

하나도 안 만족함								완전히 만족함
1	2	3	4	5	6	7	8	9
Not happy at all								Completely happy

4. **가치**: 당신에게 한국어 발음은 얼마나 중요합니까?

Value: How important is Korean pronunciation to you? It is...

하나도 안 중요함								매우 중요함
1	2	3	4	5	6	7	8	9
Not important at all								Extremely important

2 부: 한국어의 소리 Part 2: Individual Sounds

이 부분에서는 한국어의 소리에 (예: ㄱ, ㄴ, ㄷ) 대해서 생각할 것입니다. 한국어에는 28 개의 소리가 있습니다: 자음 19 개, 모음 7 개와 반모음 2 개입니다. (반모음은 '와'와 '요'의 첫 소리입니다).

For this part of the self-assessment, you will need to think about the individual sounds of Korean (ㄱ, ㄴ, ㄷ, etc.). Korean has 28 unique sounds: 19 consonants, 7 vowels, and 2 glides (*glides* are the first sound in 와 and 요).

각 소리에 대해서 (a) 그 소리를 잘 **듣는 것과** (b) 그 소리를 잘 **발음하는 것이** 얼마나 어려운지를 생각하세요. (1 = 항상 어렵다, 7 = 전혀 어렵지 않다)

For each individual sound, think about how difficult (1 = always difficult, 7 = never difficult) it is for you to (a) clearly **pronounce** the sound and (b) clearly **hear** the sound.

그 소리가 얼마나 어려운지를 잘 모르겠다면 '모르겠다'를 선택해도 됩니다.

If you are really not sure at all about a sound, you can select "Not sure".

한국어 자음 1 / Korean Consonants 1:

소리 Sound	영역 Mode	←항상 어렵다 ←Always Difficult						전혀 어렵지 않다→ Never Difficult→	모르겠다 Not sure
ㄱ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㄲ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅋ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㄷ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㄸ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅌ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?

한국어 자음 2 / Korean Consonants 2:

소리 Sound	영역 Mode	←항상 어렵다 ←Always Difficult						전혀 어렵지 않다→ Never Difficult→	모르겠다 Not sure
ㅂ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅃ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅍ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅈ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅉ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅊ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅅ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅆ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅁ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㄴ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅇ (예: 방)	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㄹ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅎ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?

한국어 모음 / Korean Vowels:

소리 Sound	영역 Mode	←항상 어렵다 ←Always Difficult						전혀 어렵지 않다→ Never Difficult→	모르겠다 Not sure
ㅏ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅣ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅋ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅗ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅜ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅡ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
ㅔ, ㅐ	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?

한국어 반모음 / Korean Glides:

소리 Sound	영역 Mode	←항상 어렵다 ←Always Difficult						전혀 어렵지 않다→ Never Difficult→	모르겠다 Not sure
/w/ (예: 와, 워, 외, 왜, 위)	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?
/j/ (예: 야, 요, 여, 유, 예, 얘, 의)	발음 Pron.	1	2	3	4	5	6	7	?
	듣기 Hearing	1	2	3	4	5	6	7	?

APPENDIX G

Independent Speaking Task

한국어 말하기

Speaking Task

설명: 질문을 잘 읽으세요. 그리고 그 질문에 대한 의견을 말하세요.
15 초동안 의견을 생각한 다음에 1 분 동안 말하세요.

Directions: Read the question below. You will give your opinion on the question. You will have 15 seconds to think about your opinion, and then you will have 1 minute to speak.

질문:

어떤 사람들은 작은 도시에 사는 것을 좋아해요. 또 어떤 사람들은 큰 도시에 사는 것을 좋아해요.

당신은 작은 도시와 큰 도시중에서 어디에 살고 싶어요? 왜요?

Question:

Some people prefer to live in a small town. Others prefer to live in a big city.

Which place would you prefer to live in and why?

APPENDIX H

Korean EIT Directions and Practice Items

In this task, you'll be asked to repeat some sentences in Korean and some sentences in English. Please follow the instructions carefully. Please do not take any notes during this exercise. Now let's begin.
이 시험에서는 한국어 문장을 듣고 그 문장을 따라할 거예요. 설명을 잘 듣고 따라해 보세요.

You are going to hear several sentences in Korean. After each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, DON'T START REPEATING THE SENTENCE UNTIL YOU HEAR THE TONE SOUND {TONE}. Now let's begin.
지금부터 한국어 연습 문장 5 개를 들을 거예요. 각 문장을 들은 다음에 삐소리가 나면 <삐소리> 그 문장을 따라 말해보세요. 문장을 완벽하게 따라하는 것이 어려워도 열심히 해 보세요. 삐소리 후 문장을 말할 시간은 충분히 있을 거예요. 하지만 삐소리를 듣기 전에 말하면 안 돼요. 자, 지금 연습 문장을 해 볼까요?

Note: response time given is roughly ~0.6s per syllable

나는 꽃이 좋아.
(6 syllables) 2.0s pause, 0.5s tone, 3.9s response time
[translation: I like flowers]

저는 편지를 써요.
(7 syllables) 2.0s pause, 0.5s tone, 4.1 s response time
[translation: I write a letter]

저는 큰 차가 필요해요.
(9 syllables) 2.0s pause, 0.5s tone, 5.6 s response time
[translation: I need a big car]

비가 와서 밖에 안 나가요.
(10 syllables) 2.0s pause, 0.5s tone, 6 s response time
[translation: As it is raining, I don't go out]

그 여자 아이는 축구를 좋아해.
(12 syllables) 2.0s pause, 0.5s tone, 7.2 s response time
[translation: That girl likes soccer]

That was the last practice sentence.
그건 마지막 연습 문장이었어요.

Now you will hear more Korean sentences. Once again, after each sentence, there will be a short pause followed by a tone sound <tone>. Your task is to try to repeat exactly what you hear in Korean. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, don't start repeating the sentence until you hear the tone sound <tone>.
지금부터 문장들을 더 들을 거예요. 각 문장을 들은 다음에 삐소리가 나면 <삐소리> 그 문장을 따라 말해보세요. 문장을 완벽하게 따라하는 것이 어려워도 열심히 해 보세요. 삐소리 후 말할 시간이 충분히 있을 거예요. 하지만 삐소리를 듣기 전에 말하면 안 돼요.

Do you have any questions?
질문이 있어요?

Now, let's begin.
그럼, 시작합시다!

APPENDIX I

Interview Protocols

Interview 1 - Learner

Orientation; Reflection on Own Pronunciation

1. Ask participants to reflect on their Korean Pronunciation / 당신의 발음을 반영해주세요
    a. "발음에 대해 어떻게 생각하세요?" / What do you think about your pronunciation?
    b. "한국어 발음의 가장 어려운 점은 뭐예요?" / What are the most difficult aspects?
2. Present participants with their self-assessments to assist with reflection.

Interpreting Results

1. Present participants with their KPD results. Ask them to read the results and share any thoughts that arise.
    a. "발음 진단 검사 결과를 보면서 의견이나 질문이 있으면 이야기해 주세요"
2. After initial reactions, follow up with the following questions:
    a. "결과에 대해 어떤 생각이 있어요?" / What do you think about the test results?
    b. "결과는 당신의 생각과 비슷해요?" / Are the results similar to your own impressions? (Can go to the self-evaluation results here)
    c. "놀라운 결과가 있어요? 왜요?" / Are there any surprising results? Why?

Learning Activity

1. Participants prompted to discuss study/practice habits:
2. "보통 발음을 어떻게 공부하거나 연습해요?" / How do you usually study or practice pronunciation?
3. "다른 발음 공부나 연습을 시도해 봤어요?"/ Have you tried any other methods?
4. "결과를 본 후에 다른 공부나 연습 방법을 시도 할 것 같아요?" / After seeing these results, do you think you'll try anything different?

Progress

1. Participants prompted to discuss pronunciation development:
2. "한국어 발음 배우기에 대한 경험을 이야기해주세요"
    a. "어련운 게 뭐예요?" "쉬운 것은? / What has been difficult for you? Easy?
    b. "한국어 발음을 배웠을 때, 어떤 단계나 과정을 통해서 경험했어요? / What process or steps did you experience when learning Korean pronunciation?
3. "한국어 발음에 대한 목표가 있어요?" / Do you have any Korean pronunciation goals?
    a. "그 목표를 다 이루웠어요? 아니면 아직 멀었어요?" / Are you far from achieving that goal? Close?
    b. (목표가 없는 경우) "왜 발음에 대한 목표가 없나요?" / Why do you not have any pronunciation goals?

Follow-up: Interview 2 – Learner

KPD Results

1. "한국어 발음 시험 결과가 기억나요? 뭐라고 써 있었어요?" Do you remember what your pronunciation test results were? Do you remember what it said?
    a. Can review the results if participant can't recall very well
2. "결과지에 대해 얼마나 생각했어요?" How much have you thought about your test results?

Learning Activity

1. Participants prompted to discuss study/practice habits:
    a. "처음 인터뷰한 후에 발음을 공부하거나 연습했어요? 어떻게 했어요?" / Since we last met, have you studied or practiced pronunciation? If so, how?
    b. "결과를 본 후에 다른 방법을 해 봤어요?" / After seeing the results, have you tried anything different?

Progress

1. Participants prompted to discuss pronunciation development:
    a. "최근에 __씨의 발음에 달라진 게 있어요? 없어요? 설명해주세요." / Recently, have you noticed any changes in your pronunciation? None? Please explain.
2. "한국어 발음에 대한 목표가 있어요?" / Do you have any Korean pronunciation goals?
    c. "그 목표를 다 이루웠어요? 아니면 아직 멀었어요?" / Are you far from achieving that goal? Close?

Now administer **Independent Speaking** AND **KPD** again

Interview 1 – Teacher

Pronunciation Teaching – 발음 교육

1. Teachers prompted to discuss pronunciation teaching practices:
    a. "보통 어떻게 발음을 가르쳐요?" / How do you usually teach pronunciation?
    b. "다른 방법으로 가르쳐 봤어요?"/ Have you tried any other methods?
    c. "학생들 발음을 가르칠 때 어떤 목표나 원칙이 있어요?" / For your students, do you have any pronunciation goals or principles?

Teacher's Observations – 교사의 착안과 평가

2. Show teachers the list of students who are participating in the study and have completed the KPD. Ask teacher to describe each student's pronunciation in turn. The student's Independent Speaking can be played if necessary, to jog the teacher's memory.
    a. "전반적으로, <이 학생>의 발음이 어때요?" / Overall, how is this student's pronunciation?
    b. "이 학생이 말을 대부분 쉽게 이해 할 수 있어요? 이해하기 얼마나 어려워요?" / Are you able to easily understand what this student says? How difficult is he/she to understand?
    c. "더 구체적으로, <이 학생>의 발음은 어떤 어려운 점이 있어요?" / More specifically, what are the difficulties this student has in pronunciation?
    d. "어느 소리/음소를 특별히 어려워해요?" / Which sounds/phonemes are especially difficult?
    e. "다른 어려운 점이 있어요?" / Are there any other challenging features?

Interpreting Results – 결과 이해하기

3. Present teacher with their students' KPD results. Ask them to read the results and share any thoughts that arise.
    a. "발음 진단 검사 결과를 보면서 의견이나 질문이 있으면 이야기해 주세요"
4. After initial reactions, follow up with the following questions about the test results in general:
    a. "결과에 대해 어떻게 생각해요?" / What do you think about the test results?
    b. "결과는 선생님 생각과 비슷해요?" / Are the results similar to your own impressions?
    c. "놀라운 결과가 있어요? 왜요?" / Are there any surprising results? Why?

Using Results – 결과 응용하기

5. Ask the teacher how he/she might address the student's pronunciation weaknesses.
    a. "결과를 본 후에 이 학생들의 취약점을 어떻게 고쳐주고 싶어요?" / After seeing these results, how would you address the students' weaknesses?
    b. "이 결과지가 유용할 것 같아요? 왜요?" / Do these score reports seem useful? Why?
    c. "이 학생들의 발음에 대해서 알고 싶은 다른 것이 있어요?" / Is there anything else you'd like to know about these students' pronunciation?

APPENDIX J

Item Statistics

Table J1

KPD Production Item Statistics

Item	IF	ID	Rasch Measure	Rasch S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
T1_01-1	0.84	0.34	1.46	0.20	0.93	-0.55	0.75	-1.21
T1_01-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_01-3	0.96	0.15	-0.23	0.39	0.99	0.10	3.64	3.19
T1_02-1	0.91	0.31	0.74	0.26	0.95	-0.21	0.68	-1.02
T1_02-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_02-3	0.98	0.10	-0.81	0.51	1.01	0.17	0.70	-0.22
T1_02-4	0.99	0.06	-2.22	1.01	1.01	0.34	0.63	0.21
T1_02-5	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_03-1	0.90	0.14	0.93	0.24	1.02	0.19	0.97	0.01
T1_03-2	0.82	0.04	1.65	0.19	1.11	1.03	1.22	1.14
T1_03-3	0.99	0.17	-2.22	1.01	0.99	0.32	0.32	-0.15
T1_03-4	0.94	0.06	0.35	0.30	1.03	0.22	1.06	0.28
T1_03-5	0.86	0.22	1.29	0.21	1.00	0.04	0.93	-0.22
T1_03-6	0.98	0.02	-0.81	0.51	1.02	0.20	1.37	0.70
T1_04-1	0.99	0.11	-1.52	0.72	1.00	0.23	0.58	-0.14
T1_04-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_04-3	0.99	0.03	-1.52	0.72	1.01	0.25	0.88	0.23
T1_04-4	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_05-1	0.78	0.11	1.90	0.18	1.10	1.06	1.16	0.98
T1_05-2	0.99	0.02	-2.22	1.01	1.01	0.34	0.87	0.40
T1_05-3	0.75	0.05	2.08	0.17	1.14	1.62	1.16	1.12
T1_05-4	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_06-1	0.92	0.11	0.60	0.27	1.03	0.19	1.00	0.11
T1_06-2	0.98	0.07	-1.10	0.59	1.01	0.20	1.07	0.36
T1_06-3	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_07-1	0.99	0.01	-1.52	0.72	1.02	0.25	0.93	0.28
T1_07-2	0.88	0.08	1.15	0.22	1.06	0.42	1.64	2.17
T1_07-3	0.98	0.08	-1.10	0.59	1.00	0.19	1.31	0.61
T1_07-4	0.99	0.01	-1.52	0.72	1.01	0.25	1.69	0.90
T1_08-1	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_08-2	0.98	0.02	-1.10	0.59	1.02	0.22	1.40	0.70
T1_08-3	0.87	0.28	1.24	0.22	0.97	-0.18	0.85	-0.55
T1_08-4	0.94	0.19	0.26	0.32	0.99	0.03	0.79	-0.41
T1_08-5	0.97	-0.05	-0.58	0.46	1.04	0.24	1.49	0.87
T1_09-1	0.95	0.18	0.04	0.35	0.98	0.04	0.67	-0.66
T1_09-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_09-3	0.95	0.29	0.04	0.35	0.95	-0.06	0.59	-0.89
T1_10-1	0.97	0.03	-0.39	0.42	1.02	0.19	1.21	0.54
T1_10-2	0.99	0.06	-2.22	1.01	1.01	0.34	0.63	0.21
T1_11-1	0.85	0.11	1.38	0.21	1.07	0.55	1.24	1.08
T1_11-2	0.99	0.09	-2.22	1.01	1.00	0.33	0.51	0.08

Table J1 (cont'd)

T1_11-3	0.97	0.08	-0.39	0.42	1.01	0.16	1.07	0.32
T1_11-4	0.98	0.10	-0.81	0.51	1.00	0.17	0.78	-0.08
T1_11-5	1.00	NA	-3.42	1.83	1.00	0.00	1.00	0.00
T1_12-1	0.92	0.15	0.67	0.27	1.02	0.14	0.88	-0.25
T1_12-2	0.98	0.11	-1.10	0.59	1.00	0.18	0.77	-0.02
T1_13-1	0.94	0.06	0.35	0.30	1.03	0.22	1.12	0.41
T1_13-2	0.84	0.05	1.46	0.20	1.09	0.76	1.52	2.17
T1_14-1	0.75	0.28	2.11	0.17	1.00	-0.03	1.01	0.12
T1_14-2	0.98	0.26	-1.10	0.59	0.97	0.13	0.38	-0.71
T1_14-3	0.97	0.26	-0.39	0.42	0.96	0.03	0.49	-0.89
T1_14-4	0.89	0.00	1.04	0.23	1.09	0.59	1.53	1.77
T1_14-5	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_14-6	0.94	0.20	0.35	0.30	0.98	0.02	0.78	-0.48
T1_15-1	0.97	0.04	-0.58	0.46	1.01	0.18	2.21	1.62
T1_15-2	0.93	0.09	0.52	0.28	1.02	0.18	0.90	-0.16
T1_15-3	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_15-4	0.99	0.02	-1.52	0.72	1.01	0.25	1.95	1.06
T1_15-5	0.96	0.07	-0.23	0.39	1.01	0.15	1.77	1.35
T1_15-6	0.94	0.19	0.26	0.32	0.99	0.05	0.70	-0.66
T1_15-7	0.91	0.32	0.81	0.25	0.94	-0.25	0.67	-1.11
T1_16-1_p	0.91	0.08	0.74	0.26	1.05	0.30	1.02	0.18
T1_16-2_p	0.92	0.17	0.60	0.27	1.00	0.08	0.81	-0.47
T1_16-3_p	0.85	0.05	1.42	0.21	1.09	0.72	1.28	1.25
T1_16-4_p	0.98	0.24	-0.81	0.51	0.97	0.10	0.45	-0.72
T1_17-1_p	0.96	0.05	-0.23	0.39	1.01	0.14	1.29	0.67
T1_17-2_p	0.98	0.04	-1.10	0.59	1.02	0.21	0.86	0.10
T1_17-3_p	0.85	0.21	1.42	0.21	1.01	0.10	0.90	-0.40
T1_18-1	0.99	0.03	-2.22	1.01	1.01	0.34	0.76	0.32
T1_18-2	0.98	-0.01	-1.10	0.59	1.02	0.23	1.40	0.70
T1_18-3	0.98	0.17	-1.10	0.59	0.99	0.17	0.51	-0.43
T1_18-4	0.99	0.06	-2.22	1.01	1.01	0.34	0.63	0.21
T1_18-5	0.93	0.20	0.44	0.29	0.99	0.06	0.90	-0.14
T1_18-6	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_18-7	0.98	0.19	-1.10	0.59	0.98	0.16	0.48	-0.50
T1_18-8	0.88	0.14	1.09	0.23	1.03	0.25	1.13	0.57
T1_19-1	0.98	0.10	-1.10	0.59	1.00	0.19	0.82	0.05
T1_19-2	0.99	0.18	-1.52	0.72	0.99	0.21	0.43	-0.36
T1_19-3	0.98	-0.09	-0.81	0.51	1.04	0.24	1.93	1.25
T1_19-4	0.99	0.09	-2.22	1.01	1.00	0.33	0.51	0.08
T1_20-1	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_20-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_20-3	0.95	0.12	0.15	0.33	1.01	0.14	0.87	-0.16
T1_20-4	0.92	0.10	0.67	0.27	1.02	0.19	2.32	3.02
T1_20-5	0.95	0.08	0.04	0.35	1.02	0.18	1.14	0.44
T1_20-6	0.89	0.25	0.99	0.24	0.98	-0.07	0.76	-0.85

Table J1 (cont'd)

T1_21-1	0.85	0.32	1.42	0.21	0.95	-0.37	0.78	-1.00
T1_21-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_21-3	0.96	0.20	-0.08	0.37	0.98	0.04	0.71	-0.48
T1_21-4	0.99	0.04	-1.52	0.72	1.01	0.24	1.46	0.73
T1_21-5	0.99	0.12	-2.22	1.01	1.00	0.33	0.42	-0.02
T1_22-1	0.99	-0.02	-2.22	1.01	1.02	0.35	1.29	0.67
T1_22-2	0.99	-0.02	-2.22	1.01	1.02	0.35	1.29	0.67
T1_22-3	0.90	0.16	0.93	0.24	1.02	0.15	0.94	-0.13
T1_22-4	0.93	0.21	0.44	0.29	0.98	-0.01	2.38	2.79
T1_22-5	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_23-1	0.90	0.10	0.87	0.25	1.04	0.27	1.05	0.27
T1_23-2	0.99	0.10	-1.52	0.72	1.00	0.24	0.62	-0.08
T1_23-3	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_24-1	0.98	0.04	-1.10	0.59	1.02	0.22	1.26	0.56
T1_24-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_24-3	0.99	-0.05	-2.22	1.01	1.02	0.35	1.73	0.90
T1_24-4	0.99	0.17	-2.22	1.01	0.99	0.32	0.32	-0.15
T1_24-5	0.98	0.15	-1.10	0.59	0.99	0.17	0.56	-0.35
T1_25-1_p	0.92	0.21	0.67	0.27	0.98	-0.02	1.02	0.16
T1_25-2_p	0.96	0.17	-0.23	0.39	0.99	0.08	0.85	-0.11
T1_26-1	0.96	0.13	-0.23	0.39	1.00	0.12	0.75	-0.33
T1_26-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_26-3	0.97	0.13	-0.58	0.46	1.00	0.14	0.74	-0.24
T1_26-4	0.99	0.06	-1.52	0.72	1.01	0.24	0.88	0.23
T1_26-5	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_27-1	0.96	0.12	-0.08	0.37	1.00	0.12	1.18	0.51
T1_27-2	0.99	0.08	-1.52	0.72	1.00	0.24	0.80	0.14
T1_27-3	0.98	0.02	-0.81	0.51	1.02	0.20	1.25	0.56
T1_27-4	0.99	0.13	-1.52	0.72	1.00	0.23	0.52	-0.22
T1_27-5	0.99	0.12	-2.22	1.01	1.00	0.33	0.44	0.00
T1_27-6	0.99	0.07	-2.22	1.01	1.01	0.34	0.60	0.17
T1_28-1	0.99	-0.03	-2.22	1.01	1.02	0.35	1.41	0.74
T1_28-2	0.97	0.24	-0.39	0.42	0.97	0.04	0.71	-0.36
T1_29-1_n	0.97	0.13	-0.58	0.46	1.00	0.13	0.92	0.09
T1_29-2_n	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_29-3_n	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_29-4_n	0.95	0.15	0.04	0.35	1.00	0.11	1.01	0.17
T1_29-5_n	0.85	0.29	1.42	0.21	0.96	-0.24	0.81	-0.86
T1_30-1	0.95	0.03	0.04	0.35	1.04	0.23	1.16	0.49
T1_30-2	0.98	0.09	-1.10	0.59	1.01	0.20	0.72	-0.10
T1_30-3	0.49	0.04	3.39	0.15	1.20	3.66	1.23	3.08
T1_30-4	0.95	-0.01	0.04	0.35	1.05	0.25	1.89	1.68
T1_30-5	0.81	0.31	1.69	0.19	0.96	-0.35	0.83	-0.88
T1_31-1	0.94	0.13	0.35	0.30	1.01	0.11	1.04	0.24
T1_31-2	0.99	0.04	-2.22	1.01	1.01	0.34	0.72	0.28

Table J1 (cont'd)

T1_31-3	0.98	0.17	-0.81	0.51	0.99	0.14	0.55	-0.51
T1_32-1	0.99	0.05	-1.52	0.72	1.01	0.25	0.78	0.12
T1_32-2	0.99	0.21	-1.52	0.72	0.98	0.20	0.37	-0.48
T1_32-3	0.99	0.13	-1.52	0.72	1.00	0.23	0.53	-0.21
T1_32-4	0.99	-0.01	-1.52	0.72	1.02	0.26	1.77	0.95
T1_32-5	0.83	0.02	1.58	0.20	1.12	1.00	1.31	1.50
T1_32-7	0.80	0.29	1.76	0.19	0.97	-0.23	1.02	0.17
T1_33-1	0.80	0.28	1.76	0.19	0.97	-0.27	0.90	-0.53
T1_33-2	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_33-3	0.98	-0.02	-1.10	0.59	1.02	0.23	1.69	0.96
T1_33-4	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_33-5	0.99	-0.03	-2.22	1.01	1.02	0.35	1.41	0.74
T1_33-6	0.97	0.11	-0.39	0.42	1.00	0.13	1.70	1.19
T1_33-7	0.99	0.00	-2.22	1.01	1.01	0.34	1.01	0.50
T1_33-8	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_33-9	0.86	0.28	1.29	0.21	0.97	-0.19	0.87	-0.48
T1_34-1	0.76	0.25	2.05	0.17	1.01	0.09	0.97	-0.16
T1_34-2	0.98	0.05	-1.10	0.59	1.01	0.21	0.95	0.22
T1_34-3	0.72	0.40	2.28	0.17	0.91	-1.32	0.87	-0.99
T1_35-1	0.98	0.07	-1.10	0.59	1.00	0.18	0.85	0.09
T1_35-2	0.91	0.22	0.74	0.26	0.98	-0.02	0.91	-0.18
T1_35-3	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T1_35-4	0.86	0.19	1.33	0.21	1.02	0.15	1.03	0.21
T2_01	0.88	0.10	1.09	0.23	1.04	0.28	1.07	0.33
T2_02	0.94	0.13	0.26	0.32	1.01	0.14	0.81	-0.35
T2_03	0.96	0.30	-0.23	0.39	0.95	-0.03	0.48	-1.03
T2_04	0.80	0.42	1.76	0.19	0.89	-1.05	0.74	-1.54
T2_05	0.70	0.28	2.37	0.16	0.98	-0.21	0.96	-0.30
T2_06	0.83	0.34	1.58	0.20	0.94	-0.50	0.77	-1.20
T2_07	0.83	0.26	1.54	0.20	0.99	-0.08	0.92	-0.35
T2_08	0.76	0.16	2.05	0.17	1.05	0.60	1.11	0.74
T2_09	0.85	0.14	1.42	0.21	1.03	0.28	1.04	0.26
T2_10	0.83	0.30	1.58	0.20	0.96	-0.28	0.85	-0.74
T2_11	0.85	0.43	1.38	0.21	0.88	-0.89	0.66	-1.67
T2_12	0.63	0.25	2.73	0.16	1.02	0.38	0.99	-0.10
T2_13	0.83	0.32	1.54	0.20	0.95	-0.38	0.73	-1.39
T2_14	0.90	0.35	0.93	0.24	0.92	-0.40	0.62	-1.40
T2_15	0.96	0.07	-0.08	0.37	1.02	0.15	1.40	0.88
T2_16	0.95	0.14	0.15	0.33	1.01	0.13	0.76	-0.46
T2_17	0.81	0.16	1.73	0.19	1.04	0.38	1.10	0.58
T2_18	0.76	0.25	2.05	0.17	1.01	0.10	0.96	-0.23
T2_19	0.91	0.29	0.74	0.26	0.96	-0.15	0.68	-1.03
T2_20	0.91	0.23	0.81	0.25	0.97	-0.10	0.90	-0.24
T2_21	0.97	0.16	-0.39	0.42	0.99	0.11	0.65	-0.50
T2_22	0.78	0.37	1.93	0.18	0.92	-0.92	0.75	-1.62

Table J1 (cont'd)

T2_23	0.48	0.34	3.42	0.15	0.94	-1.14	0.92	-1.10
T2_24	0.76	0.26	2.02	0.18	0.99	-0.14	0.90	-0.60
T2_25	0.81	0.25	1.73	0.19	0.99	-0.03	0.98	-0.06
T2_26	0.92	0.06	0.60	0.27	1.04	0.27	1.13	0.48
T2_27	0.88	0.04	1.09	0.23	1.08	0.55	1.15	0.62
T2_28	0.98	0.12	-0.81	0.51	1.00	0.15	0.95	0.17
T2_29	0.98	0.19	-1.10	0.59	0.98	0.15	0.53	-0.41
T2_30	0.61	0.29	2.81	0.16	1.01	0.19	0.99	-0.07
T2_31	0.75	0.16	2.11	0.17	1.06	0.77	0.98	-0.06
T2_32	0.51	0.23	3.30	0.15	1.04	0.73	1.02	0.32
T2_33	0.99	0.18	-2.22	1.01	0.99	0.32	0.31	-0.17
T2_34	0.99	0.18	-2.22	1.01	0.99	0.32	0.31	-0.17
T2_35	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T2_36	0.99	-0.05	-2.22	1.01	1.02	0.35	1.73	0.90
T2_37	0.98	0.10	-1.10	0.59	1.00	0.19	0.69	-0.13
T2_38	0.98	0.03	-1.10	0.59	1.02	0.21	1.56	0.85
T2_39	0.99	0.11	-2.22	1.01	1.00	0.33	0.46	0.03
T2_40	0.98	0.15	-1.10	0.59	0.99	0.17	0.66	-0.18
T2_41	0.91	0.29	0.74	0.26	0.95	-0.16	0.70	-0.92
T2_42	0.94	0.30	0.26	0.32	0.94	-0.13	1.08	0.32
T2_43	0.96	0.03	-0.23	0.39	1.03	0.19	1.40	0.85
T2_44	0.97	0.13	-0.39	0.42	1.00	0.12	0.76	-0.25
T2_45	0.98	0.15	-0.81	0.51	0.99	0.14	0.68	-0.27
T2_46	0.91	0.19	0.74	0.26	1.00	0.05	0.92	-0.15
T2_47	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T2_48	0.85	0.04	1.42	0.21	1.09	0.75	2.01	3.65
T2_49	0.96	0.20	-0.08	0.37	0.98	0.04	0.87	-0.10
T2_50	0.69	0.29	2.45	0.16	0.98	-0.29	0.91	-0.78
T2_51	1.00	NA	-3.43	1.83	1.00	0.00	1.00	0.00
T2_52	0.96	0.10	-0.23	0.39	1.00	0.13	1.37	0.80
T2_53	0.98	-0.02	-1.10	0.59	1.01	0.21	0.93	0.20
T2_54	0.77	0.29	1.99	0.18	0.98	-0.25	0.88	-0.74
T2_55	0.76	0.16	2.02	0.18	1.06	0.77	0.99	-0.05
T2_56	0.87	0.20	1.24	0.22	1.01	0.10	0.94	-0.17
T2_57_n	0.80	0.46	1.76	0.19	0.87	-1.33	0.68	-1.99
T2_58	0.98	0.08	-0.81	0.51	1.01	0.18	0.84	0.01
T2_59_n	0.66	0.18	2.58	0.16	1.06	1.08	1.10	0.99
T2_60	0.92	-0.07	0.60	0.27	1.09	0.49	1.63	1.65
T2_61	0.99	0.06	-1.51	0.72	1.01	0.24	1.15	0.48
T2_62	0.88	0.15	1.09	0.23	1.01	0.12	1.27	1.04
T2_63	0.99	-0.05	-2.22	1.01	1.02	0.35	1.73	0.90

Table J2

KPD Perception Item Statistics

Item	IF	ID	Rasch Measure	Rasch S.E.	Infit MS	Infit Z	Outfit MS	Outfit Z
T3_01_s	0.58	0.51	1.80	0.16	0.87	-2.56	0.79	-2.20
T3_02	0.46	0.18	2.39	0.16	1.22	3.47	1.23	2.52
T3_03	0.91	0.32	-0.48	0.26	0.92	-0.37	0.70	-0.74
T3_04	0.59	0.33	1.78	0.16	1.04	0.68	1.05	0.48
T3_05	0.77	0.40	0.79	0.18	0.93	-0.81	0.75	-1.39
T3_06	0.40	0.50	2.66	0.16	0.85	-2.36	0.84	-1.85
T3_07	0.69	0.42	1.24	0.17	0.92	-1.23	0.84	-1.13
T3_08	0.92	0.25	-0.62	0.27	0.96	-0.13	0.75	-0.54
T3_09	0.81	0.34	0.49	0.19	0.96	-0.39	0.76	-1.08
T3_10	0.98	0.29	-2.04	0.51	0.92	-0.02	0.29	-0.90
T3_11	0.73	0.47	1.01	0.17	0.87	-1.75	0.72	-1.85
T3_12	0.59	0.58	1.78	0.16	0.80	-4.00	0.72	-2.94
T3_13	0.33	0.41	3.03	0.17	0.95	-0.64	0.93	-0.60
T3_14	0.75	0.21	0.92	0.17	1.07	0.89	1.33	1.80
T3_15	0.66	0.46	1.42	0.16	0.89	-1.80	0.80	-1.64
T3_16	0.52	0.26	2.09	0.16	1.10	1.84	1.11	1.15
T3_17	0.99	0.13	-3.46	1.00	0.98	0.31	0.31	0.02
T3_18	0.39	0.31	2.71	0.16	1.04	0.61	1.06	0.65
T3_19	0.71	0.47	1.15	0.17	0.87	-1.84	0.77	-1.62
T3_20	0.49	0.53	2.24	0.16	0.84	-3.04	0.79	-2.48
T3_21	0.80	0.43	0.59	0.19	0.88	-1.19	0.74	-1.29
T3_22	0.29	0.51	3.29	0.17	0.82	-2.19	0.79	-1.79
T3_23	0.73	0.37	1.01	0.17	0.95	-0.58	0.85	-0.93
T3_24	0.45	0.37	2.41	0.16	0.99	-0.12	1.02	0.29
T3_25	0.46	0.34	2.39	0.16	1.01	0.19	1.02	0.26
T3_26	0.39	0.57	2.71	0.16	0.82	-2.91	0.74	-3.05
T3_27	0.77	0.36	0.76	0.18	0.95	-0.57	0.85	-0.76
T3_28	0.39	0.30	2.74	0.16	1.07	1.00	1.09	0.99
T3_29	0.18	0.41	4.06	0.20	0.85	-1.18	0.86	-0.67
T3_30	0.53	0.13	2.04	0.16	1.23	3.93	1.34	3.35
T3_31	0.34	0.32	2.98	0.17	1.08	1.08	1.14	1.36
T3_32	0.19	0.17	3.94	0.20	1.14	1.14	1.49	2.40
T3_33_n	0.42	0.38	2.56	0.16	0.99	-0.18	0.95	-0.54
T3_34	0.14	-0.17	4.38	0.22	1.52	3.12	2.48	4.55
T3_35	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T3_36	0.89	0.29	-0.23	0.24	0.96	-0.21	0.75	-0.70
T3_37	0.91	0.35	-0.41	0.25	0.90	-0.49	0.64	-1.02
T3_38	0.98	0.12	-2.34	0.58	0.99	0.16	0.62	-0.11

Table J2 (cont'd)

T3_39_s	0.91	0.24	-0.48	0.26	0.97	-0.08	0.82	-0.39
T3_40	0.75	0.27	0.89	0.18	1.03	0.34	1.00	0.05
T3_41_s	0.99	0.22	-2.76	0.71	0.95	0.16	0.26	-0.48
T3_42	0.94	0.25	-0.87	0.30	0.96	-0.10	0.65	-0.70
T3_43	0.70	0.38	1.18	0.17	0.95	-0.65	0.87	-0.90
T3_44_s	0.82	0.28	0.41	0.20	0.98	-0.17	1.11	0.52
T3_45	0.79	0.17	0.63	0.19	1.08	0.81	1.29	1.38
T3_46	0.94	0.19	-0.87	0.30	0.99	0.04	0.77	-0.37
T3_47	0.91	0.35	-0.41	0.25	0.90	-0.47	0.63	-1.05
T3_48	0.69	0.24	1.26	0.16	1.08	1.14	1.04	0.35
T3_49	0.93	0.23	-0.70	0.28	0.96	-0.10	0.87	-0.18
T3_50_s	0.92	-0.09	-0.55	0.27	1.15	0.76	1.93	1.99
T3_51	0.99	0.13	-2.76	0.71	0.98	0.20	0.48	-0.13
T3_52	0.79	0.39	0.63	0.19	0.91	-0.88	0.88	-0.56
T3_53	0.81	0.40	0.52	0.19	0.90	-0.90	0.73	-1.28
T3_54	0.90	0.24	-0.35	0.25	0.96	-0.17	1.40	1.11
T3_55	0.89	0.16	-0.23	0.24	1.03	0.23	1.08	0.34
T3_56	0.53	0.43	2.04	0.16	0.92	-1.47	0.90	-1.07
T3_57	0.98	0.20	-2.04	0.51	0.95	0.05	0.77	-0.01
T3_58	0.94	0.19	-0.87	0.30	0.98	-0.01	1.01	0.18
T3_59	0.65	0.25	1.45	0.16	1.09	1.39	1.13	1.06
T3_60	0.54	0.25	2.02	0.16	1.12	2.11	1.14	1.51
T3_61	0.42	0.35	2.58	0.16	1.03	0.43	1.04	0.51
T3_62	0.91	0.20	-0.48	0.26	1.00	0.07	0.89	-0.16
T3_63	0.78	0.32	0.70	0.18	0.97	-0.27	0.98	-0.03
T3_64	0.23	0.30	3.68	0.19	1.06	0.61	1.09	0.61
T3_65	0.74	0.07	0.95	0.17	1.20	2.34	1.41	2.19
T3_66	0.62	0.32	1.63	0.16	1.03	0.55	0.99	-0.04
T3_67	0.95	0.25	-1.19	0.35	0.95	-0.07	0.58	-0.71
T3_68_s	0.99	0.19	-2.76	0.71	0.96	0.17	0.34	-0.33
T3_69	0.67	0.41	1.37	0.16	0.93	-1.04	0.91	-0.63
T3_70	0.98	-0.05	-2.34	0.58	1.03	0.23	1.83	1.00
T3_71_s	0.77	0.17	0.79	0.18	1.09	1.01	1.49	2.34
T3_72	0.83	0.25	0.33	0.20	1.00	0.03	1.04	0.26
T4_01	0.46	0.39	2.36	0.16	0.99	-0.17	0.97	-0.31
T4_02	0.89	0.25	-0.23	0.24	0.99	0.00	0.74	-0.74
T4_03	0.99	0.10	-2.76	0.71	0.99	0.22	0.59	0.01
T4_04	0.78	0.36	0.70	0.18	0.94	-0.57	0.88	-0.59
T4_05	0.82	-0.02	0.41	0.20	1.20	1.70	1.85	2.99
T4_06	0.97	0.19	-1.62	0.42	0.96	0.03	0.69	-0.27
T4_07	0.94	0.21	-0.87	0.30	0.98	-0.01	0.75	-0.43

Table J2 (cont'd)

T4_08	0.77	0.37	0.79	0.18	0.94	-0.62	0.83	-0.88
T4_09	0.99	0.02	-3.46	1.00	1.00	0.33	0.95	0.55
T4_10	0.99	0.07	-2.76	0.71	1.00	0.22	0.79	0.22
T4_11	0.95	0.20	-1.07	0.33	0.97	-0.03	0.85	-0.13
T4_12	0.92	0.23	-0.55	0.27	0.97	-0.09	0.94	-0.02
T4_13	0.93	0.31	-0.70	0.28	0.93	-0.25	0.57	-1.05
T4_14	0.99	0.17	-3.46	1.00	0.97	0.29	0.22	-0.10
T4_15	0.93	0.20	-0.78	0.29	0.99	0.04	0.78	-0.38
T4_16	0.85	0.03	0.17	0.21	1.16	1.19	1.33	1.20
T4_17	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_18	0.93	0.22	-0.70	0.28	0.98	-0.03	0.79	-0.40
T4_19	0.94	0.21	-0.87	0.30	0.97	-0.06	0.91	-0.03
T4_20	0.83	0.30	0.33	0.20	0.97	-0.18	0.83	-0.68
T4_21	0.99	0.16	-2.76	0.71	0.98	0.19	0.40	-0.23
T4_22	0.76	0.36	0.83	0.18	0.96	-0.47	0.85	-0.82
T4_23	0.86	0.30	0.12	0.21	0.97	-0.18	0.75	-0.90
T4_24	0.91	0.24	-0.48	0.26	0.98	-0.05	0.77	-0.53
T4_25	0.93	0.24	-0.70	0.28	0.97	-0.07	0.73	-0.57
T4_26	0.89	0.20	-0.18	0.23	1.02	0.15	0.90	-0.22
T4_27	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_28	0.88	0.11	-0.07	0.23	1.09	0.60	1.06	0.28
T4_29	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_30	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_31	0.98	0.17	-2.34	0.58	0.97	0.13	0.49	-0.31
T4_32	0.94	0.17	-0.97	0.32	0.99	0.05	0.93	0.02
T4_33	0.75	0.23	0.89	0.18	1.06	0.70	1.09	0.56
T4_34	0.92	0.14	-0.55	0.27	1.04	0.26	1.03	0.21
T4_35	0.99	-0.05	-3.46	1.00	1.01	0.34	2.21	1.10
T4_36	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_37	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_38	0.97	-0.02	-1.81	0.46	1.04	0.24	1.41	0.73
T4_39	0.98	0.01	-2.34	0.58	1.01	0.21	1.80	0.98
T4_40_s	0.89	0.08	-0.23	0.24	1.08	0.54	1.22	0.73
T4_41	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_42	0.92	0.17	-0.55	0.27	1.02	0.18	0.85	-0.28
T4_43	0.98	0.22	-2.04	0.51	0.95	0.06	0.45	-0.54
T4_44_s	0.85	0.09	0.21	0.21	1.12	0.94	1.18	0.75
T4_45	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_46	0.99	0.17	-3.46	1.00	0.97	0.29	0.22	-0.10
T4_47	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_48	0.98	0.19	-2.34	0.58	0.97	0.12	0.42	-0.43

Table J2 (cont'd)

T4_49	0.99	0.03	-3.46	1.00	1.00	0.33	0.84	0.48
T4_50	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_51	0.97	0.11	-1.81	0.46	1.01	0.15	0.75	-0.12
T4_52	0.91	0.23	-0.48	0.26	0.98	-0.04	0.89	-0.17
T4_53	0.99	0.03	-3.46	1.00	1.00	0.33	0.84	0.48
T4_54	0.88	0.00	-0.12	0.23	1.14	0.89	1.46	1.40
T4_55	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_56	0.89	0.12	-0.18	0.23	1.05	0.34	1.60	1.68
T4_57	0.95	0.11	-1.07	0.33	1.02	0.15	1.12	0.39
T4_58	0.87	0.19	-0.02	0.22	1.05	0.35	0.88	-0.34
T4_59	0.99	0.13	-3.46	1.00	0.98	0.31	0.31	0.02
T4_60	0.99	0.04	-3.46	1.00	1.00	0.33	0.78	0.44
T4_61	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00
T4_62	0.98	0.15	-2.34	0.58	0.97	0.13	0.77	0.08
T4_63	1.00	NA	-4.67	1.82	1.00	0.00	1.00	0.00

REFERENCES

REFERENCES

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition, 34,* 187-214. https://doi.org/10.1017/S0272263112000022

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

Alderson, J. C., Haapakangas, E., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language.* New York: Routledge.

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236-260. https://doi.org/10.1093/applin/amt046

Alderson, J.C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301-320. https://doi.org/10.1191/0265532205lt310oa

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6*(7), 1-20. https://doi.org/10.1186/s40468-016-0030-z

Amengual, M. (2016). The perception of language-specific phonetic categories does not guarantee accurate phonological representations in the lexicon of early bilinguals. *Applied Psycholinguistics, 37,* 1221-1251. https://doi.org/10.1017/S0142716415000557

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines 2012.* Alexandria, VA: ACTFL.

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

Baker, A. (2014). Exploring teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices, and student perceptions. *TESOL Quarterly, 48,* 136-163.

Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning* (pp. 13–34). Amsterdam: John Benjamins.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences.* New York: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2006). The concept of validity. *Psychological Review, 111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bowles, M. A., Toth, P. D., & Adams, R. J. (2014). A comparison of L2-L2 and L2-heritage learner interactions in Spanish language classrooms. *Modern Language Journal, 92,* 497-517. https://doi.org/10.1111/j.1540-4781.2014.12

Brinkmann, S. (2013). *Qualitative interviewing.* New York: Oxford University Press.

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS ONE, 5*(5), e10773.

Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication, 52,* 980-995. https://doi.org/10.1016/j.specom.2010.08.010

Brown, Adam. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*(4), 593-606. https://doi.org/10.2307/3587258

Brown, Anna. (2018). Item response theory approaches to test scoring and evaluating the score accuracy. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 607-638). Hoboken, NJ: John Wiley & Sons Ltd.

Brown, J. D. (1999). Standard error vs. standard error of measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3*(1), 20-25.

Burri, M., Baker, A., & Chen, H. (2017). "I feel like having a nervous breakdown": Pre-service and in-service teachers' developing beliefs and knowledge about pronunciation instruction. *Journal of Second Language Pronunciation, 3*(1), 109-135. https://doi.org/10.1075/jslp.3.1.05bur

Bybee, J. (2001). *Phonology and language use.* Cambridge, U.K.: Cambridge University Press.

Carr, N. (2011). *Designing and analyzing language tests.* Oxford, UK: Oxford University Press.

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). New York: Cambridge University Press.

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

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chen, Y.-J. (2018). A study on production of Korean syllable-final /n/ and /ŋ/ by Taiwanese speakers. Unpublished master's thesis. Hankuk University of Foreign Studies.

Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12,* 235– 262. https://doi.org/10.1177/1362168807086293

Choi, E., Kim, E., Park, H., Jin, M., & Park, K. (2009a). 외국인을 위한 한국어 발음 (제 1 권) [Korean pronunciation for foreigners (Vol. 1)]. Seoul: SISA Hangeulpark.

Choi, E., Kim, E., Park, H., Jin, M., & Park, K. (2009b). 외국인을 위한 한국어 발음 (제 2 권) [Korean pronunciation for foreigners (Vol. 2)]. Seoul: SISA Hangeulpark.

Council of Europe. (2017). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume with new descriptors.* Retrieved from https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2

Creswell, J. W., & Plano Clark, V. L. (2011). Designing and conducting mixed methods research (2nd ed.). Thousand Oaks, CA: Sage.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, and Winston.

Crowther, D., Isaacs, T., Trofimovich, P., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *Modern Language Journal, 99*(1), 80-95. https://doi.org/10.1111/modl.12185

Cutler, A., & Clifton, C. (1999). Comprehending spoken language: A blueprint of the listener. In C. M. Brown and P. Hagoort (Eds.), *The Neurocognition of Language* (pp 123-166). Oxford: Oxford University Press.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America, 116*(6), 3668-3678. https://doi.org/10.1121/1.1810292

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15-32). New York: Routledge.

DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 49-73). Hoboken, NJ: John Wiley & Sons Ltd.

Derwing, T. M., Diepenbrooke, L. G., & Foote, J. A. (2012). How well do general-skills ESL textbooks address pronunciation? *TESL Canada Journal, 30*(1), 23-44.

Derwing, T., & Munro, M. J. (2014). Myth 1: Once you have been speaking a second language for years, it's too late to change your pronunciation. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 34-55). Ann Arbor, Michigan: Michigan University Press.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, *63*(2), 163-185. https://doi.org/10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*(3), 526-548. https://doi.org/10.1111/lang.12053

Derwing, T. M., Munro, M. J., & Weibe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393-410.

Dimova, S., & Kling, J. (2018). Assessing English-medium instruction lecturer language proficiency across disciplines. *TESOL Quarterly*, *52*(3), 634-656.

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System, 36,* 506-516. https://doi.org/10.1016/j.system.2008.03.003

Dorans, N. J. (2018). Scores, scales, and score linking. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 571-605). Hoboken, NJ: John Wiley & Sons Ltd.

Duñabeitia, J.A., Crepaldi, D., Meyer, A.S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2017). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470218.2017.1310261

Eckes, T. (2014). Examining teslet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. https://doi.org/10.1177/0265532213492969

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence.' *Language Testing, 21*(3), 259-283. https://doi.org/10.1191/0265532204lt284oa

Elder, C., & von Randow, J. (2008). Exploring the utility of a web-based English language screening tool. *Language Assessment Quarterly, 5*(3), 173-194. https://doi.org/10.1080/15434300802229334

Elo, S., & Kyngäs, H. (2007). The qualitative content analysis process. *Journal of Advanced Nursing, 62*(1), 107-115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition, 32*, 181-201. https://doi.org/10.1017/S0272263109990490

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking* (pp. 65-111). Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language* (pp. 77–151). Cambridge: Cambridge University Press.

Field, J. (2014). Myth 3: Pronunciation teaching has to establish in the minds of language learners a set of distinct consonant and vowel sounds. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 80-106). Ann Arbor, Michigan: Michigan University Press.

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to l2 phonological learning. In T. Hueber and C. Ferguson (Eds.), *Crosscurrents in second language acquisition and linguistic theories* (pp. 249-289). Amsterdam: John Benjamins.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and the critical period hypothesis* (pp. 233-277). Timonium, MD: York Press.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41*(1), 78-104. https://doi.org/10.1006/jmla.1999.2638

Foote, J., Holtby, A., & Derwing, T. M. (2011). Survey of the teaching of pronunciation in adult ESL programs in Canada, 2010. *TESL Canada Journal, 29,* 1-22. https://doi.org/10.18806/tesl.v29i1.1086

Foote, J. A., McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation, 3*(1), 34-56. https://doi.org/10.1075/jslp.3.1.02foo

Fulcher, G. (Host). (2015, July). *Issue 22: Eunice Jang on Diagnostic Language Testing* [Audio Podcast]. Retrieved from http://languagetesting.info/sage/podcasts/Diagnostic%20Language%20Testing.mp3

Friedman, D. (2012). How to collect and analyze qualitative data. In A. Mackey and S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 180-200). London: Wiley.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement (R package version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Gass, S., & Mackey, A. (2006). Input, interaction, and output: An overview. *AILA Review, 19,* 3-17.

Gilbert, J. B. (2005). *Clear speech – Pronunciation and listening comprehension in North American English: Student's book* (3rd ed.). New York: Cambridge University Press.

Ginther, A., & Yan, X. (2018). Interpreting relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35*(2), 271-295. https://doi.org/10.1177/0265532217704010

Graham, C. (2001). *Jazz chants old and new.* Oxford, England: Oxford University Press.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29-48. https://doi.org/10.1348/000711006X126600

Han, Z.-H. (2004). *Fossilization in adult second language acquisition.* Clevedon, UK: Multilingual Matters.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317-336. https://doi.org/10.1177/0265532214564505

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: quantitative and qualitative findings. *Language Learning & Technology, 8*, 34-52.

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics, 26*(4), 579–596. https://doi.org/10.1017/S0142716405050319

Hardison, D. M. (2012). Second-language speech perception: A cross-disciplinary perspective on challenges and accomplishments. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 349–363). London: Routledge.

Hardison, D. M. (2018). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse: Theoretical and pedagogical implications. *Journal of Second Language Pronunciation, 4*(2), 232-259. https://doi.org/10.1075/jslp.17006.har

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer Science+Business Media.

Holliday, J. J. (2014). The perceptual assimilation of Korean obstruents by native Mandarin listeners. *The Journal of the Acoustical Society of America, 135*, 1585-1595. https://doi.org/10.1121/1.4863653

Holliday, J. J. (2015). A longitudinal study of the second language acquisition of a three-way stop contrast. *Journal of Phonetics, 50*, 1-14. https://doi.org/10.1016/j.wocn.2015.01.004

Holliday, J. J. (2016). Second language experience can hinder the discrimination of nonnative phonological contrasts. *Phonetica, 73*, 33-51. https://doi.org/10.1159/000443312

Horgues, C., & Scheuer, S. (2014). "I understood you, but there was this pronunciation thing…": L2 pronunciation feedback in English/French tandem interactions. *Research in Language, 12*(2), 145-161. https://doi.org/10.2478/rela-2014-0005

Housen, A., & Pierrard, M. (2005). *Investigations in instructed second language acquisition.* Berlin: Mouton de Gruyter.

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America*, *117*(2), 896. https://doi.org/10.1121/1.1823291

Ingvalson, E. M., Ettlinger, M., & Wong, P. C. M. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism*, *18*(1), 35-47. https://doi.org/10.1177/1367006912456586

Irwing, P., & Hughes, D. J. (2018). Test development. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 3-47). Hoboken, NJ: John Wiley & Sons Ltd.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273-293. https://doi.org/10.1080/15434303.2018.1472264

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*, *50*(3), 347-366. https://doi.org/10.1017/S0261444817000118

Isaacs, T., & Trofimovich, P. (Eds.). (2017). *Second language pronunciation assessment: Interdisciplinary perspectives.* Bristol, UK: Multilingual Matters.

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193-216. https://doi.org/10.1177/0265532217703433

Isbell, D. R. (2017). Explaining intelligibility: What matters most in L2 speech? Paper presented at the 19th annual Second Language Research Forum, Columbus, Ohio.

Isbell, D. R., Park, O.-S., & Lee, K. (2019). Learning Korean Pronunciation: Effects of Instruction, Proficiency, and L1. *Journal of Second Language Pronunciation*, *5*(1), 13-48. https://doi.org/10.1075/jslp.17010.isb

Isbell, D. R., Winke, P. M., & Gass, S. M. (2018). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*. Online Early Access. https://doi.org/10.1177/0265532218798139

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to *LanguEdge* assessment. *Language Testing*, *26*(1), 31-73. https://doi.org/10.1177/0265532208097336

Jang, E. E., Dunlop, M., Park, G., & Van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*(3), 359-383. https://doi.org/10.1177/0265532215570924

Jang, E. E., & Wagner, M. (2014). Diagnostic feedback in language classroom. In A. Kunnan (Ed.), *Companion to language assessment* (vol. 2, pp. 693–711). New York, NY: Wiley-Blackwell.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllable for English as an international language. *Applied Linguistics, 23*(1), 83-103.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kang, O., & Ginther, A. (Eds.). (2017). *Assessment in second language pronunciation*. New York: Routledge.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly, 48*(1), 176-187. https://doi.org/10.1002/tesq.152

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., Thomson, R. I, & Moran, M. (2018a). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning, 68*(1), 115-146. https://doi.org/10.1111/lang.12270

Kang, O., Thomson, R. I., & Moran, M. (2018b). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics*, Online advance access. https://doi.org/10.1093/applin/amy053

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA.

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and visualize the results of multivariate data analyses, R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Kennedy, S., Guénette, D., Murphy, J., & Allard, S. (2015). Le rôle de la prononciation dans l'intercompréhension entre locuteurs de français lingua franca [The role of pronunciation on comprehension between speakers of French as a lingua franca]. Canadian Modern Language Review, 71, 1–25. https://doi.org/10.3138/cmlr.2139

Kennedy, S., & Trofimovich, P. (2010). Language awareness and second language pronunciation: A classroom study. *Language Awareness, 19*(3), 171-185. https://doi.org/10.1080/09658416.2010.486439

Kim, E.-A. (2006). A study on the diagnosis & evaluation for pronunciation errors of Korean language learners [한국어 학습자의 발음 오류 진단 및 평가에 관한 연구]. *Journal of Korean Language Education [한국어교육], 17*(1), 71-99.

Kim, J. E., & Silva, D. J. (2003). Accounting for back-vowel under-differentiation: An acoustically-based study of English-speaking learners of Korean. *The Korean Language in America, 8*, 51-64.

Kim, J.-Y. (2015). Second language acquisition: Phonology. In L. Brown & J. Yeon (Eds.), *The Handbook of Korean Linguistics* (pp. 373-388). Hoboken, NJ: John Wiley & Sons.

Kim, M. (2007). *Aspects of Korean second language phonology* (Doctoral dissertation). Retrieved from ProQuest. (UMI 3279086)

Kim, C.-W., & Park, S.-G. (1995). Pronunciation problems of Australian students learning Korean: Intervocalic liquid consonants. *Australian Review of Applied Linguistics, 12,* 183-202. https://doi.org/10.1075/aralss.12.12kim

Kim, M. J., Pae, S. Y., & Lee, S. E. (2005). The development of the 'Test of Articulation for Children': Concurrent validity. *Communication Sciences & Disorders, 10*(1), 82-96.

Kim, M., Kim, S.-J., & Stoel-Gammon, C. (2017). Phonological acquisition of Korean consonants in conversational speech produced by young Korean children. *Journal of Child Language, 44*, 1010-1023. https://doi.org/10.1017/S0305000916000258

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *Modern Language Journal, 100*(3), 655-673. https://doi.org/10.1111/modl.12346

King, R. S. (2015). *Cluster analysis and data mining: An introduction.* Boston, MA: Mercury Learning and Information.

Knoch, U., & Elder, C. (2016). Post-entry English language assessments at university: How diagnostic are they? In V. Aryadoust & J. Fox (eds.), *Trends in Language Assessment Research and Practice: The View from the Middle East and Pacific Rim* (pp. 210-230). Newcastle-upon-Tyne, England: Cambridge Scholars Publishing.

Ko, I. (2013). *The articulation of Korean coronal obstruents: Data from heritage speakers and second language learners* (Unpublished doctoral dissertation). University of Hawai'i, Hawai'i.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Krashen, S. (1982). *Principles and practice in second language acquisition.* New York: Prentice Hall.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Kwon, S. (2017). 한국어 발음 교육론 [Korean pronunciation pedagogy]. Seoul: SISA Hangeulpark.

Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals, 47*(2), 300-320. https://doi.org/10.1111/flan.12083

Lado, R. (1957). *Linguistics across cultures.* Ann Arbor, MI: University of Michigan Press.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests: A teacher's book.* London: Longmans, Green and Company.

Lee, A. H., & Lyster, R. (2016). Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning, 66*(4), 809-833. https://doi.org/10.1111/lang.12167

Lee, A. H., & Lyster, R. (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics, 38*, 371-393. https://doi.org/10.1017/S0142716416000254

Lee, H. (2017a). An empirical study to rethink the goals and components of teaching Korean language pronunciation. *Journal of Korean Language Education, 28*(3), 105-126.

Lee, H. (2017b). 한국어 발음 평가 연구 [Korean Pronunciation Assessment for Foreign Language Speakers]. Seoul: Jisikgwagyoyang.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 1–23. https://doi.org/10.1093/applin/amu040

Lee, S. (2012). *Orthographic influence on the phonological development of L2 learners of Korean* (Unpublished doctoral dissertation). The University of Wisconsin-Milwaukee, Milwaukee, WI.

Lee, S-H., Jang, S. B., Seo, S. K. (2017). *A frequency dictionary of Korean.* New York: Routledge.

Lee, S.-Y., Moon, J., & Long, M. H. (2009). Linguistic correlates of proficiency in Korean as a second language. *Language Research, 45*(2), 319-348.

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 299-316. https://doi.org/10.1177/0265532214565387

Lee, Y.-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 172-189. https://doi.org/10.1080/15434300902985108

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch analysis. *Language Testing, 26*(2), 245-274. https://doi.org/10.1177/0265532208101007

Levelt, W. J. M. (1993). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39,* 39-377. https://doi.org/10.2307/3588485

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics, 27,* 184-202. https://doi.org/10.1017/S0267190508070098

Levis, J., & Barriuso, T. A. (2012). Nonnative speakers' pronunciation errors in spoken and read English. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 3rd Annual Pronunciation in Second Language Learning and Teaching Conference* (pp. 187-194). Ames, IA: Iowa State University.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2005). Dichotomous & polytomous category information. *Rasch Measurement Transactions, 19*(1), 1005-1006.

Linacre, J. M. (2019). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Little, D. 2005. The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22*(3), 321–336. https://doi.org/10.1191/0265532205lt311oa

Llama, R., Cardoso, W., & Collins, L. (2010). The influence of language distance and language status on the acquisition of L3 phonology. *International Journal of Multilingualism, 7*(1), 39-57. https://doi.org/10.1080/14790710902972255

Loewen, S. (2015). *An introduction to instructed second language acquisition.* New York: Routledge.

Loewen, S., & Isbell, D. R. (2017). Pronunciation in face-to-face and oral synchronous computer mediated interaction. *Studies in Second Language Acquisition, 39*(2), 225-256. https://doi.org/10.1017/S0272263116000449

Long, M. (2013). Maturational constraints on child and adult SLA. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 3-41). Amsterdam: John Benjamins.

Lord, G. (2005). Can we teach foreign language pronunciation? The effects of a phonetics class on second language pronunciation. *Hispania, 88,* 557-567.

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals, 41,* 364-379.

Lord, G. (2010). The combined effects of immersion and instruction on second language pronunciation. *Foreign Language Annals, 43*(3), 487-503.

Ma, M., & Winke, P. (2019). Self-assessment: How reliable is it in assessing the oral proficiency of Chinese learners over time? *Foreign Language Annals, 52,* 66-86. https://doi.org/10.1111/flan.12379

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9,* 200–215.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research, 50*(4), 940-967. https://doi.org/10.1044/1092-4388(2007/067)

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324. doi:10.3758/s13428-011-0168-7

Matsumoto, Y. (2011). Successful ELF communications and implications for ELT: Sequential analysis of ELF pronunciation negotiation strategies. *Modern Language Journal, 95,* 97–114. https://doi.org/10.1111/j.1540-4781.2011.01172.x

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98-118. https://doi.org/10.1075/jslp.16034.mcc

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

McLeod, S., & Crowe, S. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology*, 1-28. https://doi.org/10.1044/2018_AJSLP-17-0100

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(3), 621-638. https://doi.org/10.1016/0010-0285(86)90015-0

Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional Item Response Theory. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 413-443). Hoboken, NJ: John Wiley & Sons Ltd.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Miles, M. B., Huberman, M. A., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Thousand Oaks, CA: SAGE.

Möttönen, R., & Watkins, K.E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. The Journal of Neuroscience, 29(31), 9819-9825. https://doi.org/10.1523/JNEUROSCI.6018-08.2009

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics, 35*(4), 418-440. https://doi.org/10.1093/applin/amu012

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 193-218). Philadelphia, PA: John Benjamins.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520-531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., Derwing, T. M., & Thomson, R. I. (2015). Setting segmental priorities for English learners: Evidence from a longitudinal study. *International Review of Applied Linguistics, 53*(1), 39-60. https://doi.org/10.1515/iral-2015-0002

Murphy, J. (2014). Myth 7: Teacher training programs provide adequate preparation in how to teach pronunciation. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 188-234). Ann Arbor, Michigan: Michigan University Press.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31,* 274-295. https://doi.org/10.1007/s00357-014-9161-z

Nation, I. S. P. (2001). Planning and running an extensive reading program. *NUCB Journal of Language Culture and Communication, 3*(1), 1-8.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

National Institute of the Korean Language. (2015, May 29). "초코렛"과 "초콜릿" ["chocoret" and "chocolate"]. Retrieved from https://www.korean.go.kr/front/mcfaq/mcfaqView.do?mn_id=&mcfaq_seq=5497&pageIndex=1

Nissen, S., & Shedl, M. (2012). Prototyping new item types. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 281-294). London: Routledge.

Nora, A., Revall, H., Kim, J.-Y., Service, E., & Salmelin, R. (2015). Distinct effects of memory retrieval and articulatory preparation when learning and accessing new word forms. PLOS One, 10(5), 1-27. https://doi.org/10.1371/journal.pone.0126652

Oh, J. S., Jun, S.-A., Knightly, L. M., & Au, T. K. M. (2003). Holding on to childhood language memory. *Cognition, 86*, B53-B64. https://doi.org/10.1016/S0010-0277(02)00175-0

Oh, Y. M., Coupé, C., Marsico, E., & Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. Journal of Phonetics, 53, 153-176. https://doi.org/10.1016/j.wocn.2015.08.003

Pearson. (2018). *PTE Academic Score Guide.* Retrieved from https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*(10), 1-8. https://doi.org10.3389/neuro.11.010.2008

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educational Psychologist, 51(1), 59-81. https://doi.org/10.1080/00461520.2016.1145550

Pennington, M. C. (1998). The teachability of phonology in adulthood: A re-examination. *International Review of Applied Linguistics, 36*(4), 323-341.

Pinkfong. (2016). 상어 가족 [Baby Shark] [music video]. Seoul: SmartStudy. Retrieved at https://youtu.be/761ae_KDg_Q

Piske, T., MacKay, I. R. A., Flege, J. E., (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191-215. https://doi.org/10.1006/jpho.2001.0134

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/10.1111/lang.12079

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research, 17*(3), 323-342. https://doi.org/10.1177/1362168813482935

Qian, M., Chukharev-Hudilainen, E., & Levis, J. (2018). A system for adaptive high-variability segmental perceptual training: Implementation, effectiveness, transfer. *Language Learning & Technology, 22*(1), 69-96. https://doi.org/10125/44582

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: MESA Press.

Redd, R. (2019). ragree: Rater agreement (R package version 0.0.4) [Computer software]. https://github.com/raredd/ragree

Revelle, W. (2018). psych: Procedures for personality and psychological research (R package version 1.8.12) [computer software]. Evanston, Illinois: Northwestern University.

Richards, J. (1969). Songs in language learning. *TESOL Quarterly, 3*(2), 161-174.

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning, 45*(2), 283-331.

Ryan, E., & Brunfaut, T. (2016). When the test developer does not speak the target language: The use of language informants in the test development process. *Language Assessment Quarterly, 13*(4), 393-408. https://doi.org/10.1080/15434303.2016.1236110

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second-language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics, 39*, 187-224. https://doi.org/10.1017/S0142716417000418

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. TESOL Quarterly, 46(4), 842-854. https://doi.org/10.1002/tesq.67

Saito, K. (2018). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/ pronunciation. *Second Language Research.* Online advance publication. https://doi.org/10.1177/0267658318768342

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation of /ɹ/ by Japanese learners of English. *Language Learning, 62*(2), 595-633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K., & Plonsky, L. (in press). Effects of second language pronunciation teaching revisted: A proposed measurement framework and meta-analysis. *Language Learning.*

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37,* 217-240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439-462. https://doi.org/10.1093/applin/amv047

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129-158.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13,* 206-226.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to Learn: Conversation in Second Language Acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Schreier, M. (2014). Qualitative content analysis. In U. Flick (Ed.), *The SAGE Handbook of Qualitative Data Analysis* (pp. 170-183). Thousand Oaks, CA: SAGE. http://doi.org/10.4135/9781446282243.n12

Seok, D.-I., Park, S.-H., Shin, H.-J., & Park, J.-H. (2002). A study on the development of Korean Standard Picture Articulation Test. *Communication Sciences & Disorders, 7*(3), 121-143.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3,* 243-261.

Shin, E. (2007). How do non-heritage students learn to make the three-way contrast of Korean stops? *The Korean Language in America, 12,* 85-105.

Shin, J., Kiaer, J., & Cha, J. (2013). *The sounds of Korean*. New York: Cambridge University Press.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Smith, B. L., Johson, E., & Hayes-Harb, R. (2019). ESL learners' intra-speaker variability in producing American English tense and lax vowels. *Journal of Second Language Pronunciation, 5*(1), 139-164. https://doi.org/10.1075/jslp.15050.smi

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 243-274). New York: Routledge.

Steinley, D. (2004). Standardizing variables in *K*-means clustering. In D. Banks, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications: Studies in Classification, Data Analysis, and Knowledge Organisation.* Berlin: Springer.

Stevens, S. S. (1946). On the theory of scales and measurement. *Science, 103*(2684), 677-680.

Stockwell, R., & Bowen, J. (1965). *The sounds of English and Spanish*. Chicago, IL: University of Chicago Press.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699.

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing, 35*(2), 217-238. https://doi.org/10.1177/0265532217690782

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics, 38*, 1229-1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning, 67*, 747-790. https://doi.org/10.1111/lang.12241

Tan, M., & Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. *Language Assessment Quarterly, 12*, 29-49. https://doi.org/10.1080/15434303.2014.1003301

Tark, E. S. (2016). *Acquisition of Korean obstruents by English-speaking second language learners of Korean and the role of pronunciation instruction* (Doctoral dissertation). Retrieved from ProQuest. (No. 10191441)

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology, 16*(3), 10-20. http://llt.msu.edu/issues/october2012/action.pdf

Thomson, R. I. (2011). Computer Assisted Pronunciation Training: Targeting second language vowel perception improves pronunciation. *CALICO Journal, 28,* 744-765.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. Language Learning, 62(4), 1231-1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R. I. (2016). English Accent Coach [Computer program]. Version 2.3. www.englishaccentcoach.com

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326–344. https://doi.org/10.1093/applin/amu076

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

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122-140. https://doi.org/10.1017/S1366728914000832

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369.

Turner, C. E., & Purpura, J. E. (2015). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), Handbook of Second Language Assessment (pp. 255-272). Boston: De Gruyter Mouton.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action.* New York, NY: Routledge.

VanPatten, B., & Rothman, J. (2015).What does current generative theory have to say about the explicit-implicit debate? In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 89-116). Amsterdam: John Benjamins.

Venkatagiri, H. S., & Levis, J. M. (2007). Phonological awareness and speech comprehensibility: An exploratory study. *Language Awareness, 16*(4), 263-277. https://doi.org/10.2167/la417.0

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. Journal of Memory and Language, 50, 1-25. https://doi.org/10.1016/S0749-596X(03)00105-0

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Yan, X., & Ginther, A. (2017). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. Language Testing (Online early access). https://doi.org/10.1177/0265532217704010

Yeldham, M., & Gruba, P. (2014). Toward an instructional approach to developing interactive second language listening. *Language Teaching Research, 18,* 33–53. https://doi.org/10.1177/1362168813505395

Yu, H. J. (2016). The development of obstruent consonants in bilingual Korean-English children (Doctoral dissertation). Retrieved from ProQuest. (No. 10163769)

Zoghbor, W. S. (2018). Teaching English pronunciation to multi-dialect first language learners: The revival of the Lingua Franca Core (LFC). *System, 78,* 1-14. https://doi.org/10.1016/j.system.2018.06.008